Numerikus módszerek

Galántai Aurél-Jeney András

2005

Tartalomjegyzék

1. fejezet

BEVEZETÉS

A numerikus módszerekre többféle irányból és többféle célból tekinthetünk. Az egyik szélen a tiszta matematika tudósa áll. Az ő nézőpontjából a numerikus analízis nem más, mint tudományos kutatások tárgya, amelynek művelése, fejlesztése maga a cél. A másik szélről viszont a numerikus módszerekben csupán a konkrét gyakorlati feladatok megoldását segítő eszközöket láthatunk.

Az a tankönyv, amit most az Olvasó a kezében tart, elsősorban műszaki egyetemi képzésben résztvevők számára készült. A megírásakor a különböző célok között a hangsúlyt a gyakorlati alkalmazhatóság irányában helyeztük el, azt is szem előtt tartva, hogy főiskolai hallgatók is haszonnal forgathassák. Ugyanakkor nem mondhattunk le teljesen az elméleti megalapozásról, definíciók, tételek pontos kimondásáról sem. Egy-két egészen új eredményt is beiktattunk, különösen a modern hibaanalízis területéről. Nem bizonyítottunk minden tételt, néhány helyen csupán a motivációt adtuk meg, de egyes helyeken teljes matematikai szigorúsággal igazoltuk az állítást. A közölt bizonyításokkal egyrészt ízelítőt kívántunk adni egyéb matematikai apparátusoknak a numerikus módszereken belüli alkalmazásaiból, másrészt a tétel szűken vett jelentését meghaladó mondanivalóra igyekeztünk a figyelmet felhívni. Mindezekkel együtt tudományegyetemi hallgatóknak sem lesz talán érdektelen a könyv lapozgatása.

A könyv anyagának oktatása előtt mintegy két féléves analízisbeli tanulmányokat tételezünk fel. A kifejtés terjedelme fejezetenként eltérő. A klasszikus hibaszámítás mellett külön figyelmet fordítottunk a lebegőpontos hibaanalízisre. Legrészletesebben a lineáris algebra numerikus eljárásait tárgyaltuk, tekintettel azok elterjedt műszaki alkalmazásaira. A gyakorlati, sokszor nagyméretű feladatok numerikus megoldása manapság szinte kizárólag számítógéppel történik. Éppen ezért kitértünk a tárolási módokra is és minden eljárást a számítástechnikai megvalósítás szempontjai szerint fogalmaztunk meg. Annak érdekében, hogy az algoritmusokat tömören, közérthetően, szó szerinti (vagy majdnem szó szerinti) program formájában is leírhassuk, a könyv tartalmazza a MATLAB programozási

nyelv legfontosabb elemeit. A szövegközi, kidolgozott példákkal a tárgyalt probléma mélyebb megértését céloztuk meg, a fejezetek végén felsorolt néhány feladat pedig, szándékunk szerint, önálló munkára serkenti az Olvasót.

A forrásmunkákon kívül széles szakmai látókörrel és gyakorlati tapasztalattal rendelkező kollégák észrevételei is segítették a könyv elkészülését. Köszönetünket fejezzük ki Balla Katalinnak, Demendy Zoltánnak és Stoyan Gisbertnek a soksok értékes megjegyzésért és tanácsért, amellyel hozzájárultak e tankönyv végleges formába öntéséhez.

A jelen kiadás a 2002-es kiadás alig módosított változata. A kisebb tipográfiai változásokon kívüli tartalmi módosítást elsősorban az újabb processzorok és a MATLAB újabb változatainak megjelenése, valamint az oktatás során szerzett tapasztalatok indokolták.

2. fejezet

A MÁTRIXSZÁMÍTÁS ELEMEI

2.1. Mátrixok és mátrixműveletek

Röviden összefoglaljuk azokat a mátrixokkal és vektorokkal kapcsolatos ismereteket, amelyekre szükségünk van.

1.1 Definíció. *Legyenek n és m pozitív egész számok. Egy A $m \times n$ típusú (valós) mátrixon valós a_{ij} számok alábbi táblázatát értjük:*

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mj} & \dots & a_{mn} \end{bmatrix}.$$

Az a_{ij} az A mátrix i-edik sorában és j-edik oszlopában álló mátrixelemet jelöli. Mátrixok szokásos jelölése még a következő:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mj} & \dots & a_{mn} \end{pmatrix}.$$

Néhány tömörebb mátrixmegadási mód:

$$A = [a_{ij}]_{i,j=1}^{m,n}, \qquad A = (a_{ij})_{i,j=1}^{m,n}.$$

Az $m \times n$ típusú valós mátrixok halmazát $\mathbb{R}^{m \times n}$ jelöli. Az A mátrixot *négyzetesnek* nevezzük, ha $m = n$. Ekkor a tömör megadási módok a következőképpen

egyszerűsödnek:

$$A = [a_{ij}]_{i,j=1}^{n}\,, \qquad A = (a_{ij})_{i,j=1}^{n}\,.$$

A mátrixok közti fontosabb műveleteket az alábbiak szerint definiáljuk.

1. Összeadás: $A, B \in \mathbb{R}^{m\times n}$,

$$C = A + B \in \mathbb{R}^{m\times n} \Leftrightarrow c_{ij} = a_{ij} + b_{ij} \quad (i = 1,\dots,m,\ j = 1,\dots,n).$$

Az összeadásra fennáll, hogy

$$A + B = B + A\,, \quad (A + B) + C = A + (B + C).$$

2. Számmal való szorzás: $A \in \mathbb{R}^{m\times n}$, λ valós szám,

$$C = \lambda A \in \mathbb{R}^{m\times n} \Leftrightarrow c_{ij} = \lambda a_{ij} \quad (i = 1,\dots,m,\ j = 1,\dots,n).$$

A számmal való szorzásra fennáll, hogy

$$\lambda(\mu A) = (\lambda\mu)A, \quad (\lambda + \mu)A = \lambda A + \mu A.$$

Jegyezzük meg, hogy megállapodás szerint $\lambda A = A\lambda$.

3. Transzponálás (tükrözés): $A \in \mathbb{R}^{m\times n}$,

$$C = A^T \in \mathbb{R}^{n\times m} \Leftrightarrow c_{ij} = a_{ji} \quad (i = 1,\dots,n,\ j = 1,\dots,m).$$

A transzponálásra fennáll, hogy

$$(A^T)^T = A, \quad (A + B)^T = A^T + B^T.$$

Az A mátrixot *szimmetrikusnak* nevezzük, ha $A^T = A$.

4. Szorzás: $A \in \mathbb{R}^{m\times k}$, $B \in \mathbb{R}^{k\times n}$,

$$C = AB \in \mathbb{R}^{m\times n} \Leftrightarrow c_{ij} = \sum_{t=1}^{k} a_{it}b_{tj} \quad (i = 1,\dots,m,\ j = 1,\dots,n).$$

Vegyük észre, hogy a szorzatmátrix (i, j) indexű elemét úgy kapjuk, hogy az i-edik sort szorozzuk a j-edik oszloppal, azaz

$$c_{ij} = [a_{i1},\dots,a_{ik}] \begin{bmatrix} b_{1j} \\ \vdots \\ b_{kj} \end{bmatrix}.$$

A mátrixszorzás fontos tulajdonságai a következők:

$$\begin{aligned} (AB)C &= A(BC), \\ A(B+C) &= AB+AC, \\ (A+B)C &= AC+BC, \\ (AB)^T &= B^TA^T. \end{aligned}$$

Fontos megjegyezni, hogy a szorzás nem kommutatív, tehát általában

$$AB \neq BA. \tag{2.1}$$

A továbbiakban a mátrix és mátrix-vektor műveletek felírásánál feltesszük, hogy az ott szereplő mátrixok, ill. vektorok méretei olyanok, amelyek lehetővé teszik az adott műveletet.

1.2 Definíció. *Az egyetlen sorból, vagy egyetlen oszlopból álló mátrixot vektornak nevezzük.*

A sorvektor szokásos megadási módja: $x = [x_1, \ldots, x_n]$. Az oszlopvektorokat az

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$$

formában szoktuk megadni, ahol $\mathbb{R}^n$ az n komponensű oszlopvektorok halmaza (tulajdonképpen $\mathbb{R}^n \equiv \mathbb{R}^{n\times 1}$). Az oszlopvektorokat meg lehet még adni $x = [x_1, \ldots, x_n]^T$, a sorvektorokat pedig az

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}^T \in \mathbb{R}^n$$

formában is.

Az i-edik egységvektornak nevezzük azt a vektort, amelynek i-edik komponense 1, a többi pedig 0. Oszlopvektornak tekintve tehát:

$$e_i = [0, \ldots, 0, 1, 0, \ldots, 0]^T \in \mathbb{R}^n.$$

1.3 Definíció. *$x, y \in \mathbb{R}^n$ skaláris szorzata a*

$$x^Ty = \sum_{i=1}^{n} x_iy_i$$

képlettel definiált valós szám.

A következőkben mátrixok részekre bontásával (particionálásával) foglalkozunk. Az A $m \times k$ típusú mátrixot sorok szerint particionáljuk, ha

$$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_i^T \\ \vdots \\ a_m^T \end{bmatrix},$$

ahol $a_i^T = [a_{i1}, \ldots, a_{ik}]$ az i-edik (k-dimenziós) sorvektort jelöli. A B $k \times n$ típusú mátrixot oszlopok szerint particionáljuk, ha

$$B = [b_1, \ldots, b_n],$$

ahol

$$b_i = \begin{bmatrix} b_{1i} \\ \vdots \\ b_{ki} \end{bmatrix}$$

az i-edik (k-dimenziós) oszlopvektort jelöli. A fenti particionálások felhasználásával az A és B mátrixok szorzata felírható

$$AB = \left[a_i^T b_j\right]_{i,j=1}^{m,n}$$

alakban is. Tehát AB a sorok és oszlopok skalárszorzataiból álló mátrix. Az AB mátrixszorzatot felírhatjuk még a következő alakokban is

$$AB = [Ab_1, \ldots, Ab_n], \quad AB = \begin{bmatrix} a_1^T B \\ \vdots \\ a_m^T B \end{bmatrix}.$$

Megjegyezzük, hogy más típusú particionálások is lehetségesek. Ezek közös alakja az $A \in \mathbb{R}^{m \times n}$ mátrix esetén

$$A = \begin{bmatrix} A_{11} & \ldots & A_{1j} & \ldots & A_{1r} \\ \vdots & & \vdots & & \vdots \\ A_{i1} & \ldots & A_{ij} & \ldots & A_{ir} \\ \vdots & & \vdots & & \vdots \\ A_{p1} & \ldots & A_{pj} & \ldots & A_{pr} \end{bmatrix},$$

ahol $A_{ij} \in \mathbb{R}^{m_i \times n_j}$ $(i = 1, \ldots, p,\ j = 1, \ldots, r)$. Az azonos sorban álló blokkok sorainak száma azonos. Hasonlóképpen, az azonos oszlopban álló blokkok oszlopainak száma azonos. Fennáll még az értelemszerű

$$\sum_{i=1}^{p} m_i = m, \quad \sum_{j=1}^{r} n_j = n$$

összefüggés. Azonosan particionált mátrixok összegzését és skalárral való szorzását blokkonként végezhetjük, úgy mintha a blokkok számok lennének. Particionált mátrixok blokkonkénti szorzásánál az első mátrix oszlopok szerinti particionálása meg kell, hogy egyezzen a második tényező sorok szerinti particionálásával. Például az

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \in \mathbb{R}^{m\times n}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \in \mathbb{R}^{n\times p}$$

particionált mátrixok szorzata a particionálás megtartásával akkor lehetséges, ha $A_{11} \in \mathbb{R}^{r\times s}$ és $B_{11} \in \mathbb{R}^{s\times\sigma}$. Ekkor

$$C = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}.$$

A következőkben speciális tulajdonságú mátrixokat vezetünk be.

1.4 Definíció. *Az $I \in \mathbb{R}^{n\times n}$ mátrix egységmátrix, ha*

$$I = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & 0 \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix}.$$

Az egységmátrixra fennáll, hogy minden $A \in \mathbb{R}^{n\times n}$ esetén

$$AI = IA = A.$$

1.5 Definíció. *A $D \in \mathbb{R}^{n\times n}$ diagonálmátrix, ha*

$$D = \begin{bmatrix} d_1 & 0 & \dots & \dots & 0 \\ 0 & d_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & d_{n-1} & 0 \\ 0 & \dots & \dots & 0 & d_n \end{bmatrix}.$$

A diagonálmátrixra fennáll, hogy minden $A \in \mathbb{R}^{n\times m}$ és $B \in \mathbb{R}^{m\times n}$ esetén

$$DA = D \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix} = \begin{bmatrix} d_1 a_1^T \\ \vdots \\ d_n a_n^T \end{bmatrix}, \quad BD = [b_1, \dots, b_n]\, D = [d_1 b_1, \dots, d_n b_n]\,.$$

A D diagonálmátrixot $\mathrm{diag}(d_1, \dots, d_n)$, vagy $\mathrm{diag}(d_i)$ $(i = 1, \dots n)$ is jelölheti.

1.6 Definíció. *Az* $0 \in \mathbb{R}^{m\times n}$ *mátrix zérusmátrix, ha minden eleme* 0*, azaz*

$$0 = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}.$$

A zérusmátrixra fennáll, hogy minden A mátrix esetén

$$A + 0 = A, \quad A0 = 0.$$

2.2. Mátrixok inverze és determinánsa

1.7 Definíció. *Az* $X \in \mathbb{R}^{n\times n}$ *mátrixot az* $A \in \mathbb{R}^{n\times n}$ *mátrix inverzének nevezzük, ha* $AX = XA = I$*.*

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

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

Jelölje $A(i)$ azt az $(n-1)\times(n-1)$-es mátrixot, amelyet az

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \dots & a_{in} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

mátrixból az első oszlop és az i-edik sor elhagyásával kapunk.

1.8 Definíció. *Az* $A \in \mathbb{R}^{n\times n}$ *(*$n \geq 2$*) négyzetes mátrix determinánsát a*

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

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

előírások definiálják.

Megjegyezzük, hogy az egy elemű $[a_{11}]$ mátrix determinánsán az a_{11} értéket értjük.

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

2.3. Vektorok és mátrixok normája

1.9 Definíció. *Az $f : \mathbb{R}^n \to \mathbb{R}$ függvényt vektornormának nevezzük, ha*

$$\begin{aligned} f(x) &\geq 0 \quad (\forall x \in \mathbb{R}^n), \quad f(x) = 0 \Leftrightarrow x = 0, && (2.2)\\ f(\lambda x) &= |\lambda| f(x) \quad (\forall x \in \mathbb{R}^n,\ \forall \lambda \in \mathbb{R}), && (2.3)\\ f(x+y) &\leq f(x) + f(y) \quad (\forall x, y \in \mathbb{R}^n). && (2.4) \end{aligned}$$

A vektornorma szokásos jelölése: $\|x\|$. A fontosabb vektornormák a következők:

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^{n} |x_i|, && (2.5)\\ \|x\|_2 &= \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} \quad \text{(euklideszi norma)}, && (2.6)\\ \|x\|_\infty &= \max_{1 \leq i \leq n} |x_i| \quad \text{(maximum norma)}. && (2.7) \end{aligned}$$

1.10 Definíció. *Az $x, y \in \mathbb{R}^n$ $(x, y \neq 0)$ vektorok szöge θ, amelynek koszinuszát a*

$$\cos(\theta) = \frac{x^T y}{\|x\|_2 \|y\|_2}$$

összefüggés definiálja.

1.11 Definíció. *Az $f : \mathbb{R}^{n \times n} \to \mathbb{R}$ függvényt mátrixnormának nevezzük, ha*

$$\begin{aligned} f(A) &\geq 0 \quad \left(\forall A \in \mathbb{R}^{n \times n}\right), \quad f(A) = 0 \Leftrightarrow A = 0, && (2.8)\\ f(\lambda A) &= |\lambda| f(A) \quad \left(\forall A \in \mathbb{R}^{n \times n},\ \forall \lambda \in \mathbb{R}\right), && (2.9)\\ f(A+B) &\leq f(A) + f(B) \quad \left(\forall A, B \in \mathbb{R}^{n \times n}\right). && (2.10) \end{aligned}$$

A mátrixnorma szokásos jelölése: $\|A\|$. A leggyakrabban használt mátrixnormák a következők:

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}| \quad \text{(oszlopösszeg norma)}, \tag{2.11}$$

$$\|A\|_2 = \{A^T A \text{ legnagyobb sajátértéke}\}^{\frac{1}{2}} \quad \text{(spektrálnorma)}, \tag{2.12}$$

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| \quad \text{(sorösszeg norma)}, \tag{2.13}$$

$$\|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 \right)^{\frac{1}{2}} \quad \text{(Frobenius norma)}. \tag{2.14}$$

A mátrixnormáknak két fontos osztályát emeljük ki.

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

$$\|A\|_M = \max \{ \|Ax\|_V : \|x\|_V = 1 \}. \tag{2.15}$$

Az indukált mátrixnorma jelentése: az egységnormájú x vektorok megnyújtásának (Ax) maximális mértéke. Másképpen fogalmazva az indukált mátrixnorma az egységgömb képének (ellipszoidnak) az origótól vett legtávolabbi pontjába mutató helyvektor normája.

Összehasonlításként megjegyezzük, hogy a determináns abszolút értéke az ellipszoid és az egységgömb térfogatának hányadosával egyenlő. Könnyen igazolható, hogy $\|A\|_1$ az $\|x\|_1$, $\|A\|_2$ az $\|x\|_2$, $\|A\|_\infty$ pedig az $\|x\|_\infty$ vektornorma által indukált mátrixnorma.

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

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

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

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

Bizonyítás. Először igazoljuk, hogy indukált mátrixnormában

$$\|Ax\| \leq \|A\| \|x\| \quad (x \in \mathbb{R}^n).$$

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

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

ahonnan

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

és a tétel állítása következik. □

Megjegyezzük, hogy az állítás nem minden mátrixnormára igaz (lásd 1.6 Feladatokban a 2. számút). Az indukált normákhoz hasonló (de velük nem azonos) fogalom a következő.

1.13 Definíció. *A* $\|.\|_M : \mathbb{R}^{n\times n} \to \mathbb{R}$ *mátrixnorma kompatibilis a* $\|.\|_V : \mathbb{R}^n \to \mathbb{R}$ *vektornormával, ha* $\|Ax\|_V \leq \|A\|_M \|x\|_V$.

Az $\|A\|_F$ Frobenius norma kompatibilis a $\|x\|_2$ vektornormával. Az 1.12 Definíció alapján világos, hogy az indukált mátrixnormák az őket indukáló vektornormákkal kompatibilisek.

2.4. Mátrixok és vektorok a MATLAB nyelvben

A MATLAB matematikai szoftver a nevét a MATrix LABoratory kifejezésből kapta. A neve is arra utal, hogy benne a mátrixszámítások rendkívül egyszerűen programozhatók. Ennek megfelelően a MATLAB legfontosabb adattípusa a valós vagy komplex elemű mátrix. Mindazonáltal a skalárokra és vektorokra nem kell (bár lehet) két indexszel hivatkozni, továbbá az újabb verziókban a többindexes tömbök és egyéb objektumok is megjelentek.

A mátrixokat legegyszerűbben elemeik sorfolytonos felsorolásával adhatjuk meg. A sorokat pontosvesszővel vagy *enter*-rel választjuk el, ugyanazon mátrixsorban álló elemeket pedig helyközzel vagy vesszővel. Az egész felsorolást szögletes zárójelbe tesszük. Pl. az

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

mátrixot megadhatjuk a következőképpen:

```
≫ A=[1 0 3;3 1 2
      4,5,6]
```

Mindjárt az értékadásra is láttunk példát. A mátrixok deklarációja a rájuk, vagy egy elemükre vonatkozó első értékadó utasítással automatikusan megtörténik. Azonosítót a szokásos módon választhatunk. Vannak foglalt azonosítók, ezek a *HELP* parancs segítségével megismerhetők. Most csak az *eps* foglalt azonosítót említjük ami a konkrét számábrázolástól függő, ún. gépi epszilon (rendszerint $\epsilon \approx 2.2204\times10^{-16}$) értékét jelenti. Nehezen kideríthető hibát okozhat a jelentésétől eltérő használata.

Egy mátrixot kisebb mátrixokból (blokkokból) is felépíthetünk. Pl. a

```
≫ B=[A,[-1 -2;-1-2;-1-2];7,7,7,-7,-7]
```

utasítás az előbbi A-val a

$$B = \begin{bmatrix} 1 & 0 & 3 & -1 & -2 \\ 3 & 1 & 2 & -1 & -2 \\ 4 & 5 & 6 & -1 & -2 \\ 7 & 7 & 7 & -7 & -7 \end{bmatrix}$$

mátrixot adja.

A mátrix valamelyik elemére kerek zárójelek segítségével hivatkozunk: `A(2,3), B(1,4)`. Példáinknál maradva előbbi értéke 2, utóbbié -1. Sor- vagy oszlopvektor (azaz $1 \times n$ vagy $n \times 1$ méretű mátrix) elemére egyetlen indexszel is hivatkozhatunk. Bizonyos esetekben nincs jelentősége annak, hogy egy vektor sor- vagy oszlopvektor-e, sokszor viszont igen. Ha a vektor egyindexes elemeit külön-külön adjuk meg, akkor az sorvektor lesz. Skalár adatot pedig más programnyelvben megismert módon írhatunk, egy legáltalánosabb példa: `-1.23e-2`.

Ha *k,h,m* már definiált számértékek, akkor a vektorok felépítésére egy speciális lehetőség is rendelkezésünkre áll:

$\gg$ `c=k:h:m`

Itt egy sorvektort definiáltunk: c elemei rendre $k, k+h, k+2h, \ldots, k+th$ lesznek. Az utolsó elem a $k+th \leq m < k+(t+1)h$, vagy a $k+(t+1)h < m \leq k+th$ relációkból adódik, attól függően, hogy h pozitív vagy negatív. A h elhagyható, ha értéke 1. A vektor (vagy mátrix) üres is lehet.

Láttuk, hogy egy mátrixot részmátrixok segítségével is megadhatunk. A fordítottját is megtehetjük: egy mátrixból kiemelhetünk egy-egy blokkot (részmátrixot). A következő

$\gg$ `C=A(p:q,r:s)`

utasítás a

$$C = \begin{bmatrix} a_{pr} & \cdots & a_{ps} \\ \vdots & & \vdots \\ a_{qr} & \cdots & a_{qs} \end{bmatrix}$$

mátrixot hozza létre. Ha $p = q$ vagy $r = s$, akkor elég az egyiket megadni. Teljes sort vagy oszlopot pedig elég a : (kettőspont)-tal jelölni.

Fontos alkalmazása a részmátrixoknak az

$\gg$ `A(i,:)=A(i,:)+t*A(k,:)`

utasítás, amely az A mátrix i-edik sorához hozzáadja a k-adik sor t-szeresét.

A részmátrixot még általánosabban is definiálja a MATLAB: ha u egy r elemű, v pedig egy s elemű vektor (mindegy, hogy sor-, vagy oszlopvektor), akkor a

$\gg$ `H=A(u,v)`

utasítás azt az $r \times s$ méretű H mátrixot hozza létre, amelynek elemeire $H(i,j) = A(u_i, v_j)$ teljesül. Pl. egyetlen utasítással felcserélhetjük A két sorát:

$\gg$ `A([j,k],:)=A([k,j],:)`

Megjegyezzük, hogy ha u vagy v elemei nem szigorúan növekedő sorozatot alkotnak, akkor `A(u,v)`-re már nem mondhatjuk, hogy részmátrixa A-nak.

Speciális mátrixokat beépített függvényekkel is létrehozhatunk. Néhány függvényhívás:

$\gg$ `eye(m,n)`, `ones(m,n)`, `zeros(m,n)`, `rand(m,n)`

Ezek a függvények rendre az $m \times n$ méretű egységmátrixot, a csupa 1-esekből álló mátrixot, a zérusmátrixot, illetve véletlenmátrixot adnak. (Akármilyen méretű

egységmátrixot a következőképpen értelmezünk: az azonos indexű elemei 1-esek, a többi eleme 0.) A *RAND* a $(0,1)$-be eső (itt most egyenletes eloszlású) véletlenszámokat állít elő. Ha $m = n$, akkor elég egy argumentumot írni, pl.: `ones(4)`.

Logikailag ide (kis erőltetéssel akár a részmátrixokhoz is) tartozik további három függvény. A *TRIL(A)* alsó háromszög-, a *TRIU(A)* pedig felső háromszögmátrixot képez az A mátrix megfelelő pozíciójában lévő elemeiből. (A háromszögmátrixok definícióját a következő fejezetben adjuk meg.) A *DIAG* tárgyalása előtt meg kell említeni, hogy a mátrixok diagonálisaira sorszámukkal is hivatkozhatunk. Az $m \times n$-es A mátrix fődiagonálisának sorszáma 0, a fölötte lévőké rendre $1, \ldots, (n-1)$, az alatta lévőké pedig $-1, \ldots, -(m-1)$. Precízebben szólva a k-adik diagonálist azon a_{ij} elemek alkotják, amelyekre $k = j - i$. Legyen v egy t hosszúságú vektor. A *DIAG(v,k)* azt a $D \in \mathbb{R}^{(t+|k|)\times(t+|k|)}$ mátrixot hozza létre, melynek k-adik diagonálisa a v elemeit tartalmazza, többi eleme pedig 0. *DIAG(A,k)* pedig azt a h oszlopvektort adja, amelynek elemei az A mátrix k-adik diagonálisában találhatók. Ha $k = 0$, akkor feltüntetése elmaradhat.

Mátrixok, vektorok, skalárok között a $+, -, *$ jelekkel az algebrában megismert műveleteket végezhetjük el és a várt eredményt kapjuk. Van egy kivétel: skalárt bármilyen méretű mátrixhoz hozzáadhatunk (kivonhatunk). Ilyenkor a skalár a mátrix minden eleméhez hozzáadódik. A `^` a hatványozás jele. Az `A^k` művelet k-szor összeszorozza A-t, ha k pozitív egész, illetve $k = -1$ esetén az inverzet adja, ha az létezik.
eredménye az $Ax = b$ lineáris egyenletrendszer megoldása (ha az egyenletrendszer konzisztens).

2.5. Feladatok

1. Melyik mátrix szimmetrikus bármely $A, B \in \mathbb{R}^{m\times n}$ mellett a következők közül?
   (i) $A^TA + B^TB$,
   (ii) $(A^TA)(B^TB)$
2. Igazoljuk, hogy $\|A\|_\Delta = \max_{1\le i,j\le n} |a_{ij}|$ mátrixnorma! Mutassuk meg továbbá, hogy az $A = B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ mátrixokra $\|AB\|_\Delta > \|A\|_\Delta \|B\|_\Delta$!
3. Igazoljuk a Strassen és Winograd algoritmusok helyességét!

3. fejezet

LINEÁRIS EGYENLETRENDSZEREK MEGOLDÁSA

3.1. Lineáris egyenletrendszerek

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

$$\begin{array}{rcl} a_{11}x_1 + \ldots + a_{1j}x_j + \ldots + a_{1n}x_n & = & b_1 \\ & \vdots & \\ a_{i1}x_1 + \ldots + a_{ij}x_j + \ldots + a_{in}x_n & = & b_i \\ & \vdots & \\ a_{m1}x_1 + \ldots + a_{mj}x_j + \ldots + a_{mn}x_n & = & b_m \end{array} \tag{3.1}$$

Az egyenletrendszert megadhatjuk a tömörebb

$$Ax = b \tag{3.2}$$

formában, ahol

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

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

2.1 Definíció. *Az $\mathbb{R}^n$ euklideszi tér d ($d \in \mathbb{R}^n$) normálvektorú és $x_0 \in \mathbb{R}^n$ ponton átmenő hipersíkját az*

$$(x - x_0)^T d = 0 \tag{3.3}$$

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

A hipersík egyenlete más formában kifejezve:

$$x^T d = x_0^T d. \tag{3.4}$$

Felhasználva az A mátrix sorok szerinti

$$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_i^T \\ \vdots \\ a_m^T \end{bmatrix}$$

felbontását, ahol $a_i^T = [a_{i1}, \ldots, a_{in}]$, az $Ax = b$ egyenletrendszert felírhatjuk az ekvivalens

$$\begin{aligned} a_1^T x &= b_1 \\ &\vdots \\ a_m^T x &= b_m \end{aligned} \tag{3.5}$$

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

(i) az egyenletrendszernek nincs megoldása,

(ii) az egyenletrendszernek pontosan egy megoldása van,

(iii) az egyenletrendszernek végtelen sok megoldása van.

2.2 Definíció. *Ha az $Ax = b$ lineáris egyenletrendszernek legalább egy megoldása van, akkor az egyenletrendszert konzisztensnek nevezzük. Ha az egyenletrendszernek nincs megoldása, akkor az egyenletrendszert inkonzisztensnek nevezzük.*

Például az $x + 2y = 1$, $x + 2y = 4$ egyenletrendszer inkonzisztens.

Az $Ax = b$ egyenletrendszert felírhatjuk az ekvivalens

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

alakban is, ahol a_i az A mátrix i-edik oszlopát jelöli. A $\sum_{i=1}^n x_i a_i$ összeget az $\{a_i\}_{i=1}^n$ vektorok *lineáris kombinációjának* nevezzük. Az egyenletrendszer akkor és csak akkor oldható meg, ha b kifejezhető az A oszlopvektorainak lineáris kombinációjaként.

2.3 Definíció. *Az $\{a_i\}_{i=1}^k \subseteq \mathbb{R}^m$ vektorok lineárisan összefüggők, ha létezik $x \in \mathbb{R}^k$ ($x \neq 0$), hogy*

$$\sum_{i=1}^{k} x_i a_i = 0. \tag{3.6}$$

Ha nincs ilyen $x \neq 0$ vektor, akkor az $\{a_i\}_{i=1}^k$ vektorokat lineárisan függetlennek nevezzük.

A megoldhatóság egy "másik jellemzését" adhatjuk a *rang* fogalmával:

$$\text{rank}(A) = \begin{array}{c}\text{lineárisan független oszlop- vagy sorvektorok} \\ \text{maximális száma}\end{array} \tag{3.7}$$

A mátrix rangjával megfogalmazva az $Ax = b$ egyenletrendszer akkor és csak akkor megoldható, ha $\text{rank}(A) = \text{rank}([A, b])$. Ha $\text{rank}(A) = \text{rank}([A, b]) = n$, akkor az $Ax = b$ egyenletrendszernek pontosan egy megoldása van.

A továbbiakban csak négyzetes egyenletrendszerekkel foglalkozunk. Feltesszük tehát, hogy $m = n$. Ismert a következő

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

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

3.2. Háromszögmátrixú egyenletrendszerek

2.4 Definíció. *Az $A = [a_{ij}]_{i,j=1}^n$ mátrix alsó háromszög alakú, ha minden $i < j$ esetén $a_{ij} = 0$.*

Az alsó háromszögmátrixok alakja sematikusan a következő

$$\begin{bmatrix} * & 0 & \dots & \dots & 0 \\ * & * & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & * & 0 \\ * & \dots & \dots & * & * \end{bmatrix}.$$

2.5 Definíció. *Az $A = [a_{ij}]_{i,j=1}^n$ mátrix felső háromszög alakú, ha minden $i > j$ esetén $a_{ij} = 0$.*

A felső háromszögmátrixok alakja:

$$\begin{bmatrix} * & * & \dots & \dots & * \\ 0 & * & & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & * & * \\ 0 & \dots & \dots & 0 & * \end{bmatrix}.$$

Megjegyzés. *Egyidejűleg alsó és felső háromszögmátrixok a diagonálmátrixok (köztük a zérusmátrix).*

Igazolható, hogy alsó vagy felső háromszögmátrixok esetén $\det(A) = a_{11}a_{22}\dots a_{nn}$. A háromszögmátrixú egyenletrendszerek megoldása igen egyszerű. Tekintsük az

$$\begin{array}{ccccccc} a_{11}x_1 & & & & & = & b_1 \\ \vdots & \ddots & & & & & \vdots \\ a_{i1}x_1+ & \dots & +a_{ii}x_i & & & = & b_i \\ \vdots & & \vdots & \ddots & & & \vdots \\ a_{n1}x_1+ & \dots & +a_{ni}x_i & \dots & +a_{nn}x_n & = & b_n \end{array}$$

alsó háromszögmátrixú egyenletrendszert! Az egyenletrendszer akkor és csak akkor oldható meg egyértelműen, ha $a_{11} \neq 0, \dots, a_{nn} \neq 0$. Az alsó háromszögmátrixú egyenletrendszer megoldását adja a következő algoritmus:

$$\begin{array}{l} x_1 = b_1/a_{11} \\ \textbf{for } i = 2 : n \\ \quad x_i = (b_i - \sum_{j=1}^{i-1} a_{ij}x_j)/a_{ii} \\ \textbf{end} \end{array} \tag{3.8}$$

Tekintsük most az

$$\begin{array}{ccccccc} a_{11}x_1+ & \dots & +a_{1i}x_i+ & \dots & +a_{1n}x_n & = & b_1 \\ \ddots & \vdots & & & \vdots & & \vdots \\ & a_{ii}x_i+ & \dots & & +a_{in}x_n & = & b_i \\ & & \ddots & & \vdots & & \vdots \\ & & & & a_{nn}x_n & = & b_n \end{array}$$

felső háromszögmátrixú egyenletrendszert! Az egyenletrendszer akkor és csak akkor oldható meg egyértelműen, ha $a_{11} \neq 0, \dots, a_{nn} \neq 0$. A felső háromszögmátrixú egyenletrendszer megoldását a következő, ún. *visszahelyettesítő* algoritmus adja:

$$\begin{array}{l} x_n = b_n/a_{nn} \\ \textbf{for } i = n-1 : -1 : 1 \\ \quad x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii} \\ \textbf{end} \end{array} \tag{3.9}$$

3.3. A Gauss-módszer

A Gauss-féle eliminációs módszer két fázisból áll:

I. Azonos átalakításokkal az $Ax = b$ egyenletrendszert felső háromszög alakúra hozzuk:

II. A kapott felső háromszögmátrixú egyenletrendszert a (3.9) Algoritmussal megoldjuk.

A felső háromszög alakra hozás a következőképpen történik.

Ha $a_{11} \neq 0$, akkor az a_{11} alatti x_1 együtthatókat nullává tesszük (kinullázzuk) úgy, hogy az i-edik sorból kivonjuk $(i = 2, \ldots, n)$ az első sor γ-szorosát:

$$(a_{i1} - \gamma a_{11})x_1 + (a_{i2} - \gamma a_{12})x_2 + \ldots + (a_{in} - \gamma a_{1n})x_n = b_i - \gamma b_1. \tag{3.10}$$

Az $a_{i1} - \gamma a_{11} = 0$ feltételből kapjuk, hogy $\gamma = \frac{a_{i1}}{a_{11}}$. Így az első oszlop a_{11} alatti kinullázását a

$$\left.\begin{array}{l} \gamma = a_{i1}/a_{11} \\ a_{ij} = a_{ij} - \gamma a_{1j} \quad (j = 2, \ldots, n) \\ b_i = b_i - \gamma b_1 \end{array}\right\} \quad (i = 2, \ldots, n) \tag{3.11}$$

algoritmussal végezhetjük el. Vegyük észre, hogy az algoritmus felülírja az A mátrix $2 \leq i, j \leq n$ indexű és a b vektor $2 \leq i \leq n$ indexű elemeit (a 0-kat viszont feleslegesen nem írja be az alsó háromszög részbe). A felülírt elemeknél is megtartva az eredeti jelölést, a kapott ekvivalens egyenletrendszer alakja:

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & b_1 \\ & & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = & b_2 \\ & & \vdots & & & & \vdots & & \vdots \\ & & a_{n2}x_2 & + & \ldots & + & a_{nn}x_n & = & b_n \end{array} . \tag{3.12}$$

Ezt szétbonthatjuk az n ismeretlent tartalmazó első egyenletre és az $n-1$ ismeretlent tartalmazó kisebb $(n-1) \times (n-1)$-es egyenletrendszerre. Ha $a_{22} \neq 0$, akkor a kisebb egyenletrendszeren megismételjük az előző lépést és így tovább. Tegyük fel, hogy a $(k-1)$-edik oszlopban a kinullázást már elvégeztük és az

$$\begin{array}{ccccccccc} a_{11}x_1 + & \ldots & \ldots & +a_{1k}x_k & + & \ldots & + & a_{1n}x_n & = & b_1 \\ & \ddots & & \vdots & & & & \vdots & & \vdots \\ & & \ddots & \vdots & & & & \vdots & & \vdots \\ & & & a_{kk}x_k & + & \ldots & + & a_{kn}x_n & = & b_k \\ & & & \vdots & & & & \vdots & & \vdots \\ & & & a_{ik}x_k & + & \ldots & + & a_{in}x_n & = & b_i \\ & & & \vdots & & & & \vdots & & \vdots \\ & & & a_{nk}x_k & + & \ldots & + & a_{nn}x_n & = & b_n \end{array}$$

egyenletrendszert kaptuk. Ha $a_{kk} \neq 0$, akkor kinullázzuk az a_{kk} alatti x_k együtthatókat. Az i-edik sorból a k-adik sort γ-szorosát kivonva az

$$(a_{ik} - \gamma a_{kk})x_k + (a_{i,k+1} - \gamma a_{k,k+1})x_{k+1} + \ldots + (a_{in} - \gamma a_{kn})x_n = b_i - \gamma b_k \tag{3.13}$$

egyenlet adódik. Az $a_{ik} - \gamma a_{kk} = 0$ feltételből kapjuk, hogy $\gamma = \frac{a_{ik}}{a_{kk}}$. A k-adik oszlop a_{kk} alatti kinullázását tehát a

$$\left.\begin{array}{l} \gamma = a_{ik}/a_{kk} \\ a_{ij} = a_{ij} - \gamma a_{kj} \quad (j = k+1, \ldots, n) \\ b_i = b_i - \gamma b_k \end{array}\right\} \quad (i = k+1, \ldots, n)$$

algoritmussal végezhetjük el.

A kinullázást mindaddig folytathatjuk, amíg az $a_{kk} \neq 0$ és $k \leq n-1$ feltételek fennállnak. Ha sikerül az A mátrixot felső háromszög alakra hozni, akkor a (3.9) Algoritmust alkalmazzuk (visszahelyettesítünk). A következőkben $A(i,j)$ az A mátrix a_{ij} elemét jelöli.

A GAUSS-MÓDSZER:
I. (eliminációs) fázis:
for $k = 1 : n-1$
for $i = k+1 : n$
$\gamma = A(i,k)/A(k,k)$
$A(i, k+1:n) = A(i, k+1:n) - \gamma * A(k, k+1:n)$
$b_i = b_i - \gamma b_k$
end
end
II. (visszahelyettesítő) fázis:
$x_n = b_n/a_{nn}$
for $i = n-1 : -1 : 1$
$x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii}$
end

3.4. A Gauss-módszer műveletigénye

A Gauss-módszer véges sok lépésben, véges sok aritmetikai alapművelet ($+, -, *, /$) elvégzése után megadja az $Ax = b$ ($A \in \mathbb{R}^{n\times n}$) egyenletrendszer megoldását. A szükséges aritmetikai műveletszám (műveletigény) az egyenletrendszer megoldó eljárások fontos minőségi jellemzője, mert az ilyen algoritmusok számítógépideje nagyjából arányos az aritmetikai műveletigénnyel. A Gauss-módszer műveletigényét az alábbiak szerint tudjuk meghatározni. A szorzás és az osztás nagyjából azonos időigényűek. Ezért ezeket összevonva, azonos multiplikatív műveletként (jele M) számoljuk. A kivonás és összeadás műveleteknél hasonló a helyzet. Itt a közös alapegység az additív művelet (jele A).

Az I. fázis k-adik lépésében a műveletek és műveletszámok a következők:

for $i = k+1 : n$		
$\gamma = A(i,k)/A(k,k)$	$\Rightarrow$	M
$A(i, k+1:n) = A(i, k+1:n) - \gamma * A(k, k+1:n)$	$\Rightarrow$	$(n-k)(M+A)$
$b_i = b_i - \gamma b_k$	$\Rightarrow$	$M+A$
end		

A ciklus műveletigénye összesen

$$(n-k)\left(M + (n-k)(M+A) + M + A\right),$$

azaz

$$\left((n-k)^2 + 2(n-k)\right)M + \left((n-k)^2 + (n-k)\right)A.$$

Ezt összegezve a $k = 1, \ldots, n-1$ lépésekre kapjuk, hogy az I. fázis műveletigénye

$$\sum_{i=1}^{n-1}(i^2+2i)M + \sum_{i=1}^{n-1}(i^2+i)A.$$

Felhasználva, hogy

$$\sum_{i=1}^{k} i^2 = \frac{k(k+1)(2k+1)}{6}$$

kapjuk, hogy

$$\left(\frac{(n-1)n(2n-1)}{6} + (n-1)n\right)M + \left(\frac{(n-1)n(2n-1)}{6} + \frac{(n-1)n}{2}\right)A,$$

illetve

$$\left(\frac{n^3}{3} + \frac{n^2}{2} - \frac{5}{6}n\right)M + \left(\frac{n^3}{3} - \frac{n}{3}\right)A.$$

A II. fázis műveletigénye:

$x_n = b_n/a_{nn}$ $\Rightarrow$ M
for $i = n-1 : -1 : 1$
  $x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii}$ $\Rightarrow$ $(n-i+1)M + (n-i)A$
end

Összegezve az $i = n-1 : -1 : 1$ lépések műveletigényét kapjuk, hogy a II. fázis összköltsége

$$M + \sum_{j=2}^{n} jM + \sum_{j=1}^{n-1} jA = \left(\frac{n^2}{2} + \frac{n}{2}\right)M + (\frac{n^2}{2} - \frac{n}{2})A.$$

Az I. és II. fázis költségét összeadva kapjuk a Gauss-módszer számítási összköltségét:

$$\left(\frac{n^3}{3} + n^2 - \frac{n}{3}\right)M + \left(\frac{n^3}{3} + \frac{n^2}{2} - \frac{5}{6}n\right)A.$$

Nagy n értékekre az $\frac{n^3}{3}$ együttható válik dominánssá mindkét zárójeles kifejezésben.

2.3a Tétel. *A Gauss-módszer műveletigénye* $\frac{n^3}{3} + O(n^2)$ *multiplikatív és ugyanennyi* $\frac{n^3}{3} + O(n^2)$ *additív művelet.*

Véges befejezésű aritmetikai műveleteket használó algoritmusok esetén az algoritmusok lényeges minőségi jellemzője az algoritmus végrehajtásához szükséges aritmetikai műveletek száma a feladatok paramétereinek (pl. ismeretlenek száma,

együttható mátrix mérete, stb.) függvényében. A régebbi számítógépek esetén a multiplikatív műveletek végrehajtási ideje lényegesen nagyobb volt, mint az additívaké. Ezért ezeket külön határozták meg. Később megfigyelték, hogy a lineáris algebra számítási eljárásaiban az additív és a multiplikatív műveletek száma nagyon gyakran közel azonos. Ezért C.B. Moler a számítási igény mérésére bevezette az *1 flop* fogalmát.

2.6a Definíció. *1 (régi) flop az a számítási munka, amely az* $s = s + x * y$ *művelet (1 összeadás + 1 szorzás) elvégzéséhez kell.*

Például az $x^T y$ skalárszorzat ($x, y \in \mathbb{R}^n$) kiszámítása az

$$s = 0, \quad s = s + x_i y_i \quad (i = 1, \ldots, n)$$

algoritmussal n flop műveletigényű.

Az újabb számítógépeken a multiplikatív és az additív műveletek ideje azonosnak tekinthető. Ezért a régi flop fogalmát a következőképpen módosították.

2.6b Definíció. *1 (új) flop az a számítási munka, amely egy* $+, -, *, /$ *aritmetikai művelet elvégzéséhez kell.*

Tehát egy régi flop 2 új floppal azonos a mai számítógépeken. Az algoritmusok flop igényét a MATLAB rendszer régi változatai mérték. Az új MATLAB változatok azonban a fent említett okok miatt már nem. Ugyanakkor a flop fogalmának erős elterjedtsége miatt a könyvben még a régi flop értelmezést használjuk.

2.3b Tétel. *A Gauss-módszer műveletigénye* $\frac{n^3}{3} + O(n^2)$ *régi flop.*

Klyuyev és Kokovkin-Shcherbak igazolta, hogy ha csak sor- és oszlopműveleteket (sor, vagy oszlop számmal való szorzása; sorok, vagy oszlopok cseréje; sorok, vagy oszlopok számszorosának sorokhoz, vagy oszlopokhoz való hozzáadása) engedünk meg, akkor nem lehet $\frac{n^3}{3} + O(n^2)$ régi flopnál kevesebb művelettel az $Ax = b$ lineáris egyenletrendszert megoldani.

Az $Ax = b$ alakú $n \times n$-es egyenletrendszerek megoldásához szükséges műveletigény gyors mátrixinvertáló eljárásokkal $O(n^{2.808})$ régi flopra leszorítható. A jelenleg ismert eljárásokat numerikus instabilitásuk miatt gyakorlatilag nem használják.

3.5. A főelemkiválasztásos Gauss-módszer

A Gauss-módszer I. fázisában előfordulhat, mondjuk a k-adik lépésben, hogy az a_{kk} elem zérus. Például a

$$\begin{array}{rcrcrcr} & & 4x_2 & + & x_3 & = & 9 \\ x_1 & + & x_2 & + & 3x_3 & = & 6 \\ 2x_1 & - & 2x_2 & + & x_3 & = & -1 \end{array}$$

rendszernél $a_{11} = 0$. Ilyen esetekben a sorok, vagy az oszlopok felcserélésével megkísérelhetjük elérni, hogy az a_{kk} helyére zérustól különböző elem kerüljön. A fenti esetben például az első és harmadik sor felcserélésével kapjuk, hogy

$$\begin{array}{rcrcrcr} 2x_1 & - & 2x_2 & + & x_3 & = & -1 \\ x_1 & + & x_2 & + & 3x_3 & = & 6 \\ & & 4x_2 & + & x_3 & = & 9 \end{array}$$

Az első és második oszlop oszlop cseréjével pedig azt kapjuk, hogy

$$\begin{array}{rcrcrcr} 4x_2 & & & + & x_3 & = & 9 \\ x_2 & + & x_1 & + & 3x_3 & = & 6 \\ 2x_2 & - & 2x_1 & + & x_3 & = & -1 \end{array}$$

A sorok cseréjénél az egyenletek (és b megfelelő komponenseinek) sorrendje, az oszlopok cseréjénél pedig a változók sorrendje változik meg. Általában, így az előző példában is, több választási lehetőségünk is van sor-, vagy oszlopcserére. Ha azonban az a_{kk} elem alatt minden együttható zérus, akkor az $[a_{ij}]_{i,j=1}^{n,k}$ részmátrix oszlopai lineárisan összefüggők, A szinguláris és az eliminációs eljárás sorcserével sem folytatható. Hasonló a helyzet, ha a_{kk} sorában, tőle jobbra, minden együttható zérus, mert ekkor A ismét szinguláris.

Az a_{kk} elemet k-adik *pivot, vagy főelemnek* nevezzük. A sorok felcserélését úgy, hogy az új pivot elem zérustól különböző legyen, *pivotálási*, vagy *főelemkiválasztási eljárásnak* nevezzük. A pivot elem megválasztása nagymértékben befolyásolja az eredmények megbízhatóságát. Példaként említjük a következő egyenletrendszert:

$$\begin{array}{rcrcr} 10^{-17}x & + & y & = & 1 \\ x & + & y & = & 2 \end{array}$$

Ha ezt a pivotálás nélküli Gauss-módszerrel számítógépen megoldjuk, akkor (15 tizedesjegy pontosságú MATLAB számítások esetén) az $x = 0$, $y = 1$ közelítő eredményt kapjuk. A helyes eredmény: $x = \frac{1}{1-10^{-17}}$, $y = 1 - \frac{10^{-17}}{1-10^{-17}}$. Az első és a második egyenlet felcserélésével kapott

$$\begin{array}{rcrcr} x & + & y & = & 2 \\ 10^{-17}x & + & y & = & 1 \end{array}$$

egyenletrendszeren ugyanaz a módszer az $x = 1$, $y = 1$ közelítő megoldást adja. Ez utóbbi közel van a pontos megoldáshoz, míg az első eredmény katasztrófálisan eltér.

Általában is igaz, hogy a közelítő megoldás pontosságát nagymértékben javítja a helyesen megválasztott pivotálás. Pivot elemnek nagy abszolút értékű elemet kell választani. Két alapvető pivotálási stratégiát használunk.

Részleges főelemkiválasztás: A k-adik lépésben a k-adik oszlop a_{jk} $(k \leq j \leq n)$ elemei közül kiválasztjuk a maximális abszolút értékűt. Ha ennek indexe i, akkor a k-adik és az i-edik sort felcseréljük. A pivotálás után teljesül, hogy

$$|a_{kk}| = \max_{k \leq i \leq n} |a_{ik}| .$$

Teljes főelemkiválasztás: A k-adik lépésben az a_{ij} $(k \leq i, j \leq n)$ mátrixelemek közül kiválasztjuk a maximális abszolút értékűt. Ha ennek indexe (i, j), akkor a k-adik és az i-edik sort, valamint a k-adik oszlopot és j-edik oszlopot felcseréljük. A pivotálás után teljesül, hogy

$$|a_{kk}| = \max_{k \leq i,j \leq n} |a_{ij}| .$$

Megjegyezzük, hogy oszlopcsere esetén változócsere is történik.

A főelemkiválasztásos Gauss-módszer esetén az I. fázis minden lépésében pivotálást hajtunk végre. A teljes főelemkiválasztást biztonságos stratégiának tekinthetjük. A részleges főelemkiválasztás egyéb technikákkal kiegészítve ugyancsak biztonságosnak tekinthető. A részleges főelemkiválasztás lényegesen kevesebb extra aritmetikai műveletet igényel mint a teljes főelemkiválasztás. Ezért a gyakorlatban inkább ezt használjuk. Az I. fázis alakja algoritmikus formában

for $k = 1 : n - 1$
  Határozzuk meg a t indexet, hogy $|A(t,k)| = \max_{k \leq i \leq n} |A(i,k)|$.
  if $k \neq t$
    Cseréljük fel a k-adik és t-edik sort!
  end
  for $i = k + 1 : n$
    $\gamma = A(i,k) / A(k,k)$
    $A(i, k+1:n) = A(i, k+1:n) - \gamma * A(k, k+1:n)$
    $b_i = b_i - \gamma b_k$
  end
end

Nem kell végrehajtani főelemkiválasztást a következő esetekben:

1. A szimmetrikus és pozitív definit ($A \in \mathbb{R}^{n \times n}$ pozitív definit$\Leftrightarrow$ $x^T A x > 0$, $\forall x \in \mathbb{R}^n$, $x \neq 0$).

2. A diagonálisan domináns a következő értelemben:

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad (1 \leq i \leq n) .$$

Szimmetrikus és pozitív definit A mátrix esetén a Gauss-eliminiáció egy később bemutatásra kerülő speciális alakját, a Cholesky-módszert használjuk az egyenletrendszer megoldására.

A Gauss-elimináció további vizsgálatához szükségünk van az eljárás következő leírására: Legyen $A^{(0)} = A$ és $b^{(0)} = b$. A Gauss-elimináció első fázisában egy

$$A^{(0)}x = b^{(0)} \to A^{(1)}x = b^{(1)} \to \ldots \to A^{(n-1)}x = b^{(n-1)}$$

ekvivalens egyenletrendszerekből álló sorozatot képezünk, ahol

$$A^{(k)} = \left[a_{ij}^{(k)}\right]_{i,j=1}^{n}.$$

Fennáll, hogy

$$A^{(n-1)} = \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \ldots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & \ldots & a_{2n}^{(1)} \\ \vdots & & \ddots & \vdots \\ 0 & \ldots & \ldots & a_{nn}^{(n-1)} \end{bmatrix},$$

ahol $a_{kk}^{(k-1)}$ a k-adik fő-, vagy pivotelem. A pivotelemek növekedési tényezője:

$$\rho = \rho_n = \max_{1 \le k \le n} \left| a_{kk}^{(k-1)} / a_{11}^{(0)} \right|.$$

Igen fontos kérdés a ρ növekedési tényező nagyságrendje, mert ez összefüggésben áll az eljárás numerikus stabilitásával. Wilkinson igazolta, hogy a közelítő megoldás hibája arányos a ρ növekedési tényezővel, amelyre teljes főelemkiválasztás esetén a

$$\rho \le \sqrt{n}\left(2 \cdot 3^{\frac{1}{2}} \cdots n^{\frac{1}{n-1}}\right)^{\frac{1}{2}} \sim c n^{\frac{1}{2}} n^{\frac{1}{4} log(n)},$$

részleges főelemkiválasztás esetén pedig a

$$\rho \le 2^{n-1}$$

korlát teljesül. Wilkinson azt sejtette, hogy teljes főelemkiválasztás esetén $\rho \le n$. Ezt kis n értékekre többen is igazolták. Véletlen mátrixokon végzett statisztikai vizsgálatok ($n \le 1024$) azt mutatják, hogy ρ nagyságrendje átlagosan $O\left(n^{2/3}\right)$ a részleges és $O\left(n^{1/2}\right)$ a teljes főelemkiválasztás esetén. Tehát a $\rho > n$ eset statisztikai értelemben ritkán fordulhat elő.

A részleges főelemkiválasztás $\rho = 2^{n-1}$ korlátja pontos és teljesül például a

Wilkinson által konstruált $n \times n$-es

$$A = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 1 \\ -1 & 1 & 0 & \dots & 0 & 1 \\ -1 & -1 & 1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & 1 & 0 & 1 \\ -1 & -1 & \dots & -1 & 1 & 1 \\ -1 & -1 & \dots & -1 & -1 & 1 \end{bmatrix}$$

mátrix esetén. Ekkor nem kell a pivotálásnál sorcseréket végrehajtani és az utolsó oszlop elemei exponenciálisan növekednek. Újabban találtak néhány, a gyakorlatban is (differenciál- és integrálegyenletek közelítő megoldásánál) előforduló esetet, amikor a részleges főelemkiválasztáson alapuló Gauss-módszer ρ növekedési tényezője exponenciálisan nő és a módszer csődöt mond.

A főelemkiválasztás nélküli Gauss elimináció esetén a növekedési tényező nagyon nagy lehet. Például az

$$A = \begin{bmatrix} 1.7846 & -0.2760 & -0.2760 & -0.2760 \\ -3.3848 & 0.7240 & -0.3492 & -0.2760 \\ -0.2760 & -0.2760 & 1.4311 & -0.2760 \\ -0.2760 & -0.2760 & -0.2760 & 0.7240 \end{bmatrix}$$

mátrix esetén a növekedési tényező $\rho = \rho_4(A) = 1.23 \times 10^5$.

3.6. Az LU-felbontás

A Gauss-módszer, amely egy egyszerű eljárást ad egyenletrendszerek megoldására, az I. fázisban egy felső háromszögmátrixot állít elő. Most azt a kérdést vizsgáljuk meg, hogy mi ez a felső háromszögmátrix és milyen kapcsolatban áll az eredeti A együttható mátrixszal. Szükségünk van a következő észrevételekre és fogalmakra:

(i) Háromszögmátrixok inverze azonos típusú háromszögmátrix;

(ii) Alsó háromszögmátrixok szorzata alsó, felső háromszögmátrixok szorzata pedig felső háromszögmátrix.

2.7 Definíció. *Az $A \in \mathbb{R}^{n\times n}$ mátrix LU-felbontásán a mátrix $A = LU$ szorzatalakban történő felbontását értjük, ahol $L \in \mathbb{R}^{n\times n}$ alsó, $U \in \mathbb{R}^{n\times n}$ pedig felső háromszögmátrix.*

Ha egy nemszinguláris A mátrixnak létezik két LU-felbontása, $A = L_1U_1$ és $A = L_2U_2$, akkor van olyan D diagonális mátrix, hogy $L_1 = L_2D$, $U_1 = D^{-1}U_2$.

Az észrevétel igazolásához tegyük fel, hogy van két LU-felbontásunk. Az $L_1U_1 = L_2U_2$ egyenlőségből $L_2^{-1}L_1 = U_2U_1^{-1}$ következik. A baloldal alsó, a jobboldal pedig felső háromszögmátrix. Ezek csak akkor lehetnek egyenlők, ha diagonálisak, azaz $L_2^{-1}L_1 = U_2U_1^{-1} = D$. Innen az állítás már következik.

Ha L egység alsó háromszögmátrix (azaz főátlójában minden elem 1), akkor az LU felbontás a fenti bizonyítás alapján egyértelmű.

Legyen

$$A^{(r)} = \begin{bmatrix} a_{11} & \dots & a_{1r} \\ \vdots & & \vdots \\ a_{r1} & \dots & a_{rr} \end{bmatrix} \quad (r = 1, \dots, n-1)\,.$$

Az $A^{(r)}$ mátrixot az A mátrix r-edik főminor mátrixának nevezzük.

2.4 Tétel. *Egy $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrixnak akkor és csak akkor létezik LU-felbontása, ha*

$$\det\left(A^{(r)}\right) \neq 0 \quad (r = 1, \dots, n-1). \tag{3.14}$$

Bizonyítás. Feltehetjük, hogy L egység alsó háromszögmátrix. Az $n = 1$ esetben az állítás triviális. Tegyük fel, hogy az állítás $n \leq k$ esetén igaz. Legyen $A \in \mathbb{R}^{k\times k}$ és

$$\widetilde{A} = \begin{bmatrix} A & c \\ b^T & \alpha \end{bmatrix} \in \mathbb{R}^{(k+1)\times(k+1)} \quad \left(b, c \in \mathbb{R}^k, \alpha \in \mathbb{R}\right).$$

Keressük az $\widetilde{A}$ mátrix LU-felbontását az utolsó sor és oszlop szerint particionált formában:

$$\widetilde{A} = \begin{bmatrix} A & c \\ b^T & \alpha \end{bmatrix} = \underbrace{\begin{bmatrix} L & 0 \\ x^T & 1 \end{bmatrix}}_{\widetilde{L}} \underbrace{\begin{bmatrix} U & y \\ 0 & \gamma \end{bmatrix}}_{\widetilde{U}} \quad \left(b, c, x, y \in \mathbb{R}^k,\ \alpha, \gamma \in \mathbb{R}\right),$$

ahol L egység alsó, U pedig felső háromszögmátrix. Ez pontosan akkor állhat fenn, ha

$$A = LU, \quad Ly = c, \quad x^T U = b^T, \quad x^T y + \gamma = \alpha.$$

Ha $\widetilde{A}$-nak van LU-felbontása, akkor szükségképpen A-nak is van. Az $\tilde{L}$ nemszinguláris, mert egység alsó háromszögmátrix. Ha $\widetilde{A}$ nem szinguláris, akkor $\widetilde{U}$ sem lehet szinguláris, mert $\widetilde{U} = \left(\widetilde{L}\right)^{-1} \widetilde{A}$. Ekkor tehát az U nemszinguláris és $\alpha - b^T A^{-1} c \neq 0$. Tehát $A = LU$ szükségképpen nemszinguláris. Az $\widetilde{A}$ mátrix LU-felbontásának alakja pedig

$$\widetilde{A} = \underbrace{\begin{bmatrix} L & 0 \\ b^T U^{-1} & 1 \end{bmatrix}}_{\widetilde{L}} \underbrace{\begin{bmatrix} U & L^{-1}c \\ 0 & \alpha - b^T A^{-1} c \end{bmatrix}}_{\widetilde{U}}.$$

Fordítva, ha A-nak van LU-felbontása, akkor $\widetilde{A}$-nek is van. Az A mátrixnak az indukciós feltevés miatt akkor és csak akkor van LU-felbontása, ha főminor mátrixai nemszingulárisak. □

A pozitív definit mátrixok összes főminor mátrixa is pozitív definit és ezért nemszinguláris. Pozitív definit mátrixoknak tehát van LU-felbontásuk. Ha a mátrix még szimmetrikus is, akkor az LU-felbontás speciális szerkezetű.

Tegyük fel, hogy A szimmetrikus és pozitív definit. Legyen az $A = LU$ felbontásban L egység alsó háromszögmátrix. Minthogy $A^T = A$, fennáll $U^T L^T = LU$ is. Ezt balról szorozva az L^{-1}, jobbról pedig az L^{-T} mátrixszal az

$$L^{-1}U^T = UL^{-T}$$

egyenlőséget kapjuk. A baloldal alsó, a jobboldal pedig felső háromszögmátrix. Ezért egyenlőség csak akkor lehetséges, ha létezik egy D diagonális mátrix, hogy

$$L^{-1}U^T = UL^{-T} = D.$$

Innen az $U = DL^T$ összefüggés azonnal adódik. Kaptuk tehát, hogy szimmetrikus pozitív definit mátrixoknak létezik az

$$A = LDL^T \tag{3.15}$$

felbontásuk, ahol L egység alsó háromszögmátrix. A $D = \mathrm{diag}(d_i)$ mátrix minden diagonális eleme pozitív. Legyen ui. $x = L^{-T}e_i$. Minthogy $x^T Ax > 0$, az $x^T Ax = e_i^T L^{-1} LDL^T L^{-T} e_i = d_i > 0$ is igaz. Legyen $\widehat{D} = \mathrm{diag}\left(\sqrt{d_i}\right)$ és $\widehat{L} = L\widehat{D}$. Akkor fennáll, hogy

$$A = \widehat{L}\widehat{L}^T, \tag{3.16}$$

ahol $\widehat{L}$ alsó háromszögmátrix, amelynek főátlójában pozitív számok vannak. Ezt a felbontást a szimmetrikus pozitív definit A mátrix *Cholesky*-féle felbontásának nevezzük.

Vannak esetek, amikor egy mátrix nemszinguláris és mégsincs LU-felbontása.
Példa. A

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

nemszinguláris mátrixnak nincs LU-felbontása.
2.8 Definíció. *A P $n \times n$ mátrix permutációmátrix, ha minden sorában és oszlopában pontosan egy darab* 1*-es van és a többi elem zérus.*
Példa.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Legyen $e_i = [0, \ldots, 0, 1, 0, \ldots, 0]^T \in \mathbb{R}^n$ az i-edik egységvektor. Egy permutációmátrixot felírhatunk a következő formában is

$$P = \begin{bmatrix} e_{i_1}^T \\ e_{i_2}^T \\ \vdots \\ e_{i_n}^T \end{bmatrix},$$

ahol $i_1, i_2, \ldots, i_n$ az $1, 2, \ldots, n$ számok valamelyik permutációja. A fenti példa esetében $P = [e_1, e_3, e_4, e_2]^T$. Vizsgáljuk most meg a permutációmátrixszal való szorzás eredményét. Az $e_i^T A = a_i^T$ összefüggést felhasználva kapjuk, hogy

$$PA = \begin{bmatrix} e_{i_1}^T \\ e_{i_2}^T \\ \vdots \\ e_{i_n}^T \end{bmatrix} A = \begin{bmatrix} e_{i_1}^T A \\ e_{i_2}^T A \\ \vdots \\ e_{i_n}^T A \end{bmatrix} = \begin{bmatrix} a_{i_1}^T \\ a_{i_2}^T \\ \vdots \\ a_{i_n}^T \end{bmatrix}.$$

Ez azt jelenti, hogy a PA szorzás felcseréli az A mátrix sorait. Ezzel szemben az AP szorzás az A mátrix oszlopait cseréli fel. Legyen a P permutációmátrix oszlopvektorok szerint particionálva, azaz legyen

$$P = [e_{j_1}, \ldots, e_{j_n}],$$

ahol $j_1, j_2, \ldots, j_n$ az $1, 2, \ldots, n$ számok valamelyik permutációja. Az A mátrix oszlopok szerinti particiója legyen $A = [a_1, \ldots, a_n]$ ($a_i \in \mathbb{R}^n$). Mármost $Ae_i = a_i$ miatt

$$AP = A\,[e_{j_1}, \ldots, e_{j_n}] = [Ae_{j_1}, \ldots, Ae_{j_n}] = [a_{j_1}, \ldots, a_{j_n}].$$

Könnyen belátható, hogy tetszőleges permutációmátrix esetén $P^T P = I$.

Amikor a részleges főelemkiválasztáson alapuló Gauss-módszer k-adik lépésében az $A^{(k-1)}$ mátrix k-adik és i-edik sorát ($i > k$) felcseréljük, akkor ez ekvivalens a $P_k A^{(k-1)} x = P_k b^{(k-1)}$ átalakítással, ahol

$$P_k = \begin{bmatrix} e_1^T \\ \vdots \\ e_{k-1}^T \\ e_i^T \\ e_{k+1}^T \\ \vdots \\ e_{i-1}^T \\ e_k^T \\ e_{i+1}^T \\ \vdots \\ e_n^T \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Itt a k-adik sorban e_i^T, az i-edik sorban pedig e_k^T áll. A teljes főelemkiválasztás esetén, amikor oszlopcserét is végezhetünk, az $P_k A^{(k-1)} Q_k \left(Q_k^T x\right) = P_k b^{(k-1)}$, ahol Q_k az oszlopcserének megfelelő permutációmátrix. Igaz a következő

2.5 Tétel. *Ha az A $n \times n$-es mátrix nemszinguláris, akkor létezik olyan P permutációmátrix, hogy a PA mátrixnak van LU-felbontása.*

3.7. Az LU-felbontás és a Gauss-módszer kapcsolata

Particionáljuk az A mátrixot az

$$A = \begin{bmatrix} a & r^T \\ c & B \end{bmatrix} \quad \left(a \in \mathbb{R},\ a \neq 0,\ c, r \in \mathbb{R}^{n-1}\right)$$

formában. A Gauss-módszer első lépése megfelel az

$$\tilde{L}_1 A = \begin{bmatrix} a & r^T \\ 0 & B_1 \end{bmatrix} = A_1$$

szorzásnak, ahol

$$\tilde{L}_1 = \begin{bmatrix} 1 & 0^T \\ -c/a & I \end{bmatrix}, \quad B_1 = B - cr^T/a.$$

Az $\tilde{L}_1$ particionált alsó háromszögmátrix inverze

$$\tilde{L}_1^{-1} = \begin{bmatrix} 1 & 0^T \\ c/a & I \end{bmatrix}$$

és

$$A = \begin{bmatrix} 1 & 0^T \\ c/a & I \end{bmatrix} \begin{bmatrix} a & r^T \\ 0 & B_1 \end{bmatrix}.$$

Ha a B_1 mátrixnak létezik LU-felbontása $B_1 = L_1 U_1$, akkor A–nak is van:

$$A = \begin{bmatrix} 1 & 0^T \\ c/a & L_1 \end{bmatrix} \begin{bmatrix} a & r^T \\ 0 & U_1 \end{bmatrix} = LU.$$

Azt még nem tudjuk, hogy mi a B_1 mátrix LU-felbontása, de azt már tudjuk, hogy az A mátrix LU-felbontásában az első oszlop, ill. sor micsoda. Ez ugyanis nem más mint

$$A = \begin{bmatrix} 1 & 0^T \\ c/a & ? \end{bmatrix} \begin{bmatrix} a & r^T \\ 0 & ? \end{bmatrix}.$$

Ha megismételjük az eliminációs lépést a B_1 mátrixon, akkor megkapjuk az L_1 első oszlopát, valamint U_1 első sorát. Az eljárást így folytatva eljutunk az U

mátrixhoz. Ha a c/a vektort egy L mátrixba (a gyakorlatban A alsó háromszög részébe) beírjuk, akkor az LU-felbontást teljes egészében megkapjuk. Összegezve: a Gauss-módszer az I. fázisban előállítja az A mátrix LU-felbontását, pontosabban az ekvivalens

$$Ux = L^{-1}b \tag{3.17}$$

egyenletrendszert. Tehát a Gauss-módszer az $A = LU$ speciális szorzatfelbontáson (faktorizáción) alapul.

Ha az eljárás során főelemkiválasztást kell végrehajtani, akkor a Gauss-módszer a PA mátrix LU-felbontását adja meg, ahol P permutációmátrix.

Konstruktívan igazoljuk az 2.5 Tételt. Ha $a_{11} = 0$, akkor van olyan P_1 permutációmátrix, hogy

$$P_1 A = \begin{bmatrix} a & r^T \\ c & B \end{bmatrix} \quad (a \neq 0)\,.$$

Ekkor

$$P_1 A = \begin{bmatrix} 1 & 0^T \\ c/a & I \end{bmatrix} \begin{bmatrix} a & r^T \\ 0 & B_1 \end{bmatrix}.$$

Ha B_1 nemszinguláris, akkor feltevésünk szerint létezik $\widehat{P}_2 \in \mathbb{R}^{(n-1)\times(n-1)}$ permutációmátrix, hogy $\widehat{P}_2 B_1 = L_1 U_1$. Tehát

$$P_1 A = \begin{bmatrix} 1 & 0 \\ c/a & I \end{bmatrix} \begin{bmatrix} a & r^T \\ 0 & \widehat{P}_2^T L_1 U_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ c/a & \widehat{P}_2^T L_1 \end{bmatrix} \begin{bmatrix} a & r^T \\ 0 & U_1 \end{bmatrix},$$

amelyet az $(n-1) \times n$ méretű alsó részmátrix sorait felcserélő

$$P_2 = \begin{bmatrix} 1 & 0^T \\ 0 & \widehat{P}_2 \end{bmatrix}$$

permutációmátrixszal szorozva a

$$P_2 P_1 A = \begin{bmatrix} 1 & 0 \\ \widehat{P}_2\,(c/a) & L_1 \end{bmatrix} \begin{bmatrix} a & r^T \\ 0 & U_1 \end{bmatrix}$$

összefüggést kapjuk. Ezzel az 2.5 Tételt igazoltuk.

Vegyük észre, hogy a síma Gauss-módszerrel ellentétben a P_2 permutációmátrixszal való szorzás a c/a sorait is felcseréli. Ez azt is jelenti, hogy a főelemkiválasztásos Gauss-módszer esetén, ahol P_2 előre ismeretlen, a további sorcserék a c/a oszlopvektor elemeit is megcserélik.

A síma Gauss-módszer esetében az L alsó háromszögmátrixot úgy kaphatjuk meg, hogy a c/a vektorokat beírjuk A főátlója alá. Ezek az elemek lesznek L

diagonális alatti elemei. Eszerint

$$L = \begin{bmatrix} 1 & 0 & & \dots & & 0 \\ a_{21}/a_{11} & \ddots & & & & \\ a_{31}/a_{11} & & 1 & & & \vdots \\ \vdots & & a_{k+1,k}/a_{kk} & \ddots & & \\ \vdots & & \vdots & & 1 & 0 \\ a_{n1}/a_{11} & \dots & a_{nk}/a_{kk} & \dots & a_{n,n-1}/a_{n-1,n-1} & 1 \end{bmatrix}$$

A főelemkiválasztás esetén ugyanezt tesszük, de a diagonális alatti elemeken is végrehajtjuk a sorcseréket.

3.8. Az LU- és Cholesky-módszerek

Legyen $A = LU$ és vizsgáljuk az $Ax = b$ megoldását. Ez az $Ax = LUx = L(Ux) = b$ összefüggés miatt felbontható az $Ly = b$ alsó háromszögmátrixú és az $Ux = y$ felső háromszögmátrixú egyenletrendszerek megoldására.

AZ LU-MÓDSZER ALGORITMUSA (I.):

1. Határozzuk meg az $A = LU$ felbontást!
2. Oldjuk meg az $Ly = b$ egyenletrendszert!
3. Oldjuk meg az $Ux = y$ egyenletrendszert!

Az eredeti Gauss-módszer I. fázisában az $A = LU$ felbontást és az $Ux = L^{-1}b$ felső háromszögmátrixú egyenletrendszert állítjuk elő. A II. fázisban ezt az egyenletrendszert oldjuk meg. Az LU-módszerben a Gauss-módszer I. fázisát két lépésre bontjuk fel. Az első lépésben az $A = LU$ felbontást állítjuk elő és értelemszerűen nem végzünk számításokat a b oszlopvektoron. Az eljárás második lépésében az $y = L^{-1}b$ vektort állítjuk elő. Az eljárás harmadik lépése megegyezik az eredeti Gauss-módszer II. fázisával.

Az LU-módszer különösen előnyös, ha ugyanazon együtthatómátrixszal egynél több

$$Ax = b_1,\ Ax = b_2, \dots,\ Ax = b_k$$

alakú egyenletrendszert kell megoldani. Ekkor elég az A mátrix LU-felbontását egyszer meghatározni, majd rendre az $Ly_i = b_i$, $Ux_i = y_i$ $(i = 1, \dots, k)$ háromszögmátrixú egyenletrendszereket megoldani. Az eljárás összköltsége: $n^3/3 + kn^2/2 + O(kn)$ régi flop.

Ezen észrevétel alapján egy $A \in \mathbb{R}^{n\times n}$ mátrix invertálását az LU-módszer segítségével a következőképpen végezhetjük:

1. Meghatározzuk az $A = LU$ felbontást.

2. Sorra meghatározzuk az $Ly_i = e_i, Ux_i = y_i$ $(i = 1, \ldots, n)$ egyenletrendszerek x_i megoldásait. Az A inverze $A^{-1} = [x_1, \ldots, x_n]$.

Az eljárás műveletigénye $O(n^3)$ régi flop.

Ha a $PA = LU$ felbontást tudjuk csak meghatározni, akkor az I. algoritmus értelemszerűen a következőképpen módosul.

AZ LU-MÓDSZER ALGORITMUSA (II.):

1. Határozzuk meg a $PA = LU$ felbontást!
2. Oldjuk meg az $Ly = Pb$ egyenletrendszert!
3. Oldjuk meg az $Ux = y$ egyenletrendszert!

Példa. Oldjuk meg az

$$\begin{array}{rcrcrcr} 1.2x_1 & + & 2.1x_2 & - & 3.2x_3 & = & -11.5 \\ 2.3x_1 & & & + & 4.5x_3 & = & 11.3 \\ 5.7x_1 & - & 3.1x_2 & + & 8.9x_3 & = & 32.8 \end{array}$$

egyenletrendszert LU-módszerrel!

Az LU-felbontást gyakorlatilag ugyanúgy végezzük, mint a Gauss-eliminációt, csak a főátló alatti számokat változtatás nélkül hagyjuk (nem írjuk be a 0-kat). A számítás során ezeket a számokat gömbölyű zárójellel jelöltük, a pivot elemeket pedig bekereteztük.

$$A = \begin{bmatrix} 1.2 & 2.1 & -3.2 \\ 2.3 & 0 & 4.5 \\ \boxed{5.7} & -3.1 & 8.9 \end{bmatrix}, \qquad P^{(0)} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad A^{(0)} = P^{(0)}A$$

$$\tilde{A} = \begin{bmatrix} 5.7 & -3.1 & 8.9 \\ (2.3) & 1.2509 & 0.9088 \\ (1.2) & \boxed{2.7506} & -5.0737 \end{bmatrix}, \; P^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad A^{(1)} = P^{(1)}\tilde{A}$$

$$A^{(2)} = \begin{bmatrix} \boxed{5.7} & -3.1 & 8.9 \\ (1.2) & \boxed{2.7506} & -5.0737 \\ (2.3) & (1.2509) & \boxed{3.2144} \end{bmatrix}, \; P = P^{(1)}P^{(0)} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0.2105 & 1 & 0 \\ 0.4035 & 0.4544 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 5.7 & -3.1 & 8.9 \\ 0 & 2.7526 & -5.0737 \\ 0 & 0 & 3.2144 \end{bmatrix}$$

Emlékeztetőül: az L az $A^{(2)}$ alsó háromszög részéből úgy keletkezett, hogy az oszlopokat osztottuk a főátlóban lévő elemekkel (azaz $A^{(2)}$ bekeretezett elemeivel), U pedig az $A^{(2)}$ felső háromszög része.

$LU = PA$, $b^{(2)} = Pb = P \begin{bmatrix} -11.5 \\ 11.3 \\ 32.8 \end{bmatrix} = \begin{bmatrix} 32.8 \\ -11.5 \\ 11.3 \end{bmatrix}$. A $PAx = Pb$ egyenletet megoldva:

$$Ly = Pb \Rightarrow y = \begin{bmatrix} 32.8 \\ -18.4053 \\ 6.4288 \end{bmatrix} \text{ és } Ux = y \Rightarrow x = \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}.$$

Láttuk, hogy ha az $A \in \mathbb{R}^{n\times n}$ mátrix szimmetrikus és pozitív definit, akkor felbontható $A = LL^T$ alakban, ahol L alsó háromszögmátrix. Ekkor nem kell tárolni az A mátrix felét és az LU-felbontás (LL^T-felbontás) kiszámításának a fele is megtakarítható. Legyen

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & \dots & l_{n1} \\ 0 & l_{22} & \dots & l_{n2} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & l_{nn} \end{bmatrix},$$

ahonnan kapjuk, hogy

$$a_{11} = l_{11}^2,\ a_{21} = l_{21}l_{11},\ \dots\ , a_{n1} = l_{n1}l_{11}.$$

Innen $l_{11} = \sqrt{a_{11}}$ és $l_{i1} = a_{i1}/l_{11} = a_{i1}/\sqrt{a_{11}}$ $(i = 2, \dots, n)$. Tegyük fel, hogy az L első $k-1$ oszlopát már meghatároztuk. Figyelembevéve, hogy az L^T mátrix k-adik oszlopában csak az első k elem nem zérus, kapjuk, hogy

$$a_{ik} = l_{i1}l_{k1} + l_{i2}l_{k2} + \dots + l_{i,k-1}l_{k,k-1} + l_{ik}l_{kk} \quad (i = k+1, \dots, n)$$

és

$$a_{kk} = l_{k1}^2 + l_{k2}^2 + \dots + l_{k,k-1}^2 + l_{kk}^2,$$

ahonnan $l_{kk} = (a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2)^{1/2}$. Ennek figyelembevételével

$$l_{ik} = (a_{ik} - \sum_{j=1}^{k-1} l_{ij}l_{kj})/l_{kk} \quad (i = k+1, \dots, n)\,.$$

Ennek alapján az eljárás algoritmusa a következő.

A CHOLESKY-MÓDSZER ALGORITMUSA:

for $k = 1 : n$
  $a_{kk} = (a_{kk} - \sum_{j=1}^{k-1} a_{kj}^2)^{1/2}$
  for $i = k+1 : n$
    $a_{ik} = (a_{ik} - \sum_{j=1}^{k-1} a_{ij}a_{kj})/a_{kk}$
  end
end

Az A mátrix alsó háromszög része fogja tartalmazni L-et. Az eljárás számítási költsége $\frac{n^3}{6} + O(n^2)$ régi flop. Az eljárás, amely a Gauss-elimináció speciális esetének tekinthető, nem igényel pivotálást. Megjegyezzük, hogy az algoritmus leírásában felhasználtuk, hogy megállapodás szerint $\sum_{j=i}^{k} s_j = 0$, ha $k < i$.

3.9. Az LU-módszer algoritmusa pointeres technikával

Az algoritmus ezen formája lényegében a 60-as évek eleje óta ismert. A P vektor tartalmazza a sorok indexeit. Induláskor $P(i) = i$ $(1 \leq i \leq n)$. Sorcserék esetén ténylegesen csak a P vektor megfelelő indexű elemeit cseréljük ki. A program a következő:

for $k = 1 : n-1$
  Határozzuk meg azt t indexet, amelyre $|A(P(t), k)| = \max_{k \leq i \leq n} |A(P(i), k)|$!
  if $k < t$
    Cseréljük fel a $P(k)$ és $P(t)$ vektorkomponensek értékét!
  end
  for $i = k+1 : n$
    $A(P(i), k) = A(P(i), k) / A(P(k), k)$
    $A(P(i), k+1 : n) = A(P(i), k+1 : n) - A(P(i), k) * A(P(k), k+1 : n)$
  end
end
for $i = 1 : n$
  $s = 0$
  for $j = 1 : i-1$
    $s = s + A(P(i), j) * x(j)$
  end
  $x(i) = b_{P(i)} - s$
end
for $i = n : -1 : 1$
  $s = 0$
  for $j = i+1 : n$
    $s = s + A(P(i), j) * x(j)$
  end
  $x(i) = (x(i) - s) / A(P(i), i)$
end

3.10. Utasítások, függvények és eljárások a MATLAB nyelvben

A MATLAB rendszerben nagyon sok beépített függvény és eljárás segíti a programozói munkát. (Eljáráson itt a hagyományos függvénykiértékelésnek nem tekinthető műveletsort – pl. képernyőre rajzolás – értünk.) Mindezen alprogramok körét

maga a programozó is bővítheti, így aztán egy MATLAB program általában kevés utasításból és sok-sok függvény, illetve eljárás behívásából áll.

3.10.1. Utasítások

Konkrét utasításból csak néhány fajta van. A már látott egyszerű értékadó utasításon kívül tulajdonképpen csak három további utasítást említhetünk: az *IF* utasítást, és két ciklusutasítást. Ezek más programnyelvből is jól ismert összetett, ún. blokkutasítások.

Az *IF* utasítás általános alakja:

if *feltétel*
  utasítások
  {**elseif**
    utasítások}
  <**else**
    utasítások>
end

A $\{\cdots\}$ akárhányszor ismétlődhet (el is maradhat), $<\cdots>$ elmaradhat.

A külön sorba írt részek egy sorba is írhatók. Ilyenkor vesszőt, vagy pontosvesszőt kell közéjük tenni. (Minden önálló utasítással is így kell eljárni.) Az *IF* utasítás szemantikája magyarázatot nem igényel, minimális programnyelvi ismerettel rá lehet jönni.

Feltételként logikai kifejezéseket használunk. Túlnyomórészt egyetlen reláció a kifejezés, de logikai műveletekkel össze is kapcsolhatjuk őket. Relációjelek:

Az utolsó kettő az "egyenlő", illetve "nemegyenlő" jel. A logikai összeadás, szorzás és a negáció jelei rendre:

A taxatív ciklusutasítás legáltalánosabb alakja:

for $i = v$
  utasítások
end

Az i a ciklusváltozó, v pedig általában egy vektor amely leggyakrabban `k:h:m` (vagy `k:m`) alakú. Szemantika (legyen pl. v elemszáma n): az i ciklusváltozó rendre felveszi a $v_1, v_2, \ldots, v_n$ értékeket és mindegyik mellett lefut a ciklusmag. (Ha történetesen v mátrix, akkor $v_1, v_2, \ldots, v_n$ rendre a mátrix oszlopvektorait jelenti.)

Az iteratív ciklus általános alakja:

while *feltétel*
  utasítások
end

A szemantika itt is ismerős.

3.10.2. Függvények

A MATLAB-ban való programozáshoz sokféle függvényt használhatunk. Egykét függvénnyel már találkoztunk, amelyek mátrixot közvetlenül (a "semmiből") hoznak létre, vagy egy mátrixot egyszerűen alakítanak át. Most néhány olyan függvényt sorolunk fel, amelyek már létező mátrixokból állítanak elő egy vagy több output értéket. Ezek inputjai is lehetnek természetesen skalárok. Bizonyos függvények alapjában véve csak skalárokra vannak értelmezve, ezeknél a függvény elemenként értendő. Pl. `sin(pi/4*ones(2,3))` azt a 2×3-as mátrixot jelenti, amelynek minden eleme $\sqrt{2}/2$. A függvények gazdag tárházából csak a legfontosabbakat említjük meg (a *HELP* segít a többit is megismerni).

Elemi matematikai függvények: *SIN, COS*, stb. Ezek a szokásos értelmezésűek, a *ROUND* szabályosan kerekít.

DET(A) az A determinánsát, *COND(A)* a mátrix kondíciószámát ("2"-es normában), *RANK(A)* pedig a rangját adja. Néhány további fontos függvény hívását mutatja a következő példa:

```
≫[m,n]=size(A), h2=norm(A),  hinf=norm(A,inf)
```

Itt az első utasítás az A mátrix (vagy vektor) méretét (sorainak és oszlopainak számát) adja (a kettő közül a nem kisebbet *LENGTH(A)*-val közvetlenül megkaphatjuk), a következő kettő pedig a "2"-es és "∞" normáját. Az "1"-es és a Frobenius norma is számolható a *NORM(A,1)*, illetve a *NORM(A,'fro')* hívással.

A következő függvények outputjai *sorvektorok* (adott esetben persze skalárak): *MAX, MIN, SUM, PROD*. Ha az input argumentum mátrix, de nem sorvektor, az eredményvektor az input oszlopainak a *legnagyobb elemét, legkisebb elemét, összegét, szorzatát* tartalmazza. Ha az input egy sorvektor, akkor a vektornak veszi a *legnagyobb elemét, legkisebb elemét, összegét,* és a *szorzatát* (az eredmény tehát itt mindenképpen skalár). Fontos tudni, hogy a *MAX*, és a *MIN* függvények két outputtal is hívhatók. Ilyenkor a második output a megfelelő elemek sorindexét tartalmazza. A *MAX, MIN* függvények egy fontos alkalmazását érdemes példán bemutatni. Legyen

$$A = \begin{bmatrix} 1 & 3 & 2 & 1 & -2 \\ 2 & 2 & 2 & 2 & 15 \\ 0 & 2 & 3 & -9 & 1 \\ 2 & 0 & 4 & 5 & 1 \end{bmatrix}.$$

`max(A)` csak a $[2, 3, 4, 5, 15]$ vektort adná, míg a

```
≫ [c, d]=max(A),  [c1,d1]=max(abs(A(3:4,2:4)))
```

utasításokkal a

$$\begin{aligned} c &= [2,3,4,5,15] \\ d &= [2,1,4,4,2] \\ c1 &= [2,4,9] \\ d1 &= [1,2,1] \end{aligned}$$

eredményeket kapjuk. A

$\gg$ `[p,j]=max(max(A))`

eredménye pedig a $p = 15$, $j = 5$, azaz A legnagyobb eleme és annak oszlopindexe. Az A maximális elemének sorindexét ezek után `d(j)` adja, ami most természetesen 2.

A beépített függvények között találunk néhány olyat is, amelyek a lineáris algebra egy-egy bonyolult algoritmusát hajtják végre. Ezek közül megemlítjük a legfontosabbakat.

$\gg$ `[L,U,P]=lu(A), [T,F]=lu(A), H=chol(C)`

Az első kettő az LU-felbontás két verziója, az utolsó a C mátrix (aminek szimmetrikus, pozitív definitnek kell lennie) Cholesky-felbontását adja. Az eredményekre igaz a következő: L, PT, H alsó háromszögmátrixok, U, F felső háromszögmátrixok, P permutációmátrix, továbbá

$$L = PT, \quad U = F, \quad LU = PA, \quad TF = A, \quad HH^T = C.$$

$\gg$ `X=inv(A), [Q,R]=qr(A)`

Itt X az A inverze (ha az létezik), Q ortogonális, R felső háromszögmátrix és $QR = A$.

$\gg$ `s=eig(A), [V,D]=eig(A)`

Az első utasítás A sajátértékeit helyezi az s vektorba. A második után a sajátértékek a diagonális D mátrixban jelennek meg, a V mátrix i-edik oszlopa pedig a d_{ii}-hez tartozó sajátvektor (aminek "2"-es normája 1) lesz.

Bár nem a beépített függvények közé tartoznak, de itt említjük meg a numerikus módszerek nemlineáris eljárásaira készült függvényeket. Néhány közülük: az *FSOLVE* nemlineáris egyenletrendszerek megoldását végzi, a *QUAD* integrált számol adaptív Simpson-formulával, a *SPLINE* természetes *spline* interpolációt végez.

3.10.3. M-adatállományok, eljárások

MATLAB utasítások sorozata egy MATLAB programot alkot. A programot $.M$ kiterjesztésű adatállományban (M-adatállomány) kell elhelyeznünk. A behívása egyszerűen a nevével történik. Pl. a

≫ `prog`
utasítás betölti és lefuttatja a *prog.m* adatállományban lévő programot. Az egyszerű program (*script-file*) változói globális változók.

Lehetőségünk van egy M-adatállományban a beépített függvényekkel egyező módon hívható saját függvények definiálására. Ekkor az utasítássorozatot

$$\textbf{function}\ [k_1,\ldots,k_t] = fnév(b_1,\ldots,b_s)$$

sorral kell kezdenünk. A $k_1,\ldots,k_t$ kimenő és a $b_1,\ldots,b_s$ bemenő paraméterek száma (elvben) tetszőleges. A függvény azonosítóját és az adatállomány nevét érdemes egyeztetni, azaz a függvényt az *fnév.m* állományban elhelyezni. Az állományban (*script-file*-ban is) bárhol elhelyezhetünk *RETURN* utasítást, ennek hatása a szokásos. A függvény definíciós programjában szereplő, nem paraméter változók lokálisak. Egy-egy változó globálissá tehető a *GLOBAL* utasítással. Bemenő paraméterként szerepelhet függvénynév is. Ilyenkor a **function** programban a *FEVAL* eljárást kell hívni. Pl. ha az f paraméter egy kétváltozós függvény, akkor `z = f(x,y)` hibás, helyette a `z = feval(f,x,y)` a helyes utasítás.

Ismerkedjünk meg néhány további lehetőséggel is.

A *DISP(változó)* vagy *DISP('szöveg')* eljárás a képernyőn megjeleníti az argumentumát.

A *DIARY oda* utasítás kiadása után minden képernyőüzenet *oda* (tetszőleges adatállomány lehet) is kiíródik. Ezen naplózást a `diary off` utasítással fejezhetjük be, `diary on` ott folytatja, ahol előzőleg abbahagyta.

Ha x és y egy-egy azonos hosszúságú vektor, akkor a `plot(x,y)` hatására a képernyőn egy görbét kapunk az (x_i, y_i) pontokból. A *PLOT* eljárást más formában is hívhatjuk.

Az időmérést egy példán keresztül mutatjuk be:

≫ `t=clock %vagy t=cputime`
≫ {*utasítások*}
≫`tt=etime(clock,t) %vagy tt=cputime-t`

A *CLOCK* egy 6 elemű vektorban a dátum és idő adatokat tárolja, az *ETIME* az argumentumainak különbségét adja másodpercekben (század pontossággal), a *CPUTIME* pedig a MATLAB indítása óta eltelt időt méri.

3.11. Feladatok

1. Mekkora a lehetséges legnagyobb aritmetikai műveletigénye a részleges és a teljes főelemkiválasztásnak?
2. Állítson elő a MATLAB-ban egy véletlen 4×4-es A mátrixot és egy csupa 1-es, 4 elemű y vektort, majd a b vektort, melyre $b = Ay$. Program és az LU

függvény használata nélkül, egyszerre csak egy-egy utasítást adva ki, lépésről-lépésre hajtva végre az algoritmusokat, számítsa ki az $Ax = b$ egyenletrendszer megoldását Gauss-módszerrel

(i) pivotálás nélkül,

(ii) részleges főelemkiválasztással,

(iii) teljes főelemkiválasztással,

(iv) valamint LU-módszerrel is!

(v) Határozza meg a pivot elemek növekedési tényezőjét az első három esetben!

(vi) Hasonlítsa össze a megoldásokat az $x = [1, 1, 1, 1]^T$ pontos megoldással!

3. Az előző feladat A mátrixával számítsa ki a $B = A^T A$ mátrixot, majd

(i) határozza meg a $B = LU$ felbontást Gauss-módszerrel!

(ii) Adja meg azt a D diagonál mátrixot, melyre $L_1 = LD$ és $D^{-1}U = L_1^T$ teljesül!

(iii) A *CHOL* függvénnyel számítsa ki B Cholesky-felbontását és hasonlítsa össze az előbb kapott L_1 illetve L_1^T-vel!

4. fejezet

A KLASSZIKUS HIBASZÁMÍTÁS ELEMEI

A klasszikus hibaszámítás alapmodellje a következő. A pontos értékeket nem ismerjük, csak adott hibakorlátú közelítéseiket. A közelítő értékekkel pontosan végzett műveletek eredményét az ismeretlen elméleti eredmény közelítésének tekintjük és azt vizsgáljuk, hogy mekkora a közelítés hibája. Például a $\sqrt{2} \approx 1.41$ közelítés hibája legfeljebb 0.01.

A következő jelöléseket és elnevezéseket használjuk: x pontos érték, a az x közelítése ($a \approx x$), $\Delta a = x - a$ a közelítés hibája, δa az a közelítő érték abszolút hibakorlátja, ha fennáll $|x - a| = |\Delta a| \leq \delta a$.

4.1. Az aritmetikai műveletek abszolút hibái

Legyen x és y két pontos érték, a az x, b pedig az y közelítése. Tegyük fel, hogy az a és b közelítések abszolút hibakorlátai δa, ill. δb. Ezekre fennáll, hogy

$$|x - a| = |\Delta a| \leq \delta a, \quad |y - b| = |\Delta b| \leq \delta b.$$

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

$$|\Delta (a \Diamond b)| = |(x \Diamond y) - (a \Diamond b)| \leq \delta (a \Diamond b)$$

mennyiségre keresünk becsléseket, ahol $\Delta (a \Diamond b)$ a művelet hibáját, $\delta(a \Diamond b)$ pedig abszolút hibakorlátját jelöli.

6.1 Tétel. *Igazak a következő becslések:*

$$\delta (a + b) \leq \delta a + \delta b, \tag{4.1}$$

$$\delta (a - b) \leq \delta a + \delta b. \tag{4.2}$$

Bizonyítás. Az összeg esetében fennáll, hogy

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

amiből a fenti állítás következik. A különbség esetében hasonlóan kapjuk, hogy

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

ami bizonyítandó volt. □

A szorzat abszolút hibakorlátjára kapjuk, hogy

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

Ha $|a| \gg \delta a$ és $|b| \gg \delta b$, akkor a $\delta a\delta b$ másodrendű hibatagot elhanyagolhatjuk és a

$$\delta(ab) \approx |a|\,\delta b+|b|\,\delta a \tag{4.3}$$

közelítő becslést kapjuk.

Az osztás esetén azt kapjuk, hogy

$$\begin{aligned} \left|\frac{x}{y}-\frac{a}{b}\right| &= \left|\frac{a+\Delta a}{b+\Delta b}-\frac{a}{b}\right| \\ &= \left|\frac{-a\Delta b+b\Delta a}{b(b+\Delta b)}\right| \leq \frac{|a|\,|\Delta b|+|b|\,|\Delta a|}{|b|^2\,(1-\frac{|\Delta b|}{|b|})} \leq \frac{|a|\,\delta b+|b|\,\delta a}{|b|^2\,(1-\frac{\delta b}{|b|})}. \end{aligned}$$

Ha $|b| \gg \delta b$, akkor a nevezőben lévő $\frac{\delta b}{|b|}$ tagot elhanyagolhatjuk és a

$$\delta(a/b) \approx \frac{|a|\,\delta b+|b|\,\delta a}{|b|^2} \tag{4.4}$$

közelítő becslést kapjuk.

6.2 Tétel. *Fennállnak az alábbi közelítő egyenlőségek:*

$$\delta(ab) \approx |a|\,\delta b+|b|\,\delta a \quad (|a| \gg \delta a,\ |b| \gg \delta b)\,, \tag{4.5}$$

$$\delta(a/b) \approx \frac{|a|\,\delta b+|b|\,\delta a}{|b|^2} \quad (b \neq 0,\ |b| \gg \delta b)\,. \tag{4.6}$$

Figyeljük meg, hogy osztás abszolút hibakorlátja 0-hoz közeli b esetén rendkívül nagy lehet!

4.2. Függvényértékek hibája

Külön foglalkozunk az egy- és a többváltozós esetekkel. Legyen $f : \mathbb{R} \to \mathbb{R}$ legalább kétszer folyonosan differenciálható függvény, $x \approx a$. Az $f(x)$ helyett $f(a)$-t számoljuk. Az

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

másodrendű Taylor-formulából kapjuk, hogy

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

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

$$\delta\left(f\left(a\right)\right) \approx |f'(a)|\,\delta a. \tag{4.7}$$

Legyen $f : \mathbb{R}^n \to \mathbb{R}$ legalább kétszer folytonosan differenciálható függvény, $x, a \in \mathbb{R}^n$ és $x \approx a$. Legyen $\Delta a = x - a$, $|x_i - a_i| = |\Delta a_i| \leq \delta a_i$ $(i = 1, \ldots, n)$ és $\delta a = [\delta a_1, \ldots, \delta a_n]^T$. Az $f(x)$ helyett az $f(a)$ függvényértéket számoljuk. A többváltozós

$$f(x) = f(a) + \nabla f(a)^T (x-a) + \frac{1}{2}(x-a)^T H(a + \xi(x-a))(x-a) \quad (0 < \xi < 1)\,,$$

Taylor-formulát használjuk, ahol $\nabla f\left(x\right) = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]^T$ és

$$H(x) = \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]_{i,j=1}^{n}$$

az ún. Hesse-mátrix. Tegyük fel, hogy $\|H(a + \xi(x-a))\| \leq M$. Ekkor az $\left|x^T y\right| \leq \|x\|_2 \|y\|_2$ és $\|Ax\|_2 \leq \|A\|_F \|x\|_2$ egyenlőtlenségek alapján

$$\begin{aligned} \left|(x-a)^T H(a + \xi(x-a))(x-a)\right| &\leq \|x-a\|\,\|H(a + \xi(x-a))(x-a)\| \\ &\leq \|H(a + \xi(x-a))\|\,\|x-a\|^2\,. \end{aligned}$$

Ezt felhasználva kapjuk, hogy

$$\begin{aligned} |f(x) - f(a)| &\leq \left|\nabla f(a)^T (x-a)\right| + \frac{1}{2} M \|x-a\|^2 \\ &\leq \sum_{i=1}^{n} \left|\frac{\partial f(a)}{\partial x_i}\right| \delta a_i + \frac{1}{2} M \|x-a\|^2 \end{aligned}$$

ahonnan a másodrendű $M\|x-a\|^2 = M\|\Delta a\|^2 \leq M\sum_{i=1}^n (\delta a_i)^2$ tagot elhanyagolva kapjuk a

$$\delta\left(f\left(a\right)\right) \approx \sum_{i=1}^{n} \left|\frac{\partial f(a)}{\partial x_i}\right| \delta a_i \tag{4.8}$$

becslést.

4.3. Az aritmetikai műveletek relatív hibái

Az abszolút hiba sok esetben semmitmondó. Például egy 0.001 nagyságrendű elméleti mennyiség 0.05 abszolút hibakorlátú közelítése nem sokat ér. A $\pi \approx 22/7$ közelítés sok esetben jó lehet, de például a csillagászatban már bizonyosan nem.

6.1 Definíció. *Az x szám valamely a közelítő értékének relatív hibája a $\frac{\delta a}{|x|}$ mennyiség.*

Az x pontos érték általában nem ismeretes, ezért a $\frac{\delta a}{|x|}$ helyett a $\frac{\delta a}{|a|}$ közelítést használjuk. Ennek hibájára fennáll, hogy

$$\left|\frac{\delta a}{|x|} - \frac{\delta a}{|a|}\right| = \delta a \frac{||a|-|x||}{|a|\,|x|} \leq \delta a \frac{|a-x|}{|a|\,|x|} \leq \frac{(\delta a)^2}{|a|\,|x|}. \tag{4.9}$$

A hiba elhanyagolható, ha $|x|$ és $|a|$ lényegesen nagyobb a másodrendű $(\delta a)^2$ mennyiségnél. Ennek figyelembe vételével kaphatjuk a következő eredményeket:

$$\frac{\delta(a+b)}{|a+b|} = \max\left\{\frac{\delta a}{|a|}, \frac{\delta b}{|b|}\right\} \quad (ab>0)\,, \tag{4.10}$$

$$\frac{\delta(a-b)}{|a-b|} = \frac{\delta a + \delta b}{|a-b|} \quad (ab>0)\,, \tag{4.11}$$

$$\frac{\delta(ab)}{|ab|} \approx \frac{\delta a}{|a|} + \frac{\delta b}{|b|}, \tag{4.12}$$

$$\frac{\delta(\frac{a}{b})}{\left|\frac{a}{b}\right|} \approx \frac{\delta a}{|a|} + \frac{\delta b}{|b|}. \tag{4.13}$$

Figyeljük meg, hogy egymáshoz közeli a és b esetén a kivonás relatív hibája rendkívül nagy lehet!

Példa. Számítsuk ki a $\sqrt{1996} - \sqrt{1995}$ mennyiséget, ha ismertek a $\sqrt{1996} \approx 44.67$ és $\sqrt{1995} \approx 44.66$ közelítő értékek, amelyek közös abszolút hibakorlátja 0.01, a közös relatív hibakorlát pedig 0.022%. A kivonás elvégzésével kapjuk, hogy $\sqrt{1996} - \sqrt{1995} \approx 0.01$, amelynek relatív hibakorlátja az általános formulából

$$\frac{0.01 + 0.01}{0.01} = 2,$$

azaz 200%. Most lehetőségünk van az elméleti relatív hiba kiszámolására is, ami "csak" 10.66%. Ez a valóságos hiba is jelentős mértékű, a kiinduló adatok hibájához képest kb. 5×10^2-szoros. A különbség képzését elkerülhetjük a

$$\sqrt{1996} - \sqrt{1995} = \frac{1996 - 1995}{\sqrt{1996} + \sqrt{1995}} = \frac{1}{\sqrt{1996} + \sqrt{1995}} \approx \frac{1}{89.33} \approx 0.01119$$

átalakítással. A számláló pontos érték. A nevező abszolút hibája 0.02, a hányados relatív hibája pedig $0.02/89.33 \approx 0.00022 = 0.022\%$. Ez összhangban van a kiinduló adatok relatív hibáival és lényegesen kisebb, mint amit a közvetlen kivonásnál kaptunk.

Hasonló fogásokat lehet alkalmazni más esetekben is.

4.4. Függvényértékek relatív hibája és a kondíciószám

Legyen ismét $f : \mathbb{R} \to \mathbb{R}$ kétszer folytonosan differenciálható és $x \approx a$. Az $f(x)$ pontos érték helyett számított $f(a)$ érték relatív hibája

$$\frac{\delta(f(a))}{|f(a)|} \approx \frac{|f'(a)|\,\delta a}{|f(a)|}. \tag{4.14}$$

Érdekesebb ez a mennyiség az a közelítő érték relatív hibájával való összehasonlításban. A

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

mennyiség, amely a relatív hibák hányadosa, azt méri, hogy az a adatban fellépő bizonytalanságot az $f(a)$ függvénybe való behelyettesítés mennyire "nagyítja" meg. Egyszerű átalakításokkal adódik, hogy

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

6.2 Definíció. *A*

$$c(f, a) = \frac{|f'(a)|\,|a|}{|f(a)|} \tag{4.15}$$

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

Egy függvényt numerikusan instabilnak, vagy rosszul kondicionáltnak nevezünk, ha nagy a kondíciószáma. A függvény stabil, vagy jól kondicionált, ha a kondíciószám kicsi. Természetesen a kicsi és nagy jelző relatív. Mint később látni fogjuk, ezek a relatív jelzők adott feladat esetén a rendelkezésre álló számítógép aritmetikájától és a közelítés megkövetelt pontosságától függenek.

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

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

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

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

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

6.3 Definíció. *Egy $f(x)$ mennyiség $O(g(x)^\gamma)$ nagyságrendű, ha alkalmas $K > 0$ számmal fennáll $|f(x)| \leq K|g(x)|^\gamma$ minden szóbajövő x értékre. Egy $h(x)$ mennyiség $o(g(x))$ nagyságrendű, ha $\frac{\|h(x)\|}{g(x)} \to 0$ $(x \to x_0)$.*

Tekintsük az $F = [f_1, \ldots f_n]^T : \mathbb{R}^m \to \mathbb{R}^n$ többváltozós függvényt. Az F'-t a következőképpen értelmezzük:

$$F'(x) = \left[\frac{\partial f_i}{\partial x_j}\right]_{i,j=1}^{n,m}$$

Legyen most F olyan, amelyre fennáll, hogy

$$F(x) = F(a) + F'(a)(x-a) + o(\|x-a\|), \quad (x \to a).$$

Az $F(x)$ helyett $F(a)$-t számoljuk. Az $F(a)$ relatív hibája az a vektor relatív hibájához viszonyítva a következő

$$\frac{\|F(x) - F(a)\|}{\|F(a)\|} : \frac{\|x-a\|}{\|a\|}.$$

A "nagyítási szám" korlátja az a pont egy $\varepsilon > 0$ sugarú nyílt környezetében a

$$c(F,a,\varepsilon) = \sup\left\{\frac{\|F(x) - F(a)\|}{\|F(a)\|} : \frac{\|x-a\|}{\|a\|} \mid \|x-a\| < \varepsilon,\ x \neq a\right\}$$

mennyiség. Az

$$\frac{\|F(x) - F(a)\|}{\|F(a)\|} : \frac{\|x-a\|}{\|a\|} = \frac{(\|F'(a)(x-a)\| + o(\|x-a\|))}{\|x-a\|} \cdot \frac{\|a\|}{\|F(a)\|}.$$

és

$$\frac{\|F'(a)(x-a)\|}{\|x-a\|} \leq \|F'(a)\|$$

összefüggések miatt

$$\lim_{\varepsilon\to 0} c(F,a,\varepsilon) = \frac{\|a\|\,\|F'(a)\|}{\|F(a)\|}.$$

6.4 Definíció. *Az $F:\mathbb{R}^m\to\mathbb{R}^n$ függvény valamely a pontbeli kondíciószáma a*

$$c(F,a) = \frac{\|a\|\,\|F'(a)\|}{\|F(a)\|} \tag{4.16}$$

mennyiség.

Definiáljuk az $F:\mathbb{R}^n\to\mathbb{R}^n$ leképezést az $Ay=x$ egyenletrendszer megoldásával, azaz legyen $F(x)=A^{-1}x$ ($A\in\mathbb{R}^{n\times n}$, $\det(A)\neq 0$). Ekkor $F'\equiv A^{-1}$ és

$$c(A^{-1},a) = \frac{\|a\|\,\|A^{-1}\|}{\|A^{-1}a\|} = \frac{\|Ay\|\,\|A^{-1}\|}{\|y\|} \le \|A\|\left\|A^{-1}\right\| \quad (Ay=a).$$

A jobboldali felső korlátot az A mátrix kondíciószámának hívjuk. Ez a korlát pontos, mert létezik olyan $a\in\mathbb{R}^n$, hogy $c(A^{-1},a)=\|A\|\,\|A^{-1}\|$.

4.5. Direkt és inverz hibák

Vizsgáljuk egy $f(x)$ függvényérték kiszámítását. Ha az $\hat{y}$ közelítést számoljuk a pontos $y=f(x)$ érték helyett, akkor a *direkt (forward) hiba* $\Delta y=\hat{y}-y$. Ha egy $x+\Delta x$ értékre fennáll, hogy $\hat{y}=f(x+\Delta x)$, azaz $\hat{y}$ a perturbált (megváltoztatott) $x+\Delta x$ értékhez tartozó pontos függvényérték, akkor a Δx értéket *inverz (backward) hibának* nevezzük. A kétfajta hibát mutatja a következő ábra:

Az $y=f(x)$ értéket számító algoritmust *inverz stabilnak* nevezzük, ha bármely x értékre olyan $\hat{y}$ számított értéket ad, amelyre a Δx inverz hiba kicsi. A "kicsi" jelző környezetfüggő.

Vizsgáljuk most a direkt és az inverz hiba kapcsolatát. Tegyük fel, hogy $\hat{y}=f(x+\Delta x)$ és f kétszer folytonosan differenciálható. Ekkor

$$\hat{y}-y = f(x+\Delta x)-f(x) = f'(x)\,\Delta x + \frac{f''(x+\vartheta\Delta x)}{2!}(\Delta x)^2 \quad (\vartheta\in(0,1))$$

és a számított megoldás relatív hibája

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

Innen kapjuk az alábbi, hibaszámítási ökölszabálynak is nevezett

$$\frac{\delta(\hat{y})}{|y|} \le c(f,x)\,\frac{\delta(x)}{|x|} \tag{4.17}$$

közelítő egyenlőtlenséget, amely szóban kifejezve a következő:

$$\text{relatív direkt hiba} \leq \text{kondíciószám} \times \text{relatív inverz hiba}. \tag{4.18}$$

Az egyenlőtlenség azt mutatja, hogy egy rosszul kondicionált probléma számított megoldásának nagy lehet a (relatív) direkt hibája. Egy algoritmust *direkt stabilnak* nevezünk, ha a direkt hiba kicsi. Egy direkt stabil módszer nem feltétlenül inverz stabil. Ha az inverz hiba és a kondíciószám kicsi, akkor az algoritmus direkt stabil.

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

4.6. Feladatok

1. Vizsgáljuk meg $x^2 - y^2$ kiszámításának alábbi módjait:

$$F = x * x - y * y,$$
$$G = (x - y) * (x + y),$$
$$H1 = (x + y) * x,\ H2 = (x + y) * y,\ H = H1 - H2.$$

Legyen $x \approx a$ és $y \approx b$, valamint r_a és r_b a hozzájuk tartozó relatív hibakorlátok! Adjuk meg az F, G és H relatív hibakorlátait r_a és r_b függvényében! Melyik eljárás mikor jó?

2. Mutassuk meg, hogy a következő kifejezések numerikusan instabilak $x \approx 0$ esetén:

(1) $\frac{1-\cos x}{\sin^2 x}$,

(2) $\sin(100\pi + x) - \sin(x)$,

(3) $2 - \sin x - \cos x - e^{-x}$.

Számítsuk ki a fenti kifejezések értékét $x = 10^{-3}, 10^{-5}, 10^{-7}$ esetén és becsüljük a hibát! Alakítsuk át a kifejezéseket numerikusan stabillá!

3. Mutassuk meg, hogy a következő kifejezések numerikusan instabilak nagyon nagy x értékek esetén:

(1) $x - \sqrt{x^3 - 1}$,

(2) $e^{-\frac{2}{x}} - \frac{x}{1+x}$,

(3) $\sin(\frac{100x^5\pi}{3+x^5})$.

Számítsuk ki a fenti kifejezések értékét $x = 10^3, 10^5, 10^7$ esetén és becsüljük a hibát! Alakítsuk át a kifejezéseket numerikusan stabillá!

5. fejezet

A LEBEGŐPONTOS HIBAANALÍZIS

A digitális számítógépek egy F véges számhalmazt ábrázolnak és az aritmetikai műveleteket ezekkel a számokkal végzik. Ha a művelet eredménye az F halmazbeli szám, akkor a művelet eredményét pontosan kapjuk meg. Egyébként pedig három eset léphet fel:

- kerekítés ábrázolható (nemzérus) számhoz,
- alulcsordulás (kerekítés 0-hoz),
- túlcsordulás.

A tudományos-műszaki számítások zömét ún. *lebegőpontos aritmetikában* végezzük. Ennek legáltalánosabban elfogadott modellje a következő.

7.1 Definíció. *A lebegőpontos számok halmaza*

$$F(\beta,t,L,U) = \\ = \left\{\pm m\times\beta^e \mid \tfrac{1}{\beta}\le m<1,\ m=0.d_1d_2\dots d_t,\ L\le e\le U\right\}\cup\{0\}, \tag{5.1}$$

ahol

- β *a számrendszer alapja,*
- m *a lebegőpontos szám mantisszája a* β *alapú számrendszerben,*
- e *az ábrázolt szám kitevője (karakterisztikája, exponense),*
- t *a mantissza hossza (az aritmetika pontossága),*
- L *a legkisebb kitevő (alulcsordulási határ kitevője),*
- U *a legnagyobb kitevő (túlcsordulási határ kitevője).*

A három leggyakrabban használt számrendszer a következő:

elnevezés	β	felhasználás
bináris	2	legtöbb számítógép
decimális	10	legtöbb számológép
hexadecimális	16	IBM és hasonló nagyszámítógépek

A mantisszát felírhatjuk az

$$m = 0.d_1 d_2 \dots d_t = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \dots + \frac{d_t}{\beta^t} \tag{5.2}$$

alakban is. Innen látható, hogy az $\frac{1}{\beta} \leq m < 1$ feltétel miatt az első jegyre teljesülnie kell az $1 \leq d_1 \leq \beta - 1$ egyenlőtlenségnek. A többi számjegyre fennáll, hogy $0 \leq d_i \leq \beta - 1$ $(i = 2, \dots, t)$. Az ilyen számrendszereket *normalizáltaknak* nevezzük. A 0 jegyet és a tizedespontot értelemszerűen nem szokás ábrázolni. Ha $\beta = 2$, akkor az első jegy csak az 1 lehet, amelyet szintén nem ábrázolnak. A (5.2) felírást használva a $F = F(\beta, t, L, U)$ halmazt megadhatjuk az

$$F = \left\{ \pm \left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \dots + \frac{d_t}{\beta^t} \right) \beta^e \mid \ L \leq e \leq U \right\} \cup \{0\} \tag{5.3}$$

alakban is, ahol $0 \leq d_i \leq \beta - 1$ $(i = 1, \dots, t)$ és $1 \leq d_1$.

Példa. Határozzuk meg az $F(\beta, t, L, U)$ halmaz elemeinek számát! A mantissza t számjegye közül az első $\beta - 1$ féle lehet (0 nem!), a többi viszont a $0, 1, \dots, \beta - 1$ bármelyike. Előjeltől eltekintve tehát $(\beta - 1)\beta^{t-1}$ különböző mantissza állítható össze. Az $L \leq e \leq U$ miatt $U - L + 1$ különböző kitevőnk van. Könnyű belátni, hogy ha két szám akár a mantisszában, akár a kitevőben (vagy mindkettőben) különbözik, akkor nem egyenlőek. Mostmár az előjelet, valamint a zérust is beszámítva, kapjuk az elemek számát: $2(\beta - 1)\beta^{t-1}(U - L + 1) + 1$.

Az F halmaz elemei nem egyenletesen helyezkednek el a számegyenesen! Például $\beta = 2$, $t = 3$, $L = -1$ és $U = 2$ esetén a 33 elemű F halmaz pozitív része

$$\left\{ \frac{1}{4}, \frac{5}{16}, \frac{6}{16}, \frac{7}{16}, \frac{1}{2}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}, 1, \frac{10}{8}, \frac{12}{8}, \frac{14}{8}, 2, \frac{20}{8}, 3, \frac{28}{8} \right\}.$$

Általában, a β alapú számrendszerben az m mantisszára fennáll, hogy

$$\frac{1}{\beta} \leq m = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \dots + \frac{d_t}{\beta^t} \leq \frac{\beta - 1}{\beta} + \frac{\beta - 1}{\beta^2} + \dots + \frac{\beta - 1}{\beta^t} = 1 - \frac{1}{\beta^t}.$$

Az $\left[\frac{1}{\beta}, 1\right]$ intervallumba eső F-beli szomszédos számok távolsága β^{-t}. Minthogy az F halmaz elemeit a $\pm m \times \beta^e$ számok alkotják, a szomszédos F-beli számok távolsága az exponens értékének megfelelően változik. A szomszédos elemek legnagyobb távolsága β^{U-t}, a legkisebb pedig β^{L-t}.

Az ábrázolható számok nagyságrendjét adja meg a következő

7.1 Tétel. *Ha $a \in F$, $a \neq 0$, akkor $M_L \leq |a| \leq M_U$, ahol*

$$M_L = \beta^{L-1}, \quad M_U = \beta^U(1 - \beta^{-t}).$$

Bizonyítás. Tetszőleges $a \in F$, $a \neq 0$ számra fennáll, hogy

$$|a| = m\beta^e \quad \left(m \in \left[\frac{1}{\beta}, 1 - \frac{1}{\beta^t}\right]\right),$$

ahonnan $L \leq e \leq U$ miatt

$$\beta^{L-1} \leq \frac{1}{\beta}\beta^e \leq m\beta^e \leq \beta^e\left(1 - \beta^{-t}\right) \leq \beta^U\left(1 - \beta^{-t}\right).$$

Ezzel az állitást igazoltuk. □

Legyen $a, b \in F$ és jelölje $\Diamond$ a négy aritmetikai művelet $(+, -, *, /)$ bármelyikét. A következő esetek lehetségesek:

(1) $a \Diamond b \in F$ (pontos eredmény),

(2) $|a \Diamond b| > M_U$ (aritmetikai túlcsordulás),

(3) $0 < |a \Diamond b| < M_L$ (aritmetikai alulcsordulás),

(4) $a \Diamond b \notin F$, $M_L < |a \Diamond b| < M_U$ (nem ábrázolható eredmény).

Az utolsó két esetben a lebegőpontos aritmetika az $a \Diamond b$ eredményhez hozzárendeli a legközelebbi F-beli számot. Ha két szomszédos F-beli szám az $a \Diamond b$ eredménytől egyformán távol van, akkor általában a nagyobbik számhoz kerekítünk.

Például ötjegyű decimális aritmetika esetén a 2.6457513 számot a 2.6458 számhoz kerekítjük.

Legyen $G = [-M_U, M_U]$. Világos, hogy $F \subset G$. Legyen $x \in G$. A kerekítéssel x-hez rendelt F-beli számot jelölje $fl(x)$. Az $x \to fl(x)$ leképezést *kerekítésnek* nevezzük. Legyen $u = \frac{1}{2}\beta^{1-t}$ az *egységnyi kerekítés mértéke*. Igaz a

7.2 Tétel. *Ha $x \in G$, akkor*

$$fl(x) = x(1 + \varepsilon), \quad |\varepsilon| \leq u.$$

Bizonyítás. Az általánosság megszorítása nélkül feltehetjük, hogy $x > 0$. Tegyük fel, hogy az x számot közrefogó szomszédos F-beli számok $m_1\beta^e$ és $m_2\beta^e$. Ekkor fennáll, hogy

$$m_1\beta^e \leq x \leq m_2\beta^e,$$

ahol $\frac{1}{\beta} \leq m_1 < m_2 \leq 1 - \beta^{-t}$ és $m_2 - m_1 = \beta^{-t}$. Mármost $fl(x) = m_1\beta^e$, vagy $fl(x) = m_2\beta^e$. Bármelyik választás esetében igaz, hogy

$$|fl(x) - x| \leq \frac{|m_2 - m_1|}{2}\beta^e = \frac{\beta^{e-t}}{2}.$$

Ezért a kerekítés relatív hibájára fennáll, hogy

$$\frac{|fl(x) - x|}{|x|} \leq \frac{|fl(x) - x|}{m_1\beta^e} \leq \frac{\beta^{e-t}}{2m_1\beta^e} = \frac{\beta^{-t}}{2m_1} \leq \frac{1}{2}\beta^{1-t} = u.$$

Fennáll tehát, hogy $fl\,(x) - x = \lambda x u$, ahol $|\lambda| \leq 1$. Ezt átrendezve kapjuk, hogy

$$fl\,(x) = x(1 + \lambda u),$$

ahol az $\varepsilon = \lambda u$ számra teljesül, hogy $|\varepsilon| = |\lambda u| \leq u$. Ha $(1 - \beta^{-t})\,\beta^e \leq x \leq \beta^e$, akkor a bizonyítás az $m_2 = 1$ választással érvényben marad. Ezzel a tételt maradéktalanul igazoltuk. □

A tétel tulajdonképpen azt mondja ki, hogy a lebegőpontos aritmetikában a kerekítés relatív hibája korlátos és ez a korlát u, az egységnyi kerekítés mértéke.

A $\epsilon_M = 2u = \beta^{1-t}$ értéket szokás *gépi epszilonnak* is nevezni. Az ϵ_M az 1 és a hozzá legközelebbi 1-nél nagyobb szám távolsága. Bináris alap esetén az

$x = 1$
while $1 + x > 1$
  $x = x/2$
end

algoritmussal határozhatjuk meg $\epsilon_M/2$ értékét. A MATLAB rendszerben $\epsilon_M \approx 2.2204 \times 10^{-16}$.

A lebegőpontos aritmetikai műveletek eredményére vonatkozóan a következő feltevéssel élünk (szabvány modell):

$$fl(a \Diamond b) = (a \Diamond b)\,(1 + \varepsilon)\,, \quad |\varepsilon| \leq u \quad (a, b \in F)\,. \tag{5.4}$$

Az IEEE aritmetikai szabvány, amelyet később ismertetünk, kielégíti ezt a feltevést. A feltevés fontos következménye, hogy $a \Diamond b \neq 0$ esetén a műveletek relatív hibájára ugyancsak teljesül, hogy

$$\frac{|fl(a \Diamond b) - (a \Diamond b)|}{|a \Diamond b|} \leq u.$$

Tehát az aritmetikai műveletek relatív hibája kicsi.

Vannak bizonyos lebegőpontos aritmetikák, amelyek nem elégítik ki a (5.4) feltevést. Ennek az az oka, hogy a kivonásnál nincs egy ún. ellenőrző jegyük.

Az egyszerűség kedvéért vizsgáljuk az $1 - 0.111$ különbséget háromjegyű bináris aritmetikában. Az első lépésben a kitevőket azonos értékre hozzuk

$$\begin{array}{rrcrcccc} & 2 & \times & 0 & . & 1 & 0 & 0 \\ - & 2 & \times & 0 & . & 0 & 1 & 1 & 1 \end{array}$$

Ha a számítást négy értékes jegyre végezzük, akkor az eredmény

$$\begin{array}{rrcrcccc} & 2^1 & \times & 0 & . & 1 & 0 & 0 \\ - & 2^1 & \times & 0 & . & 0 & 1 & 1 & 1 \\ \hline & 2^1 & \times & 0 & . & 0 & 0 & 0 & 1 \end{array}$$

amelyből a normalizált eredmény $2^{-2} \times 0.100$. Vegyük észre, hogy a kivonásra került szám nem normalizált, mert első jegye 0. A felhasznált ideiglenes negyedik mantisszajegyet, ellenőrző jegynek nevezzük. Ha nincs ilyen ellenőrző jegy, akkor a megfelelő számítások

$$\begin{array}{rccccccc} & 2^1 & \times & 0 & . & 1 & 0 & 0 \\ - & 2^1 & \times & 0 & . & 0 & 1 & 1 \\ \hline & 2^1 & \times & 0 & . & 0 & 0 & 1 \end{array}$$

amelyből a normalizált eredmény $2^{-1} \times 0.100$. Ennek relatív hibája 100%. Nincs ellenőrző jegye a CRAY szuperszámítógépeknek, valamint egy sor zsebkalkulátornak.

Ha nincs ellenőrző jegy, akkor a műveletek eredményeire az

$$fl\left(x \pm y\right) = x\left(1+\alpha\right) \pm y\left(1+\beta\right), \quad |\alpha|, |\beta| \leq u, \tag{5.5}$$

$$fl\left(x \Diamond y\right) = \left(x \Diamond y\right)\left(1+\delta\right), \quad |\delta| \leq u, \quad \Diamond = *, /. \tag{5.6}$$

összefüggések teljesülnek.

Tegyük fel a továbbiakban, hogy van ellenőrző jegy a kivonásnál és teljesül a (5.4) feltevés. Vezessük be a következő jelöléseket:

$$|z| = \left[|z_1|, \ldots, |z_n|\right]^T \quad \left(z \in \mathbb{R}^n\right), \tag{5.7}$$

$$|A| = \left[|a_{ij}|\right]_{i,j=1}^{m,n} \quad \left(A \in \mathbb{R}^{m \times n}\right), \tag{5.8}$$

$$A \leq B \Leftrightarrow a_{ij} \leq b_{ij} \quad \left(A, B \in \mathbb{R}^{m \times n}\right). \tag{5.9}$$

Igazolhatók az alábbi eredmények, ahol E az aktuális művelet hibáját (hibamátrixát) jelöli:

$$\left|fl\left(x^T y\right) - x^T y\right| \leq 1.01nu\left|x\right|^T\left|y\right| \quad \left(nu \leq 0.01\right), \tag{5.10}$$

$$fl\left(\alpha A\right) = \alpha A + E \quad \left(|E| \leq u\left|\alpha A\right|\right), \tag{5.11}$$

$$fl\left(A + B\right) = \left(A + B\right) + E \quad \left(|E| \leq u\left|A + B\right|\right), \tag{5.12}$$

$$fl\left(AB\right) = AB + E \quad \left(|E| \leq nu\left|A\right|\left|B\right| + O\left(u^2\right)\right). \tag{5.13}$$

A szabvány modellnek eleget tevő lebegőpontos aritmetikáknak számos sajátos tulajdonsága van. Fontos tulajdonságuk, hogy az összeadás a kerekítés miatt nem asszociatív. Ezt mutatják a következő MATLAB példák, amelyeket a *format long e* utasítás kiadása után ellenőrízhetünk.

Példa. MATLAB rendszerben

$$fl\left(fl\left(10^{-16}+1\right)-1\right) = 0, \quad fl\left(10^{-16} + fl\left(1-1\right)\right) = 10^{-16}.$$

Példa. Ha $a = 1$, $b = c = 3 \times 10^{-16}$, akkor a MATLAB 6.1 rendszerben Pentium 4 processzorú számítógépen

$$(a+b)+c \neq a+(b+c),$$

amelyet az $((a+b)+c)-(a+(b+c))$ utasítás segítségével ellenőrizhetünk.

Nagyszámú adat összegzésénél a kommutativitással (tulajdonképpen asszociativitással) is probléma lehet. Vizsgáljuk most a $\sum_{i=1}^{n} x_i$ összeg kiszámítását! A természetes algoritmus az ún. rekurzív összegzés:

$s = 0$
for $i = 1 : n$
  $s = s + x_i$
end

Példa. Számítsuk ki az

$$s_n = 1 + \sum_{i=1}^{n} \frac{1}{i^2 + i}$$

összeget $n = 4999$ esetén. A rekurzív összegzéssel kapott MATLAB eredmény

$$1.999800000000002e + 000.$$

Ha az összegzést fordított (azaz nagyság szerint növekedő) sorrendben végezzük, akkor az eredmény

$$1.999800000000000e + 000.$$

Ha a kétféle képpen kapott értékeket összevetjük az elméleti $s_n = 2 - \frac{1}{n+1}$ összeggel, akkor láthatjuk, hogy a második összegzés adott pontos eredményt. Ennek magyarázata az, hogy amikor a kisebb tagokkal kezdjük, akkor ezek összegei értékes jegyeket érnek a végső eredményben.

Nagy mennyiségű, előjelben és nagyságrendben eltérő szám nagy pontosságú összeadása nem egyszerű feladat. A következő algoritmus, amely az egyik legérdekesebb ilyen célra kifejlesztett eljárás, W. Kahantól származik.

A KOMPENZÁLT ÖSSZEGZÉS ALGORITMUSA:

$s = 0$
$e = 0$
for $j = 1 : n$
  $temp = s$
  $y = x(j) + e$
  $s = temp + y$
  $e = (temp - s) + y$
end

Példa. Kahan algoritmusát az

$$s_{4999} = \sum_{j=1}^{4999} \frac{1}{j^2 + j}$$

összegre az $x(j) = 1/(j^2 + j)$ szereposztással alkalmazva a MATLAB 6.1 a pontos 9.998000000000000e-001 értéket adja.

5.1. A lebegőpontos aritmetikai szabvány

Az ANSI/IEEE Std 754-1985 bináris ($\beta = 2$) lebegőpontos aritmetikai szabványt 1985-ben hozták nyilvánosságra. A szabvány specifikálja az alapvető lebegőpontos műveleteket, összehasonlításokat, kerekítési módokat, az aritmetikai kivételeket és kezelésüket, valamint a különböző aritmetikai formák közti konverziót. A négyzetgyökvonás az alapvető műveletek közé tartozik. A szabvány nem mond semmit az exponenciális és transzcendens függvényekről.

A szabvány két fő lebegőpontos formátumot ismer: az egyszeres és a dupla pontosságút.

típus	méret	mantissza	e	u	$[M_L, M_U] \approx$
egyszeres	32 bit	23+1 bit	8 bit	$2^{-24} \approx 5.96 \times 10^{-8}$	$10^{\pm 38}$
dupla	64 bit	52+1 bit	11 bit	$2^{-53} \approx 1.11 \times 10^{-16}$	$10^{\pm 308}$

Mindkét formátumban egy bitet az előjelnek tartanak fenn. Minthogy a lebegőpontos számok normalizálva vannak és az első jegy mindig 1, ez a jegy nincs tárolva. A mantisszában szereplő +1 ezt a rejtett bitet jelzi.

Az aritmetikai kivételek kezelése a következő: Az IEEE aritmetika zárt rendszer. Minden aritmetikai műveletnek van matematikailag értelmes vagy értelmetlen eredménye. A kivételes műveletek esetén jelzést ad ki, amely után a számításokat előírásszerűen folytatja. Az IEEE aritmetikai szabvány kielégíti a (5.4) modellt.

Az IEEE szabvány korai hardver megvalósításai közül ki kell emelni az Intel 80x87 matematikai koprocesszorokat, a DEC Alpha, a HP (Precision Architecture), az IBM RS/6000, az INMOS T800 és T900 processzorokat, a Motorola (680x0) és Sun (SPARCstation) processzorokat, valamint a HP tudományos kalkulátorait.

Végül megjegyezzük a következőket. Egyszeres pontosság esetén a mantissza hossza kb. 7 értékes jegyet enged meg a tizes számrendszerbe átszámolva. Ugyanez dupla pontosság esetén kb. 16 értékes jegyet jelent. Létezik még egy 80 biten ábrázolt, ún. kiterjesztett pontosság is, ahol $t = 63$, a kitevő pedig 15 bites.

5.2. Feladatok

1. Ábrázoljuk az

$$r(x) = \frac{622 - x(751 - x(324 - x(59 - 4x)))}{112 - x(151 - x(72 - x(14 - x)))}$$

függvényt az $x = 1.606 + (k-1)\,2^{-52}$ pontokban ($k = 1, \ldots 361$)! Mit tapasztalunk és mi a magyarázatuk?

2. Számítsuk ki x_{75} értékét, ha $x_{i+2} = -\frac{13}{6}x_{i+1} + \frac{5}{2}x_i$ $(i = 1, \ldots)$ és $x_1 = 30$, $x_2 = 25$. Hasonlítsuk össze a kapott eredményt az elméleti $x_i = 36\left(\frac{5}{6}\right)^i$ $(i = 1, \ldots)$ megoldással! Ábrázoljuk grafikusan is a számított és elméleti sorozatokat! Mi lehet az eredmények magyarázata?

3. Írjunk egy olyan programot, amely kísérleti úton becsli a következő függvények kondíciószámait a megadott pontokban:

$$f(x) = \prod_{i=1}^{11}(x-i) - 10^{-7}x^9 \quad (x = 0, 1.5, 3, 7, 10, 11) \tag{i}$$

$$f(x) = 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1 \quad (x = 0, 0.5, 1.0, 2.5) \tag{ii}$$

$$f(x) = \frac{\left(3\sqrt{x-2.6} + 1.26\right)\sin\left[(x+1)\,e^{0.2(x+1)}\right]\cos\left[(x+1)\,e^{-0.3(x+1)}\right]}{\left[1+(x-2)^4\right]\left[\left(\sqrt{|x|}+1\right)/\left(1+3\sqrt{|x|}\right)\right] + \log\left[(1+x^2)/\left(1+\sin^2 x\right)\right]}, \tag{iii}$$

ahol $x = -1, 0, 1, 2, 2.6, 200$.

6. fejezet

LINEÁRIS EGYENLETRENDSZEREK HIBAANALÍZISE

A vizsgálatban a direkt és az inverz hibákat elemezzük. Az $Ax = b$ egyenletrendszer megoldásával kapcsolatban a következő jelöléseket és fogalmakat használjuk. Az elméleti megoldást x, a közelítő megoldásokat pedig $\hat{x}$ jelöli. A közelítő megoldás direkt hibája $\Delta x = \hat{x} - x$. Az $r = r(y) = Ay - b$ mennyiséget *reziduális hibának* nevezzük. Az elméleti megoldás esetén $r(x) = 0$, a közelítő megoldás esetén pedig

$$r(\hat{x}) = A\hat{x} - b = A(\hat{x} - x) = A\Delta x.$$

Az inverz hiba meghatározásához különféle modelleket használunk. A legáltalánosabb esetben feltesszük, hogy az $\hat{x}$ számított megoldás kielégíti az $\hat{A}\hat{x} = \hat{b}$ egyenletrendszert, ahol $\hat{A} = A + \Delta A$ és $\hat{b} = b + \Delta b$. A ΔA és Δb mennyiségeket inverz hibáknak nevezzük.

Meg kell különböztetnünk a probléma érzékenységét és a megoldó algoritmusok stabilitását. Egy adott probléma érzékenységén a megoldás változásának mértékét értjük a probléma (input) paramétereinek függvényében. Egy algoritmus érzékenységén, vagy stabilitásán a számítási hibák végeredményre gyakorolt hatásának mértékét értjük. Egy problémát, vagy algoritmust annál stabilabbnak tekintünk mennél kisebb az input paraméterek, ill. számítási hibák megoldásra (számított megoldásra) gyakorolt hatása. Az érzékenység, ill. stabilitás fogalmának egyik jellemzési formája a korábban látott kondíciószám, amely az eltérések relatív hibáit hasonlítja össze.

Algoritmusok felhasználásának a következő általános elveit lehet megfogalmazni:

- A gyakorlatban csak stabil (jól kondicionált) algoritmusokat használunk.

- Instabil (inkorrekt kitűzésű), vagy rosszul kondicionált feladatot általános célú algoritmusokkal általában nem tudunk megoldani.

6.1. Érzékenységvizsgálat

Tegyük fel, hogy az $Ax = b$ egyenlet helyett a perturbált

$$A\hat{x} = b + \Delta b \tag{6.1}$$

egyenletrendszert oldjuk meg. Legyen $\widehat{x} = x + \Delta x$ és vizsgáljuk a két megoldás eltérését!

8.1 Tétel. *Ha A nemszinguláris és $b \neq 0$, akkor*

$$\frac{\|\Delta x\|}{\|x\|} \leq \operatorname{cond}(A) \frac{\|\Delta b\|}{\|b\|}, \tag{6.2}$$

ahol $\operatorname{cond}(A) = \|A\| \, \|A^{-1}\|$ *az A mátrix ún. kondíciószáma.*

Bizonyítás. Az $A\hat{x} = A(x + \Delta x) = b + \Delta b$ egyenlőségből $\Delta x = A^{-1}\Delta b$, ahonnan $\|\Delta x\| \leq \|A^{-1}\| \, \|\Delta b\|$ következik. Másrészt $\|b\| \leq \|A\| \, \|x\|$, ahonnan $\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$. A két egyenlőtlenséget összeszorozva kapjuk, hogy

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A\| \left\|A^{-1}\right\| \frac{\|\Delta b\|}{\|b\|},$$

ami éppen a bizonyítandó állítás. □

A tételből látható, hogy az A mátrix kondíciószáma erősen befolyásolhatja az $\widehat{x}$ perturbált megoldás relatív hibáját. Egy rendszert jól kondicionáltnak nevezünk, ha $\operatorname{cond}(A)$ kicsi és rosszul kondicionáltnak nevezünk, ha $\operatorname{cond}(A)$ nagy. Értelemszerűen a nagy és kicsi jelzők relatívak és környezetfüggők. A kondíciószám függ a normától. Ha a normától való függés lényeges, akkor ezt külön jelöljük. Ennek megfelelően például $\operatorname{cond}_\infty(A) = \|A\|_\infty \, \|A^{-1}\|_\infty$. A kondicionáltság egy lehetséges geometriai jellemzését adja a következő példa.

Példa. A

$$\begin{array}{rcrcl} 1000x_1 & + & 999x_2 & = & b_1 \\ 999x_1 & + & 998x_2 & = & b_2 \end{array}$$

egyenletrendszer rosszul kondicionált ($\operatorname{cond}_\infty(A) = 3.99 \times 10^6$). A két egyenes majdnem párhuzamos. Ezért, ha perturbáljuk a jobboldalt, az új metszéspont messze lesz az előzőtől.

A most vizsgált modellben az inverz hiba Δb, a 8.1 Tétel pedig a relatív direkt hibára ad becslést. Ez teljes összhangban van a hibaszámítási ökölszabállyal. A tétel állítása az $r(\hat{x}) = A\hat{x} - b = \hat{b} - b = \Delta b$ összefüggés miatt átírható az ekvivalens

$$\frac{\|\hat{x} - x\|}{\|x\|} \leq \operatorname{cond}(A) \frac{\|r(\hat{x})\|}{\|b\|} \tag{6.3}$$

alakba. Az egyenlőtlenség jelentése az, hogy az $\hat{x}$ perturbált megoldás relatív hibája kicsi, ha A kondíciószáma kicsi és az $\|r(\hat{x})\| / \|b\|$ relatív reziduális hiba kicsi. Ha azonban a rendszer rosszul kondicionált, akkor ez nem szükségképpen igaz.

Példa. Vizsgáljuk az $Ax = b$ egyenletrendszert, ahol

$$A = \begin{bmatrix} 1+\epsilon & 1 \\ 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Legyen $\hat{x} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$. Ekkor $r = \begin{bmatrix} 2\epsilon \\ 0 \end{bmatrix}$ és $\|r\|_\infty / \|b\|_\infty = 2\epsilon$, de $\|\hat{x} - x\|_\infty / \|x\|_\infty = 2$.

Tegyük most fel, hogy az $Ax = b$ egyenletrendszer helyett a perturbált

$$(A + \Delta A)\hat{x} = b \tag{6.4}$$

egyenletrendszert oldjuk meg. Legyen ismét $\hat{x} = x + \Delta x$ és $r(\hat{x}) = A\hat{x} - b$! Felmerül a kérdés, hogy adott $\hat{x}$ esetén van-e egyáltalán ΔA inverz hiba úgy, hogy $(A + \Delta A)\hat{x} = b$. Igaz a következő

8.2 Tétel *(Rigal-Gaches). Tegyük fel, hogy $\widehat{x}, r(\widehat{x}) \neq 0$. Ekkor $\Delta A = -r(\widehat{x})\widehat{x}^T / \widehat{x}^T\widehat{x}$ minimális spektrálnormájú inverz hiba.*

Bizonyítás. Behelyettesítéssel kapjuk, hogy

$$(A + \Delta A)\widehat{x} = \left(A - r(\widehat{x})\widehat{x}^T / \widehat{x}^T\widehat{x}\right)\widehat{x} = A\widehat{x} - r(\widehat{x}) = b.$$

Tehát ΔA inverz hiba. A spektrálnorma definíciója alapján igazolható, hogy $\|\Delta A\|_2 = \|r(\widehat{x})\|_2 / \|\widehat{x}\|_2$. Ha E tetszőleges inverz hiba, azaz $(A + E)\widehat{x} = b$, akkor $\|r(\widehat{x})\|_2 = \|E\widehat{x}\|_2 \leq \|E\|_2 \|r(\widehat{x})\|_2$ miatt $\|E\|_2 \geq \|r(\widehat{x})\|_2 / \|\widehat{x}\|_2$. Tehát ΔA valóban minimális spektrálnormájú inverz hiba. □

Megjegyezzük, hogy $r(\widehat{x})^T\widehat{x} \neq 0$ esetén $\Delta A = -r(\widehat{x})\, r(\widehat{x})^T / r(\widehat{x})^T\widehat{x}$ is inverz hiba, amely azonban már nem minimális normájú.

A következő tétel azt mondja ki, hogy kis relatív reziduális hiba esetén a relatív inverz hiba is kicsi.

8.3 Tétel. *Tegyük fel, hogy $\hat{x} \neq 0$ az $Ax = b$ egyenletrendszer közelítő megoldása,* $\det(A) \neq 0$ *és $b \neq 0$. Ha $\|r(\hat{x})\|_2 / \|b\|_2 = \alpha < 1$, akkor a $\Delta A = -r(\hat{x})\hat{x}^T / \hat{x}^T\hat{x}$ mátrixra fennáll, hogy $(A + \Delta A)\hat{x} = b$ és $\|\Delta A\|_2 / \|A\|_2 \leq \alpha / (1 - \alpha)$.*

Bizonyítás. Csak az előző tételben nem szereplő állításrészt igazoljuk. Az $r + b = A\hat{x}$ egyenlőségből és az $\alpha < 1$ feltételből kapjuk, hogy

$$0 < \|b\|_2 - \|r\|_2 \leq \|r + b\|_2 \leq \|A\|_2 \|\hat{x}\|_2,$$

ahonnan

$$\frac{1}{\|A\|_2 \|\hat{x}\|_2} \leq \frac{1}{\|b\|_2 - \|r\|_2},$$

illetve

$$\frac{\|\Delta A\|_2}{\|A\|_2} = \frac{\|r\|_2}{\|A\|_2 \|\hat{x}\|_2} \leq \frac{\|r\|_2}{\|b\|_2 - \|r\|_2} = \frac{\frac{\|r\|_2}{\|b\|_2}}{1 - \frac{\|r\|_2}{\|b\|_2}} = \frac{\alpha}{1-\alpha}. \square$$

Ha a relatív inverz hiba kicsi és A kondíciószáma kicsi, akkor a relatív reziduális hiba is kicsi. Erre mutat rá a következő tétel.

Ha A rosszul kondicionált, akkor a 8.4 Tétel nem igaz.

Példa. Legyen $A = \begin{bmatrix} 1+\epsilon & 1 \\ 1 & 1-\epsilon \end{bmatrix}$, $\Delta A = \begin{bmatrix} 0 & 0 \\ 0 & \epsilon^2 \end{bmatrix}$ és $b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $(0 < \epsilon \ll 1)$. Ekkor $\mathrm{cond}_\infty(A) = (2+\epsilon)^2/\epsilon^2 \approx 4/\epsilon^2$ és $\|\Delta A\|_\infty / \|A\|_\infty = \epsilon^2/(2+\epsilon) \approx \epsilon^2/2$. Legyen most

$$\hat{x} = (A + \Delta A)^{-1} b = \frac{1}{\epsilon^3} \begin{bmatrix} 2 - \epsilon + \epsilon^2 \\ -2 - \epsilon \end{bmatrix} \approx \begin{bmatrix} 2/\epsilon^3 \\ -2/\epsilon^3 \end{bmatrix}.$$

Ekkor $r(\hat{x}) = A\hat{x} - b = \begin{bmatrix} 0 \\ \frac{2}{\epsilon} + 1 \end{bmatrix}$. Ezért $\|r(\hat{x})\|_\infty / \|b\|_\infty = \frac{2}{\epsilon} + 1$, ami nem kicsi.

Tegyük fel, hogy az $Ax = b$ egyenlet helyett a perturbált

$$(A + \Delta A)\hat{x} = b + \Delta b \tag{6.5}$$

egyenletrendszert oldjuk meg. Legyen $\widehat{x} = x + \Delta x$. Igaz a következő

8.5 Tétel. *Ha A nemszinguláris,* $\mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|} < 1$ *és $b \neq 0$, akkor*

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right)}{1 - \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}}. \tag{6.6}$$

A tételt nem igazoljuk. A tételből következik a következő "ökölszabály".

Ökölszabály. *Tegyük fel, hogy $Ax = b$. Ha A és b elemeit s decimális jegyre pontosak és* $\mathrm{cond}(A) \sim 10^t$*, ahol $t < s$, akkor a számított megoldás kb. $s - t$ jegyre lesz pontos.*

Az ökölszabály következő heurisztikus levezetését adjuk. A feltevések miatt

$$\frac{\|\Delta A\|}{\|A\|} \approx 10^{-s}, \quad \frac{\|\Delta b\|}{\|b\|} \approx 10^{-s}$$

és

$$\mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|} \approx 10^{t-s} \ll 1.$$

Ezért feltehetjük, hogy $1 - \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|} \approx 1$. A 8.5 Tétel alapján

$$\frac{\|\Delta x\|}{\|x\|} \approx \mathrm{cond}(A)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right) \approx 10^{t-s},$$

amiből az ökölszabály következik.

A 8.5 Tétel $\mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|} < 1$ feltételének jelentése szemléletes: azt biztosítja, hogy az $A + \Delta A$ mátrix ne legyen szinguláris. A $\mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|} < 1$ egyenlőtlenség ugyanis ekvivalens a $\|\Delta A\| < \frac{1}{\|A^{-1}\|}$ feltétellel és az A nemszinguláris mátrix legközelebbi szinguláris mátrixtól való távolsága éppen $\frac{1}{\|A^{-1}\|}$. Ennek igazolásához szükségünk van a következő eredményre.

8.1 Lemma. *Legyen $F \in \mathbb{R}^{n\times n}$ olyan, hogy $\|F\| < 1$. Akkor $I - F$ nemszinguláris,*

$$(I - F)^{-1} = \sum_{k=0}^{\infty} F^k \tag{6.7}$$

és

$$\left\|(I - F)^{-1}\right\| \leq \frac{1}{1 - \|F\|}. \tag{6.8}$$

8.6 Tétel *(Eckart-Young-Gastinel). Legyen $A \in \mathbb{R}^{n\times n}$ nemszinguláris, $P \in \mathbb{R}^{n\times n}$ pedig tetszőleges szinguláris mátrix. Akkor fennáll, hogy*

$$\|A - P\| \geq \frac{1}{\|A^{-1}\|}. \tag{6.9}$$

Létezik továbbá olyan P szinguláris mátrix, amelyre egyenlőség áll fenn.

Bizonyítás. Az állítást csak spektrálnormában igazoljuk. Legyen $P = A + \Delta A$. Ha $\|\Delta A\| < 1/\|A^{-1}\|$, akkor $A + \Delta A$ a fenti lemma miatt invertálható, ui.

$$(A + \Delta A)^{-1} = \left(I + A^{-1}\Delta A\right)^{-1} A^{-1}, \qquad \left\|A^{-1}\Delta A\right\| \leq \left\|A^{-1}\right\| \|\Delta A\| < 1.$$

Következésképpen $\|\Delta A\| \geq 1/\|A^{-1}\|$ teljesül. Megmutatjuk, hogy alkalmas ΔA választással egyenlőség is fennáll. Legyen $y \in \mathbb{R}^n$ olyan, hogy $\|y\|_2 = 1$ és $\|A^{-1}y\|_2 = \|A^{-1}\|_2$. Legyen továbbá $x = \frac{1}{\|A^{-1}\|_2^2} A^{-1}y$. Ekkor igaz, hogy $x^T A^{-1}y = 1$, ui.

$$x^T A^{-1}y = \frac{1}{\|A^{-1}\|_2^2} \underbrace{y^T A^{-T}}_{z^T} \underbrace{A^{-1}y}_{z} = \frac{\|A^{-1}y\|_2^2}{\|A^{-1}\|_2^2} = 1.$$

Legyen $\Delta A = -yx^T$. Minthogy $(A + \Delta A)A^{-1}y = y - y\left(x^T A^{-1}y\right) = 0$, az $A + \Delta A$ mátrix szinguláris. Felhasználva a korábban megmutatott

$$\left\|yx^T\right\|_2 = \|x\|_2 \|y\|_2,$$

egyenlőséget, egyszerű számítással

$$\|\Delta A\|_2 = \frac{1}{\|A^{-1}\|_2^2} \left\|A^{-1}y\right\|_2 = \frac{1}{\|A^{-1}\|_2}$$

adódik. Ezzel állításunkat igazoltuk. □

A 8.6 Tétel segítségével a kondíciószám egy újabb jellemzését adhatjuk:

$$\frac{1}{\operatorname{cond}(A)} = \min_{A+E \text{ szinguláris}} \frac{\|E\|}{\|A\|}. \tag{6.10}$$

Eszerint, ha egy mátrix rosszul kondicionált, akkor közel van egy szinguláris mátrixhoz. Megjegyezzük, hogy mátrixok kondíciószámát az $F(x) = A^{-1}x$ leképezés kondíciószámának felső becsléseként már korábban is megkaptuk.

Vezessük be a következő definíciót.

8.1 Definíció. *Egy lineáris egyenletrendszer megoldó eljárást gyengén stabilnak nevezünk egy H mátrixosztályon, ha minden jólkondicionált $A \in H$ mátrix és minden b esetén az $Ax = b$ egyenletrendszer $\hat{x}$ számított megoldásának $\|\hat{x} - x\| / \|x\|$ relatív hibája kicsi.*

Ha a 8.3-8.5 Tételeket összerakjuk, akkor a következő eredményt kapjuk.

8.7 Tétel *(Bunch). Egy lineáris egyenletrendszer megoldó algoritmus gyengén stabil a H mátrixosztályon, ha minden jólkondicionált $A \in H$ mátrix és minden b esetén az $Ax = b$ egyenletrendszer $\hat{x}$ számított megoldására fennáll az alábbi feltételek bármelyike:*

(1) $\|\hat{x} - x\| / \|x\|$ kicsi;

(2) $\|r(\hat{x})\| / \|b\|$ kicsi;

(3) Létezik ΔA úgy, hogy $(A + \Delta A)\hat{x} = b$ és $\|\Delta A\| / \|A\|$ kicsi.

A 8.5 Tétel becslését a gyakorlatban akkor tudjuk jól használni, ha ismert $\Delta b, \Delta A$ és $\operatorname{cond}(A)$, vagy ezek egy becslése. Ezek hiányában a közelítő megoldás hibáját csak utólagosan (a posteriori) lehet becsülni.

A következőkben komponensenkénti hibabecslésekkel foglalkozunk. Elsőként a közelítő megoldás abszolút hibájára adunk becslést az inverz hibák komponenseinek segítségével.

8.8 Tétel *(Bauer-Skeel). Legyen $A \in \mathbb{R}^n$ nemszinguláris és tegyük fel, hogy az $Ax = b$ egyenletrendszer $\widehat{x}$ közelítő megoldása kielégíti az $(A + E)\widehat{x} = b + e$ egyenletrendszert. Ha $S \in \mathbb{R}^{n\times n}$, $s \in \mathbb{R}^n$ és $\varepsilon > 0$ olyanok, hogy $S \geq 0$, $s \geq 0$, $|E| \leq \varepsilon S$, $|e| \leq \varepsilon s$, valamint $\varepsilon \left\| |A^{-1}|\, S \right\|_\infty < 1$, akkor*

$$\|\widehat{x} - x\|_\infty \leq \frac{\varepsilon \left\| |A^{-1}| \left(S|x| + s\right) \right\|_\infty}{1 - \varepsilon \left\| |A^{-1}|\, S \right\|_\infty} \tag{6.11}$$

Bizonyítás. Az $(A + E)\widehat{x} = b + e$ egyenlőséget balról megszorozva az A^{-1} mátrixszal kapjuk, hogy $\widehat{x} + A^{-1}E\widehat{x} = A^{-1}b + A^{-1}e = x + A^{-1}e$. Innen

$$\widehat{x} - x = A^{-1}e - A^{-1}E\widehat{x} = -A^{-1}Ex + A^{-1}e - A^{-1}E(\widehat{x} - x)$$

adódik. Mindkét oldalon a vektor abszolút értékét véve kapjuk, hogy

$$|\widehat{x} - x| \leq \varepsilon \left|A^{-1}\right| S|x| + \varepsilon \left|A^{-1}\right| s + \varepsilon \left|A^{-1}\right| S|\widehat{x} - x|\,.$$

Figyelembevéve, hogy $\|\,|y|\,\|_\infty = \|y\|_\infty$ $(y \in \mathbb{R}^n)$, az

$$\|\widehat{x} - x\|_\infty \leq \varepsilon \left\| \left|A^{-1}\right| (S\,|x| + s) \right\|_\infty + \varepsilon \left\| \left|A^{-1}\right| S \right\|_\infty \|\widehat{x} - x\|_\infty$$

egyenlőtlenséget kapjuk, ahonnan a tétel állítása már átrendezéssel adódik. □

Ha $e = 0$ $(s = 0)$, $S = |A|$ és

$$k_r(A) = \left\| \left|A^{-1}\right| |A| \right\|_\infty < 1, \tag{6.12}$$

akkor a

$$\|\widehat{x} - x\|_\infty \leq \frac{\varepsilon k_r(A)}{1 - \varepsilon k_r(A)} \tag{6.13}$$

becslést kapjuk. A $k_r(A)$ mennyiséget Skeel-féle normának nevezzük, noha a korábban definiált értelemben nem norma. A Skeel-féle norma kielégíti a

$$k_r(A) \leq \mathrm{cond}_\infty(A) = \|A\|_\infty \left\|A^{-1}\right\|_\infty \tag{6.14}$$

egyenlőtlenséget. Tehát a fenti becslés nem rosszabb mint a hagyományos kondíciószámot használó becslés. Az inverz hiba komponensenkénti becslését teszi lehetővé Oettli és Práger következő eredménye.

Legyenek adottak az $A, \delta A \in \mathbb{R}^{n\times n}$ mátrixok és a $b, \delta b \in \mathbb{R}^n$ vektorok. Tegyük fel, hogy $\delta A \geq 0$ és $\delta b \geq 0$. Legyen továbbá

$$D = \left\{ \Delta A \in \mathbb{R}^{n\times n} : |\Delta A| \leq \delta A \right\}, \quad G = \left\{ \Delta b \in \mathbb{R}^n : |\Delta b| \leq \delta b \right\}.$$

8.9 Tétel *(Oettli-Práger). Egy $\hat{x}$ számított megoldás akkor és csak akkor megoldása egy $(A + \Delta A)\,\hat{x} = b + \Delta b$ perturbált egyenletrendszernek, ahol $\Delta A \in D$ és $\Delta b \in G$, ha*

$$|r(\hat{x})| = |A\hat{x} - b| \leq \delta A\,|\hat{x}| + \delta b. \tag{6.15}$$

Bizonyítás. Tegyük fel, hogy létezik $\Delta A \in D$ és $\Delta b \in G$ úgy, hogy $(A + \Delta A)\,\hat{x} = b + \Delta b$. Ekkor

$$\begin{aligned} |A\hat{x} - b| &= |(A + \Delta A)\,\hat{x} - \Delta A\hat{x} - (b + \Delta b) + \Delta b| \\ &\leq |(A + \Delta A)\,\hat{x} - (b + \Delta b)| + |\Delta A\hat{x}| + |\Delta b| \\ &\leq \delta A\,|\hat{x}| + \delta b. \end{aligned}$$

Fordítva tegyük fel, hogy $\hat{x}$ kielégíti az (6.15) egyenlőtlenséget. Legyen $r = A\hat{x} - b$, $s = \delta A\,|\hat{x}| + \delta b$ és

$$t_i = \begin{cases} \frac{r_i}{s_i}, & \text{ha } s_i \neq 0 \\ 0, & \text{ha } s_i = 0 \end{cases}.$$

Fennáll, hogy $|t_i| \leq 1$ és az $s_i = 0$ feltételből következik, hogy $r_i = 0$. Tehát

$$r_i = s_i t_i = \left(\delta b_i + \sum_{j=1}^{n} \delta a_{ij} |\hat{x}_j| \right) t_i = \delta b_i t_i + \sum_{j=1}^{n} \delta a_{ij} sign(\hat{x}_j)\, t_i \hat{x}_j.$$

Legyen $\varepsilon_i = \delta b_i t_i$ és $e_{ij} = -\delta a_{ij} sign\left(\hat{x}_j\right) t_i$, akkor az r definíciója alapján

$$r_i = \sum_{j=1}^{n} a_{ij}\hat{x}_j - b_i = -\sum_{j=1}^{n} e_{ij}\hat{x}_j + \varepsilon_i,$$

vagy ekvivalens formában

$$\sum_{j=1}^{n} \left(a_{ij} + e_{ij}\right)\hat{x}_j = b_i + \varepsilon_i.$$

A $\Delta A = [e_{ij}]_{i,j=1}^{n}$ mátrixra komponensenként fennáll, hogy $|e_{ij}| \leq \delta a_{ij}$, a $\Delta b = [\varepsilon_1, \ldots, \varepsilon_n]^T$ vektorra pedig fennáll, hogy $|\Delta b| \leq \delta b$. Ezzel a tételt igazoltuk. □

A tétel alkalmazásához nem kell a kondíciószám ismerete. A gyakorlatban δA és δb elemei a gépi epszilonnal arányosak.

6.2. Wilkinson tétele

Wilkinson igazolta, hogy *az $Ax = b$ egyenletrendszer Gauss-módszerrel lebegőpontos aritmetikában kapott $\hat{x}$ közelítő megoldása kielégíti az*

$$\left(A + \Delta A\right)\widehat{x} = b \tag{6.16}$$

egyenletrendszert, ahol

$$\left\|\Delta A\right\|_\infty \leq 8n^3\rho_n \left\|A\right\|_\infty u + O(u^2). \tag{6.17}$$

A ρ_n a pivot elemek *növekedési tényezője*. Minthogy a gyakorlatban ρ_n kicsi, a

$$\frac{\left\|\Delta A\right\|_\infty}{\left\|A\right\|_\infty} \leq 8n^3\rho_n u + O(u^2)$$

relatív inverz hiba is az. Ezért a Bunch tétel (8.6 Tétel) alapján a Gauss-elimináció gyengén stabil mind a teljes, mind pedig a parciális főelemválasztás esetén.

A Wilkinson tételből kapjuk, hogy

$$\mathrm{cond}_\infty(A)\frac{\left\|\Delta A\right\|_\infty}{\left\|A\right\|_\infty} \leq 8n^3\rho_n \mathrm{cond}_\infty\left(A\right)u + O\left(u^2\right).$$

Kis kondíciószám esetén feltehetjük, hogy $1 - \mathrm{cond}_\infty(A)\frac{\left\|\Delta A\right\|_\infty}{\left\|A\right\|_\infty} \approx 1$. A 8.5 Tétel felhasználásával ($\Delta b = 0$ eset) a direkt hibára az alábbi közelítő becslést kapjuk:

$$\frac{\left\|\Delta x\right\|_\infty}{\left\|x\right\|_\infty} \leq 8n^3\rho_n \mathrm{cond}_\infty\left(A\right)u. \tag{6.18}$$

Ez az ökölszabály helyességét támasztja alá a Gauss-módszer esetén.

Tekintsük a következő példát, amelynek együtthatóit pontosan tudjuk ábrázolni:

$$888445x_1 + 887112x_2 = 1,$$
$$887112x_1 + 885781x_2 = 0.$$

Itt $\text{cond}(A)_\infty$ ugyan nagy, de $\text{cond}_\infty(A)\frac{\|\Delta A\|_\infty}{\|A\|_\infty}$ elhanyagolható az 1 mellett. A feladat pontos megoldása $x_1 = 885781$, $x_2 = -887112$. A MATLAB által adott közelítő megoldás $\hat{x}_1 = 885827.23$, $\hat{x}_2 = -887158.30$, amelynek relatív hibája

$$\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} = 5.22 \times 10^{-5}.$$

Minthogy $s \approx 16$ és $\text{cond}(A)_\infty \approx 3.15 \times 10^{12}$ az eredmény lényegében megfelel a Wilkinson tételnek, ill. az ökölszabálynak. A Wilkinson tétel az inverz hiba mértékére a

$$\|\Delta A\|_\infty \leq 1.26 \times 10^{-8}$$

becslést adja. Az Oettli-Práger tételt a $\delta A = \epsilon_M |A|$ és $\delta b = \epsilon_M |b|$ választással alkalmazva kapjuk, hogy $|r(\hat{x})| \leq \delta A |\hat{x}| + \delta b$. Minthogy $\|\delta A\|_\infty = 3.94 \times 10^{-10}$, ez jobb becslést ad a $\|\Delta A\|_\infty$ inverz hibára, mint a Wilkinson-féle eredmény.

6.3. Utólagos hibabecslések

A valamilyen módszerrel kapott közelítő megoldás hibájának utólagos becslésére azért van szükség, hogy valamilyen adatunk legyen az eredmény megbizhatóságáról. A következőkben három ilyen módszert ismertetünk.

6.3.1. A direkt hiba becslése a reziduális segítségével

8.10 Tétel *(Auchmuty). Jelölje $\widehat{x}$ az $Ax = b$ egyenletrendszer valamilyen módon kiszámított közelítő megoldását. Ekkor igaz, hogy*

$$\|x - \hat{x}\|_2 = \frac{c\,\|r(\hat{x})\|_2^2}{\|A^T r(\hat{x})\|_2},$$

ahol $c \geq 1$ konstans, amely A-tól függ.

Bizonyítás. Tegyük fel, hogy A nemszinguláris, $\hat{x} \neq x = A^{-1}b$ és legyen

$$C_2(A) = \max_{\|y\|_2=1} \left\|A^T y\right\|_2 \left\|A^{-1} y\right\|_2.$$

Azt igazoljuk, hogy

$$\frac{\|r(\hat{x})\|_2^2}{\|A^T r(\hat{x})\|_2} \leq \|x-\hat{x}\|_2 \leq C_2(A)\,\frac{\|r(\hat{x})\|_2^2}{\|A^T r(\hat{x})\|_2},$$

amiből az állítás már következik. Definíció szerint $r(\hat{x}) = A\hat{x} - b = A(\hat{x}-x)$ és

$$\begin{aligned}\|r(\hat{x})\|_2^2 &= r(\hat{x})^T r(\hat{x}) = r(\hat{x})^T A(\hat{x}-x) = \left|\left(A^T r\left(\hat{x}\right)\right)^T(\hat{x}-x)\right| \leq \\ &\leq \left\|A^T r\left(\hat{x}\right)\right\|_2 \|\hat{x}-x\|_2 .\end{aligned}$$

Ezt osztva az $\left\|A^T r(\hat{x})\right\|_2$ mennyiséggel kapjuk az egyenlőtlenség baloldalát. A $C_2(A)$ mennyiség felírható az ekvivalens

$$C_2(A) = \max_{y\neq 0} \frac{\left\|A^T y\right\|_2 \|A^{-1}y\|_2}{\|y\|_2^2}$$

alakban is. Mármost az $y = r(\hat{x}) \neq 0$ helyettesítéssel kapjuk, hogy

$$c = \frac{\left\|A^T r(\hat{x})\right\|_2 \|A^{-1} r(\hat{x})\|_2}{\|r(\hat{x})\|_2^2} \leq C_2(A).$$

Átrendezéssel és az $A^{-1}r(\hat{x}) = \hat{x} - x$ összefüggés figyelembevételével kapjuk a bizonyítandó egyenlőtlenség jobboldalát. □

Megemlítjük, hogy a c hibakonstans értékét nem befolyásolja az $\hat{x}-x$ hibavektor nagysága, csak az iránya. Igaz továbbá, hogy

$$C_2(A) = \frac{1}{2}\left(\mathrm{cond}_2(A) + \frac{1}{\mathrm{cond}_2(A)}\right) \leq \mathrm{cond}_2(A).$$

A számítógépes tapasztalatok azt mutatják, hogy c nem túl nagy szám, gyakran $c \leq 100$. Az összefüggés alapján a reziduális és az $\|r(\hat{x})\|_2^2 / \left\|A^T r(\hat{x})\right\|_2$ hányados meghatározásával nagyságrendileg helyesen becsülhetjük a közelítő megoldás abszolút hibáját.

6.3.2. Az $\left\|A^{-1}\right\|$ LINPACK becslése

A LINPACK programcsomagban használják a következő eljárást $\|A^{-1}\|$ becslésére. Oldjuk meg rendre az $A^T y = d$, $Aw = y$ egyenletrendszereket! Az $\|A^{-1}\|$ becslése:

$$\left\|A^{-1}\right\| \approx \frac{\|w\|}{\|y\|} \quad \left(\leq \left\|A^{-1}\right\|\right). \tag{6.19}$$

Minthogy

$$\frac{\|w\|}{\|y\|} = \frac{\left\|A^{-1}\left(A^{-T}d\right)\right\|}{\|A^{-T}d\|},$$

az eljárás felfogható a sajátértékszámítási fejezetben ismertetésre kerülő hatványmódszer alkalmazásának. A becslés az $1, 2$ és ∞ normák esetén alkalmazható. A d vektor javasolt komponensei ± 1 értékűek. Az előjeleket lehet véletlenszerűen választani. A LINPACK rendszerben az előjelek megválasztása egy adaptív stratégiával történik.

Ha az $Ax = b$ egyenletrendszert az LU-módszerrel oldottuk meg, akkor a további egyenletrendszerek megoldása $O(n^2)$ régi flop rendszerenként. Így a LINPACK becslő eljárás költsége kicsi marad. Az $\|A^{-1}\|$ becslés segítségével $\mathrm{cond}(A)$ és a közelítő megoldás hibája már könnyen becsülhető (v.ö. a 8.5 Tétellel, vagy az ökölszabállyal). Megjegyezzük, hogy más hasonló eljárások is ismertek.

6.3.3. Az inverz hiba Oettli-Práger-féle becslése

Az inverz hiba becsléséhez az Oettli-Práger eredményt a következő formában szokás használni. Legyen $r(\hat{x}) = A\hat{x} - b$ a reziduális hiba, $E \in \mathbb{R}^{n\times n}$ és $f \in \mathbb{R}^n$ adottak úgy, hogy $E \geq 0$ és $f \geq 0$. Legyen

$$\omega = \max_i \frac{|r(\hat{x})_i|}{(E|\hat{x}| + f)_i},$$

ahol 0/0-át 0-nak, ρ/0-át pedig ∞-nek definiáljuk, ha $\rho \neq 0$. Az $(y)_i$ jelölés az y vektor i-edik komponensét jelöli. Ha $\omega \neq \infty$, akkor van egy ΔA mátrix és egy Δb vektor, amelyekkel fennáll

$$|\Delta A| \leq \omega E, \quad |\Delta b| \leq \omega f$$

és

$$(A + \Delta A)\,\hat{x} = b + \Delta b.$$

Továbbá ω a legkisebb olyan szám, amelyre a fenti tulajdonságú ΔA és Δb létezik. Az ω mennyiség az E és f mennyiségekben kifejezett relatív inverz hibát méri. Ha egy adott E, f és $\hat{x}$ esetén ω kicsi, akkor a perturbált probléma is (és ennek megoldása is) közel van az eredetihez (és ennek megoldásához). A gyakorlatban az $E = |A|$, $f = |b|$ választást preferálják.

6.4. A közelítő megoldás iteratív javítása

Az eljárás első ismert alkalmazása Fox, Goodwin, Turing és Wilkinson nevéhez fűződik (1946). Jelölje $\hat{x}$ az $Ax = b$ egyenletrendszer közelítő megoldását. Legyen

$r(y) = Ay - b$ az y pontbeli reziduális hiba. Az $\hat{x}$ közelítő megoldás pontosságát a következő iteratív eljárással lehet javítani.

AZ ITERATÍV JAVÍTÁS ALGORITMUSA:

$i = 1$, $x_1 = \hat{x}$
for $i = 1, \ldots$
$r = Ax_i - b$
Számítsuk ki az $Ad = r$ egyenletrendszer $\hat{d}$ közelítő megoldását az LU-módszer segítségével!
$x_{i+1} = x_i - \hat{d}$
Ha $\left\|\hat{d}\right\| / \|x_i\| < tol$, akkor vége.
end

Az eljárás különféle változatai ismertek. Az LU-módszer helyett más módszer is használható.

Legyen η az a legkisebb relatív inverz hibakorlát, amelyre

$$(A + \Delta A)\,\hat{x} = b + \Delta b, \quad |\Delta A| \le \eta\,|A|\,, \quad |\Delta b| \le \eta\,|b|\,.$$

Legyen továbbá

$$\sigma\,(A, x) = \max_i \left(|A|\,|x|\right)_i / \min_i \left(|A|\,|x|\right)_i, \quad \min_i \left(|A|\,|x|\right)_i > 0.$$

Igaz a következő

8.11 Tétel *(Skeel). Ha* $k_r\,(A^{-1})\,\sigma\,(A, x) \le c_1 < 1/\epsilon_M$, *akkor elég nagy* i *esetén fennáll, hogy*

$$(A + \Delta A)\,x_i = b + \Delta b, \quad |\Delta A| \le 4\eta\epsilon_M\,|A|\,, \quad |\Delta b| \le 4\eta\epsilon_M\,|b|\,. \tag{6.20}$$

Ez az eredmény gyakran az első iteráció után, az $i = 2$ értékre teljesül. Jankowski és Wozniakowski az iteratív javitást a Gauss-elimináció helyett tetszőleges olyan ϕ módszerre vizsgálták, amely az $Ax = b$ egyenletrendszer 1-nél kisebb relatív hibájú $\hat{x}$ közelítését állítja elő, azaz amelyre $\|\hat{x} - x\| \le q\,\|x\|$ $(q < 1)$. Igazolták, hogy az iteratív javítás még egyszeres pontosságú lebegőpontos aritmetikában is javítja a közelítő megoldás pontosságát és a ϕ módszert gyengén stabillá teszi.

6.5. Feladatok

1. Oldjuk meg az

$$\begin{bmatrix} 1 & -1 & 1 \\ -1 & 888445 & 887112 \\ 1 & 887112 & 885781 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1772893 \\ 1 \\ 0 \end{bmatrix}$$

egyenletrendszert és vizsgáljuk a megoldás hibáját.
2. Oldjuk meg az $Ax = b$ egyenletrendszereket a síma és a részleges főelemkiválasztásos Gauss módszerrel az alábbi adatok esetén. Határozzuk meg a kondíciószámokat és vizsgáljuk meg a kapott megoldások hibáját a fejezet módszereivel iteratív javítással és anélkül.

$$A = \begin{bmatrix} 3\epsilon & 2 & 1 \\ 2 & 2 & 2 \\ 1 & 2 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 3+3\epsilon \\ 6 \\ 2 \end{bmatrix}, \quad \epsilon \ll 1. \tag{6.21}$$

$$A = \begin{bmatrix} 3 & 2 & 1 \\ 2 \times 10^6 & 2 & 2 \\ 10^6 & 2 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 3+3\epsilon \\ 6 \\ 2 \end{bmatrix}, \tag{6.22}$$

$$A = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 2 \times 10^{-6} & 2 \times 10^{-6} \\ 1 & 2 \times 10^{-6} & -10^{-6} \end{bmatrix}, \quad b = \begin{bmatrix} 3 + 3 \times 10^{-6} \\ 6 \times 10^{-6} \\ 2 \times 10^{-6} \end{bmatrix}. \tag{6.23}$$

Mi a kapcsolata az egyes egyenletrendszereknek?
3. Legyen A 10×10-es mátrix és válasszuk prekondicionáló mátrixnak az A főátlójából és két mellékátlójából álló sávmátrixot. Mennyit javul a rendszer kondíciószáma, ha (i) A véletlen mátrix; (ii) A Hilbert-mátrix?

7. fejezet

A SAJÁTÉRTÉK-PROBLÉMA

A mátrixok sajátérték-feladatának megfogalmazásához szükségünk van a komplex elemű mátrixok és vektorok bevezetésére. A komplex elemű n-dimenziós oszlopvektorok halmazát $\mathbb{C}^n$-el jelöljük. Hasonlóképpen az $m \times n$ típusú komplex elemű mátrixok halmazát $\mathbb{C}^{m\times n}$ jelöli. Nyilvánvalóan fennáll, hogy $\mathbb{R}^n \subset \mathbb{C}^n$ és $\mathbb{R}^{m\times n} \subset \mathbb{C}^{m\times n}$. A valós elemű vektorokra és mátrixokra bevezetett műveletek és a determináns értelemszerűen ugyanazok maradnak a komplex esetben is. Ezekhez jön még két művelet: a konjugálás és a hermitikus transzponálás. Egy $A \in \mathbb{C}^{m\times n}$ mátrix konjugáltján az $\overline{A} = [\overline{a}_{ij}]_{i,j=1}^{m,n} \in \mathbb{C}^{m\times n}$ mátrixot értjük. Egy $A \in \mathbb{C}^{m\times n}$ hermitikus (vagy konjugált) transzponáltján az

$$A^H = \overline{A}^T \in \mathbb{C}^{n\times m}$$

mátrixot értjük. A valós mátrixok esetén a konjugálás a mátrixot nem változtatja meg. Ezért valós mátrixok esetén a hermitikus transzponált egybeesik a közönséges transzponálással. A komplex vektorok és mátrixok normájának definíciója az $\mathbb{R}^n$, illetve $\mathbb{R}^{n\times n}$ halmazok $\mathbb{C}^n$-re, ill. $\mathbb{C}^{n\times n}$-re történő cseréjétől eltekintve változatlan marad. Az 1. Fejezetben bevezetett konkrét normák esetén a valós szám abszolút értéke helyett a komplex szám abszolút értékét kell érteni. Az euklideszi és a Frobenius norma definíciója azonban az alábbiak szerint módosul

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}}, \tag{7.1}$$

$$\|A\|_F = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}}, \tag{7.2}$$

ahol a definíciókban szereplő abszolút értéken a komplex számok abszolút értékét kell érteni. Vegyük észre, hogy $x \in \mathbb{C}^n$ esetén $\|x\|_2 = \left(x^H x\right)^{1/2}$.

9.1 Definíció. *Legyen $A \in \mathbb{C}^{n\times n}$ tetszőleges mátrix. A $\lambda \in \mathbb{C}$ számot az A mátrix sajátértékének, az $x \in \mathbb{C}^n$ ($x \neq 0$) vektort pedig a λ sajátértékhez tartozó (jobboldali) sajátvektornak nevezzük, ha*

$$Ax = \lambda x. \tag{7.3}$$

A sajátvektor egy olyan vektor, amelyet az $x \to Ax$ leképezés a saját hatásvonalán hagy (irányítás, nagyság változhat). A sajátérték-feladat megoldása a sajátértékek és a hozzájuk tartozó sajátvektorok meghatározását jelenti.

Egy x sajátvektor t-szerese is sajátvektor $t \neq 0$ esetén, ui. $A(tx) = tAx = t\lambda x = \lambda(tx)$. Az $Ax = \lambda x$ egyenletrendszer átrendezéssel az ekvivalens $(A-\lambda I)x = 0$ alakra hozható. Ennek a homogén egyenletrendszernek akkor és csak akkor van zérustól különböző megoldása, ha $\det(A - \lambda I) = 0$. A

$$\phi(\lambda) = \det(A - \lambda I) = \det\left(\begin{bmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{bmatrix}\right) = 0 \tag{7.4}$$

egyenletet az A mátrix *karakterisztikus egyenletének* nevezzük. A determinánst kifejtve a λ változó n-ed fokú polinomját, azaz a

$$\phi(\lambda) = (-1)^n(\lambda^n - p_1\lambda^{n-1} - \ldots - p_{n-1}\lambda - p_n) \tag{7.5}$$

polinomot kapjuk. Az algebra alaptétele miatt az A $n \times n$ mátrixnak a multiplicitásokat is beleszámítva pontosan n sajátértéke van. Egy λ_k sajátértékhez tartozó lineárisan független sajátvektorok száma legfeljebb annyi, mint λ_k multiplicitása a $\phi(\lambda) = 0$ karakterisztikus egyenletben. Különböző sajátértékekhez tartozó sajátvektorok lineárisan függetlenek.

Legyen $A \in \mathbb{C}^{n\times n}$ tetszőleges. Megmutatható, hogy A^H sajátértékei A sajátértékeinek konjugáltjai. Legyen $\overline{\lambda}$ az A^H tetszőleges sajátértéke és $y \in \mathbb{C}^n$ egy ehhez tartozó sajátvektor: $A^H y = \overline{\lambda} y$. Mindkét oldal hermitikus transzponáltját véve kapjuk, hogy

$$y^H A = \lambda y^H.$$

Az $y^H \in \mathbb{C}^{1\times n}$ sorvektort az A mátrix λ sajátértékhez tartozó baloldali sajátvektorának nevezzük. Ha λ az A mátrix egyszeres sajátértéke, x és y a hozzátartozó jobb- és baloldali sajátvektorok, akkor $y^H x \neq 0$.

Legyen A valós mátrix és vegyük az $Ax = \lambda x$ ($A \in \mathbb{R}^{n\times n}$, $\lambda \in \mathbb{C}$, $x \in \mathbb{C}^n$) egyenlet mindkét oldalának komplex konjugáltját: $A\overline{x} = \overline{\lambda}\overline{x}$. Azt kapjuk, hogy ha λ az A komplex sajátértéke (és x a hozzátartozó komplex sajátvektor), akkor $\overline{\lambda}$ az A-nak szintén sajátértéke (és $\overline{x}$ a hozzátartozó sajátvektor). Tehát valós mátrixok

sajátértékei vagy valósak, vagy komplex konjugált sajátértékpárok lehetnek. Valós mátrixok komplex sajátértékeit és sajátvektorait valós aritmetika használata esetén csak speciális technikákkal kaphatjuk meg.

9.1 Tétel. *Legyen λ az A mátrix tetszőleges sajátértéke. Tetszőleges indukált mátrixnormában fennáll, hogy $|\lambda| \leq \|A\|$.*

Bizonyítás. $\|\lambda x\| = |\lambda| \, \|x\| = \|Ax\| \leq \|A\| \, \|x\|$, ahonnan $x \neq 0$ miatt $|\lambda| \leq \|A\|$ adódik. □

9.2 Tétel. *Legyen $p(z) = b_k z^k + \ldots + b_1 z + b_0$ tetszőleges polinom. Ha az A mátrix sajátértékei $\lambda_1, \lambda_2, \ldots, \lambda_n$, akkor a $p(A) = b_k A^k + \ldots + b_1 A + b_0 I$ mátrix sajátértékei: $p(\lambda_1), p(\lambda_2), \ldots, p(\lambda_n)$.*

Következmény. *Az A^k mátrix sajátértékei $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$.*

9.3 Tétel. *Tegyük fel, hogy a T $n \times n$ mátrixra* $\det(T) \neq 0$ *teljesül. Ekkor az A és $B = T^{-1}AT$ mátrix sajátértékei megegyeznek. Ha x az A sajátvektora, akkor $y = T^{-1}x$ a B sajátvektora.*

Bizonyítás. Definíció szerint

$$Ax = \lambda x \Leftrightarrow T^{-1}Ax = \lambda T^{-1}x \Leftrightarrow T^{-1}AT(T^{-1}x) = \lambda(T^{-1}x) \Leftrightarrow By = \lambda y,$$

ami bizonyítandó volt. □

Az $A \to T^{-1}AT$ leképezést *hasonlósági transzformációnak* nevezzük. Az A és B mátrix *hasonló*, ha van egy T mátrix úgy, hogy $B = T^{-1}AT$. A 9.3. Tétel tartalma az, hogy hasonló mátrixok sajátértékei megegyeznek.

9.2 Definíció. *Egy A mátrix diagonalizálható, ha van olyan T mátrix (*$\det(T) \neq 0$*), hogy a $T^{-1}AT$ hasonló mátrix diagonális.*

A diagonalizálhatóság pontos feltételét meglehetősen bonyolult ellenőrizni.

9.3 Definíció. *Azt a legalacsonyabb fokú $p(x)$ polinomot, amelyet A kielégít, azaz $p(A) = 0$, az A mátrix minimálpolinomjának nevezzük.*

A minimálpolinom konstans szorzótól eltekintve egyértelmű. Igaz a következő

9.4 Tétel. *Egy A mátrix akkor és csak akkor diagonalizálható, ha minimálpolinomjának minden gyöke egyszeres.*

A tételnek csak elméleti jelentősége van. Általános esetben a következő fontos eredmény igaz.

9.5 Tétel *(Jordan). Legyen A tetszőleges $n \times n$ mátrix. Létezik olyan nemszinguláris X mátrix és m természetes szám, hogy*

$$X^{-1}AX = \operatorname{diag}(J_1, \ldots, J_m) = \begin{bmatrix} J_1 & 0 & \ldots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & J_m \end{bmatrix}, \tag{7.6}$$

ahol

$$J_i = \begin{bmatrix} \lambda_k & 1 & 0 & \dots & \dots & 0 \\ 0 & \lambda_k & 1 & 0 & & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & 0 & \lambda_k & 1 \\ 0 & \dots & \dots & \dots & 0 & \lambda_k \end{bmatrix} \tag{7.7}$$

ún. Jordan blokk $n_i \times n_i$ *méretű és* $\sum_{i=1}^{m} n_i = n$. *A blokkok száma és mérete sorrendjüktől eltekintve egyértelmű.*

Az $X^{-1}AX = \text{diag}(J_1, \dots, J_m)$ előállítást *Jordan-féle normálalaknak* nevezzük. Az n_i blokkméret nem feltétlenül azonos λ_k multiplicitásával. Minden sajátérték szerepel legalább egy blokkban, de egy λ_k sajátértékhez több különböző méretű Jordan blokk is tartozhat. A blokkok mérete csak bonyolult vizsgálatokkal határozható meg. Diagonalizálható mátrixok esetén a J_i Jordan blokkok természetesen 1×1 mátrixok és $m = n$.

Az összes sajátérték és sajátvektor meghatározása elvileg sem könnyű feladat. A nehézségeket tovább fokozza az a tény, hogy a sajátértékek és sajátvektorok a mátrixelemek megváltozására erősen érzékenyek. Az eredeti A mátrix és a perturbált $A + \delta A$ mátrix sajátértékei jelentősen eltérhetnek egymástól és a sajátértékek multiplicitása is megváltozhat. A sajátérték-feladat érzékenységét a következő tételekkel és példákkal jellemezhetjük.

9.6 Tétel *(Ostrowski-Elsner). Az* $A \in \mathbb{C}^{n \times n}$ *mátrix minden* λ_i *sajátértékhez létezik a perturbált* $A + \delta A$ *mátrix egy olyan* μ_k *sajátértéke, hogy*

$$|\lambda_i - \mu_k| \leq (2n-1) \left(\|A\|_2 + \|A + \delta A\|_2 \right)^{1-\frac{1}{n}} \|\delta A\|_2^{\frac{1}{n}} . \tag{7.8}$$

A tétel azt mutatja, hogy a sajátértékek folytonosan változnak és azt, hogy a megváltozás mértéke arányos a δA perturbáció mértékének n-edik gyökével.

Példa. Vizsgáljuk meg egy μ sajátértékhez tartozó $r \times r$ méretű Jordan mátrix alábbi perturbációját:

$$\begin{bmatrix} \mu & 1 & 0 & \dots & 0 \\ 0 & \mu & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & \mu & 1 \\ \epsilon & 0 & \dots & 0 & \mu \end{bmatrix} .$$

A perturbált mátrixhoz tartozó karakterisztikus egyenlet $(\lambda - \mu)^r = \epsilon$, ahonnan az eredeti Jordan mátrix r-szeres μ sajátértéke helyett az r különböző

$$\lambda_s = \mu + \epsilon^{1/r} \left(\cos(2s\pi/r) + i \sin(2s\pi/r) \right) \quad (s = 0, \dots, r-1)$$

sajátértéket kapjuk. A sajátértékek megváltozásának mértéke $\epsilon^{1/r}$, amely megfelel a 9.6 Tétel állításának. Ha például $|\mu| \approx 1$, $r = 16$ és $\epsilon = \epsilon_M \approx 2.2204 \times 10^{-16}$ értékeket vesszük, akkor a sajátértékek eltérése ≈ 0.1051. Ez pedig a mátrix perturbációjához képest a sajátértékek egy igen jelentős mértékű megváltozását jelenti.

Speciális tulajdonságú mátrixok és perturbációk esetén a sajátértékek megváltozása az Ostrowski-Elsner tételben látottnál jóval kisebb is lehet.

9.7 Tétel *(Bauer-Fike). Legyen $A \in \mathbb{C}^{n\times n}$ diagonalizálható mátrix, $X^{-1}AX = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ és jelölje μ az $A + \delta A$ mátrix sajátértékét. Ekkor*

$$\min_{1\le i\le n} |\lambda_i - \mu| \le \mathrm{cond}_2(X)\, \|\delta A\|_2 . \tag{7.9}$$

Tehát diagonalizálható mátrix perturbációja esetén a sajátértékek megváltozásának mértéke arányos az általában ismeretlen X mátrix kondíciószámával és $\|\delta A\|_2$-val. Ez aszimptotikusan lényegesen jobb eredmény mint az Ostrowski-Elsner tétel, azonban ne felejtsük, hogy $\mathrm{cond}_2(X)$ igen nagy is lehet.

Az A mátrix sajátértékei folytonos függvényei a mátrix elemeinek. Ez a normált sajátvektorokra is igaz, ha a sajátértékek egyszeresek. A következő példa is mutatja, hogy az utóbbi állítás többszörös sajátértékek esetén már nem igaz.

Példa. Legyen

$$A(t) = \begin{bmatrix} 1 + t\cos\left(\frac{2}{t}\right) & -t\sin\left(\frac{2}{t}\right) \\ -t\sin\left(\frac{2}{t}\right) & 1 - t\cos\left(\frac{2}{t}\right) \end{bmatrix} \quad (t \neq 0).$$

Az $A(t)$ mátrix sajátértékei $\lambda_1 = 1 + t$ és $\lambda_2 = 1 - t$. A λ_1 sajátértékhez tartozó sajátvektor $[\sin(1/t), \cos(1/t)]^T$, a λ_2-höz tartozó pedig $[\cos(1/t), -\sin(1/t)]^T$. Ha $t \to 0$, akkor

$$A(t) \to I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \lambda_1, \lambda_2 \to 1,$$

míg a sajátvektorok sehova sem tartanak.

Legyen λ az $A \in \mathbb{C}^{n\times n}$ mátrix egyszeres sajátértéke, x és y pedig a hozzátartozó jobb- és baloldali sajátvektorok. Az $\widetilde{A} = A + E$ perturbált mátrixnak létezik pontosan egy $\widetilde{\lambda}$ sajátértéke, hogy

$$\widetilde{\lambda} = \lambda + \frac{y^H E x}{y^H x} + O\left(\|E\|_2^2\right). \tag{7.10}$$

Innen könnyen kapjuk, hogy

$$\left|\widetilde{\lambda} - \lambda\right| \le \frac{\|y\|_2 \|x\|_2}{|y^H x|} \|E\|_2 .$$

A

$$\nu = \nu(\lambda) = \frac{\|y\|_2 \|x\|_2}{|y^H x|}$$

mennyiséget a λ egyszeres sajátérték kondíciószámának nevezzük. Többszörös sajátértékek kondíciószáma nem véges. Igaz a következő

9.8 Tétel *(Bauer-Fike). Legyenek $A \in \mathbb{C}^{n\times n}$ sajátértékei egyszeresek és $X^{-1}AX = X^{-1}AX =$diag$(\lambda_1, \ldots, \lambda_n)$. Ekkor igaz, hogy*

$$1 \leq \nu(\lambda_i) \leq \frac{1}{2}\left(\mathrm{cond}_2(X) + \frac{1}{\mathrm{cond}_2(X)}\right). \tag{7.11}$$

A sajátértékek elhelyezkedésének geometriáját jellemzi a következő

9.9 Tétel *(Gersgorin). Legyen $A \in \mathbb{C}^{n\times n}$,*

$$r_i = \sum_{j=1, j\neq i}^{n} |a_{ij}| \quad (i = 1, \ldots, n) \tag{7.12}$$

és

$$D_i = \{z \in \mathbb{C} \,|\, |z - a_{ii}| \leq r_i\} \quad (i = 1, \ldots, n). \tag{7.13}$$

Ekkor az A mátrix minden λ sajátértékére fennáll, hogy

$$\lambda \in \cup_{i=1}^{n} D_i. \tag{7.14}$$

Bizonyítás. Legyen $v = [v_1, \ldots, v_n]^T$ a λ-hoz tartozó sajátvektor, i pedig az az index, amelyre fennáll, hogy $\|v\|_\infty = |v_i| > 0$. Az $Av = \lambda v$ egyenletrendszer i-edik egyenlete

$$\lambda v_i = a_{ii} v_i + \sum_{j=1, j\neq i}^{n} a_{ij} v_j,$$

ahonnan átrendezéssel

$$|\lambda v_i - a_{ii} v_i| = |\lambda - a_{ii}|\,|v_i| \leq \left|\sum_{j=1, j\neq i}^{n} a_{ij} v_j\right| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|\,|v_i|$$

adódik. Mindkét oldalt $|v_i|$-vel elosztva kapjuk, hogy

$$|\lambda - a_{ii}| \leq r_i.$$

Minthogy minden λ sajátérték benne van valamelyik D_i körlemezben, az összes sajátérték benne van az egyesítésükben. □

A sajátértékeket tartalmazó körök egy részének sugarát sok esetben csökkenthetjük, ha a Gersgorin tételt az A-hoz hasonló $S^{-1}AS$ mátrixra alkalmazzuk, ahol S egy alkalmasan megválasztott diagonális mátrix.

Példa. Az

$$A = \begin{bmatrix} 1+i & 2i & i/2 \\ 1/2 & 4 & i/2 \\ 1/2 & 1/2 & 6+i \end{bmatrix}$$

mátrix esetén $r_1 = 2.5$, $r_2 = r_3 = 1$. Legyen $S = \text{diag}(1, 1/2, 1/2)$. Az $S^{-1}AS$ mátrix esetén a megfelelő körsugarak: $r_1 = 1.25$, $r_2 = r_3 = 1.5$. Tehát egy körsugár csökkent, míg a többiek növekedtek.

Tekintsük a $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ diagonális mátrix $D + E$ alakú perturbációját. A Gersgorin tétel miatt a $D + E$ mátrix tetszőleges μ sajátértéke benne van valamelyik körlemezben. Ha ez az i-edik, akkor

$$|\mu - \lambda_i - e_{ii}| \leq \sum_{j=1,\ j\neq i}^{n} |e_{ij}| .$$

Az $|a - b| \leq |a| + |b|$ egyenlőtlenség miatt ebből következik, hogy

$$|\mu - \lambda_i| \leq \sum_{j=1}^{n} |e_{ij}| \leq \max_i \left(\sum_{j=1}^{n} |e_{ij}| \right) = \|E\|_\infty .$$

Kimondhatjuk tehát, hogy $\min_{1\leq i\leq n} |\mu - \lambda_i| \leq \|E\|_\infty$. Legyen az $A \in \mathbb{C}^{n\times n}$ mátrix diagonalizálható, azaz $X^{-1}AX = D = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Az A mátrix $A+E$ perturbációjára fennáll, hogy

$$A + E = XDX^{-1} + E = X\left(D + X^{-1}EX\right)X^{-1}.$$

Minthogy $A + E$ és $D + X^{-1}EX$ hasonlóak, sajátértékeik is megegyeznek. $D + X^{-1}EX$ a D diagonális mátrix perturbációja. Ezért az előző okfejtés alapján $A+E$ tetszőleges μ sajátértékére fennáll, hogy

$$\min_{1\leq i\leq n} |\mu - \lambda_i| \leq \left\|X^{-1}EX\right\|_\infty \leq \text{cond}_\infty (X) \|E\|_\infty . \tag{7.15}$$

A kapott egyenlőtlenség pedig nem más mint a Bauer-Fike tétel ∞-normában.

Végül a közelítő sajátérték-sajátvektor párok jóságának megítélésével foglalkozunk. Legyen μ az A mátrix közelítő sajátértéke és x a hozzá tartozó közelítő sajátvektor, amelyre fennáll, hogy $\|x\|_2 = 1$. Azt vizsgáljuk, hogy mennyire teljesül az $Ax = \mu x$ egyenlet. Jelölje $r = Ax - \mu x$ a közelítés reziduális hibáját. Ha $r = 0$, akkor μ és x pontos sajátérték, ill. sajátvektor. Ha $r \neq 0$, akkor kérdés, hogy $\|r\|_2$ kicsinysége mond-e valamit a közelítés pontosságáról. Sajnos, általában semmire sem lehet következtetni. Tekintsük az

$$A(\epsilon) = \begin{bmatrix} 1 & 1 \\ \epsilon & 1 \end{bmatrix}$$

mátrixot, ahol $\epsilon \approx 0$ kicsi. Az $A(\epsilon)$ mátrix sajátértékei $1 \pm \sqrt{\epsilon}$. Legyen $\mu = 1$ és $x = [1, 0]^T$. Ekkor a közelítő sajátérték hibája $\pm\sqrt{\epsilon}$ és

$$\|r\|_2 = \left\| \begin{bmatrix} 0 \\ \epsilon \end{bmatrix} \right\|_2 = \epsilon.$$

Ha most például $\epsilon = 10^{-10}$, akkor a reziduális öt nagyságrenddel alulbecsli a tényleges 10^{-5} hibát.

Legyen $x \neq 0$ és $\mu \in \mathbb{C}$ tetszőleges, $r = Ax - \mu x$. Ekkor az

$$E = -\frac{rx^H}{x^H x} \tag{7.16}$$

mátrix olyan, hogy $(A + E)x = \mu x$ és $\|E\|_2 = \|r\|_2 / \|x\|_2$. Tehát E a (μ, x) sajátérték-sajátvektor közelítés inverz hibája. Ha μ a λ egyszeres sajátérték közelítése, akkor még az is fennáll, hogy

$$|\lambda - \mu| \cong \nu(\lambda) \|E\|_2 . \tag{7.17}$$

Az előző számpéldában valójában a $\nu(1 \pm \sqrt{\epsilon})$ kondíciószám nagy és ez okozta az öt nagyságrendű alulbecslést. Az $1 \pm \sqrt{\epsilon}$ sajátértékhez tartozó bal-, illetve jobboldali sajátvektorok $[1, \pm 1/\sqrt{\epsilon}]^T$, ill. $[1, \pm\sqrt{\epsilon}]^T$. Ennek megfelelően $\nu(1 \pm \sqrt{\epsilon}) = \frac{1}{2}\sqrt{2 + \epsilon + \frac{1}{\epsilon}}$. Az $\epsilon = 10^{-10}$ érték esetén $\nu(\lambda_{1,2}) \approx 5 \times 10^{-4}$. Ez nagyjából megfelel az alulbecslés mértékének. Vegyük még észre, hogy $\epsilon \to 0$ esetén a $\nu(1 \pm \sqrt{\epsilon})$ kondíciószám végtelenhez tart. A határértékként kapott $A(0)$ mátrixnak az 1 kétszeres sajátértéke.

Legyen az $A \in \mathbb{C}^{n \times n}$ mátrix közelítő sajátvektora $x \neq 0$. Azt vizsgáljuk, hogy milyen μ közelítő sajátérték minimalizálja az $\|r\|_2 = \|Ax - \mu x\|_2$ reziduális hibát és ennek következtében az $E = -rx^H / \left(x^H x\right)$ inverz hibát. Fennáll, hogy

$$\|r\|_2^2 = \|Ax\|_2^2 + \left| \mu \|x\|_2 - \frac{x^H Ax}{\|x\|_2} \right|^2 - \frac{\left|x^H Ax\right|^2}{\|x\|_2^2} \geq \|Ax\|_2^2 - \frac{\left|x^H Ax\right|^2}{\|x\|_2^2}.$$

Az egyenlőség pontosan akkor teljesül, ha $\mu \|x\|_2 - \frac{x^H Ax}{\|x\|_2} = 0$. Tehát a reziduális hibát a

$$\mu = R(x) = \frac{x^H Ax}{x^H x} \tag{7.18}$$

érték minimalizálja. Az $R(x)$ értéket Rayleigh-féle hányadosnak nevezük. Ha x a λ sajátértékhez tartozó pontos sajátvektor, akkor $R(x) = \lambda$.

7.1. Feladatok

1. Tekintsük a következő 10×10-es mátrixot

$$A(\varepsilon) = \begin{bmatrix} 10 & 10 & & & & & \\ & 9 & 10 & & & & \\ & & 8 & 10 & & & \\ & & & \ddots & \ddots & & \\ & & & & 2 & 10 & \\ \varepsilon & & & & & 1 & \end{bmatrix}.$$

Vizsgáljuk meg a sajátértékek megváltozását $\varepsilon = 0, 10^{-5}, 10^{-6}, 10^{-7}$ esetén!

2. Vizsgáljuk meg a Bauer-Fike tétel becslését az 1. Feladat $A = A(0)$ mátrixához képest!

3. Legyen $a, b \in \mathbb{R}^n$. Az $\|A\|_2 = \left[\lambda_{\max}(A^T A)\right]^{1/2}$ definíció alapján igazoljuk, hogy $\left\|ab^T\right\|_2 = \|a\|_2 \|b\|_2$! (Az állítást az 1.3 fejezet egyik példájaként más módon már beláttuk.)

8. fejezet

SAJÁTÉRTÉK-PROBLÉMÁK ITERATÍV MEGOLDÁSA

Csak valós elemű mátrixok valós sajátértékeit és sajátvektorait vizsgáljuk. A módszerek értelemszerű módosításokkal kiterjeszthetők a komplex esetre is.

8.1. A hatványmódszer

A von Mieses-től származó módszer alapgondolata a következő. Tegyük fel, hogy az A $n \times n$ típusú valós mátrixnak pontosan n különböző valós sajátértéke van. Ekkor a $\lambda_1, \ldots, \lambda_n$ sajátértékekhez tartozó $x_1, \ldots, x_n$ sajátvektorok lineárisan függetlenek. Tegyük fel, hogy a sajátértékek kielégítik a

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$$

feltételt és legyen $v^{(0)} \in \mathbb{R}^n$ adott! A sajátvektorok lineáris függetlensége miatt $v^{(0)}$ egyértelműen előáll $v^{(0)} = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n$ alakban. Tegyük fel, hogy $\alpha_1 \neq 0$. Képezzük a $v^{(k)} = Av^{(k-1)} = A^k v^{(0)}$ $(k = 1, 2, \ldots)$ sorozatot! A kiinduló feltevések miatt

$$\begin{aligned} v^{(1)} = Av^{(0)} &= A(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n) \\ &= \alpha_1 \lambda_1 x_1 + \alpha_2 \lambda_2 x_2 + \ldots + \alpha_n \lambda_n x_n, \end{aligned}$$

illetve

$$\begin{aligned} v^{(k)} = Av^{(k-1)} &= A(\alpha_1 \lambda_1^{k-1} x_1 + \alpha_2 \lambda_2^{k-1} x_2 + \ldots + \alpha_n \lambda_n^{k-1} x_n) \\ &= \alpha_1 \lambda_1^k x_1 + \alpha_2 \lambda_2^k x_2 + \ldots + \alpha_n \lambda_n^k x_n \\ &= \lambda_1^k \left(\alpha_1 x_1 + \alpha_2 \left(\tfrac{\lambda_2}{\lambda_1}\right)^k x_2 + \ldots + \alpha_n \left(\tfrac{\lambda_n}{\lambda_1}\right)^k x_n \right). \end{aligned}$$

Legyen $y \in \mathbb{R}^n$ tetszőleges olyan vektor, amelyre $y^T x_1 \neq 0$. Ekkor

$$\frac{y^T A v^{(k)}}{y^T v^{(k)}} = \frac{y^T v^{(k+1)}}{y^T v^{(k)}} = \frac{\lambda_1^{k+1}\left(\alpha_1 y^T x_1 + \sum_{i=2}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^{k+1} y^T x_i\right)}{\lambda_1^k\left(\alpha_1 y^T x_1 + \sum_{i=2}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^{k} y^T x_i\right)} \to \lambda_1,$$

ha $k \to \infty$. Minthogy $|\lambda_k/\lambda_1| \leq |\lambda_2/\lambda_1| < 1$ $(k \geq 2)$ miatt

$$\frac{v^{(k)}}{\lambda_1^k} \to \alpha_1 x_1$$

és $\left\|v^{(k)}\right\|_\infty / (\left|\lambda_1^k\right| \|\alpha_1 x_1\|_\infty) \to 1$, a λ_1 sajátértékhez tartozó sajátvektort is megkaphatjuk. Ha ugyanis az $v^{(k)}$ vektor helyett az egységnormájú $v^{(k)}/\left\|v^{(k)}\right\|_\infty$ vektort képezzük, akkor fennáll, hogy

$$\frac{v^{(k)}}{\|v^{(k)}\|_\infty} \sim \frac{\lambda_1^k \alpha_1 x_1}{\left|\lambda_1^k\right| \|\alpha_1 x_1\|_\infty} = sign\,(\lambda_1)^k\, sign\,(\alpha_1)\, \frac{x_1}{\|x_1\|_\infty} = z_k.$$

A z_k vektor sorozat minden eleme A sajátvektora. Negatív λ_1 esetén azonban a z_k és az őt megközelítő $v^{(k)}/\left\|v^{(k)}\right\|_\infty$ vektorsorozatok irányítása váltakozó lesz. Legyen $v^{(k)} = \left[v_1^{(k)}, \ldots, v_n^{(k)}\right]^T$. Ekkor az algoritmus alakja a következő.

A HATVÁNYMÓDSZER ALGORITMUSA ($v^{(0)} \in \mathbb{R}^n$ input adat):

for $k = 1, 2, \ldots$

$z^{(k)} = A v^{(k-1)}$

$\gamma_k = y^T z^{(k)} / y^T v^{(k-1)}$, $\quad (y^T \in \mathbb{R}^n$, lépésenként változhat, $y^T v^{(k-1)} \neq 0)$

$v^{(k)} = z^{(k)} / \left\|z^{(k)}\right\|_\infty$

end

A fentiek alapján fennáll, hogy

$$v^{(k)} \to x_1, \quad \gamma_k \to \lambda_1. \tag{8.1}$$

A $v^{(k)} \to x_1$ konvergencián itt azt értjük, hogy $v^{(k)}$ hatásvonala tart x_1 hatásvonalához. Az y vektort általában egységvektornak választjuk úgy, hogy ha $\left|\left(v^{(k)}\right)_i\right| = \left\|v^{(k)}\right\|_\infty$, akkor legyen $y = e_i$. Ha az $y = v^{(k-1)}$ választást használjuk, akkor $\gamma_k = v^{(k-1)T} A v^{(k-1)} / \left(v^{(k-1)T} v^{(k-1)}\right)$ a $v^{(k)}$ vektorhoz tartozó $R\left(v^{(k-1)}\right)$ Rayleigh hányadossal azonos. Az előző szakaszban már láttuk, hogy ez a választás adja a λ_1 sajátérték minimális reziduális hibájú közelítését.

Az eljárás a $|\lambda_2/\lambda_1|$ nagyságrendtől függő konvergencia sebességgel rendelkezik. A módszer erősen érzékeny a $v^{(0)}$ kezdővektor megválasztására is. Ha $\alpha_1 = 0$, akkor az eljárás nem konvergál a λ_1 domináns sajátértékhez. Bizonyos mátrixosztályok esetén igazolták, hogy véletlenül választott $v^{(0)}$ kezdővektorok esetén 1

valószínűséggel konvergál az eljárás. Komplex sajátértékek, illetve többszörös λ_1 esetén az eljárást módosítani kell. Az eljárás konvergenciáját gyorsítani lehet, ha az $A-\sigma I$ ún. eltolt mátrixra alkalmazzuk, ahol σ alkalmasan megválasztott szám. Az $A-\sigma I$ mátrix sajátértékei ui. $\lambda_1-\sigma, \lambda_2-\sigma, \ldots, \lambda_n-\sigma$ és a megfelelő konvergencia tényező: $|\lambda_2-\sigma| / |\lambda_1-\sigma|$. Ez utóbbi σ ügyes megválasztásával kisebbé tehető mint $|\lambda_2/\lambda_1|$.

A hatványmódszert az

$$\|E_k\|_2 = \frac{\|r_k\|_2}{\|v^{(k)}\|_2} = \frac{\left\|Av^{(k)} - \gamma_k v^{(k)}\right\|_2}{\|v^{(k)}\|_2} \leq \epsilon$$

kilépési feltétellel szokás leállítani. Ha a hatványmódszert szimultán alkalmazzuk az A^T mátrixra és $w_k = \left(A^T\right)^k w_0$, akkor

$$\nu(\lambda_1) \approx \frac{\left\|w^{(k)}\right\|_2 \left\|v^{(k)}\right\|_2}{|w_k^T v_k|}$$

a λ_1 kondíciószámának egy becslését adja. Ekkor értelemszerűen a

$$\nu(\lambda_1) \|E_k\|_2 \leq \epsilon$$

kilépési feltételt használjuk.

A hatványmódszert, amely igen előnyös lehet nagyméretű ritka mátrixok esetén, leginkább a legnagyobb, ill. a legkisebb abszolút értékű sajátértékek meghatározására használjuk. Ez utóbbi a következőképpen történhet. Az A^{-1} sajátértékei: $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}$. Ezek közül a legnagyobb abszolút értékű sajátérték $\frac{1}{\lambda_n}$ lesz. Alkalmazzuk a hatványmódszert A^{-1}-re a következő formában:

AZ INVERZ HATVÁNYMÓDSZER ($v^{(0)} \in \mathbb{R}^n$, $y \in \mathbb{R}^n$):

for $k = 1, 2, \ldots$

Oldjuk meg az $Az^{(k)} = v^{(k-1)}$ egyenletrendszert $z^{(k)}$-ra!

$\gamma_k = y^T z^{(k)} / y^T v^{(k-1)}$

$v^{(k)} = z^{(k)} / \left\|z^{(k)}\right\|_\infty$

end

Nyilvánvaló, hogy alkalmas feltételek mellett $\gamma_k \to 1/\lambda_n$ és $v^{(k)} \to x_n$. Az $Az^{(k)} = v^{(k-1)}$ egyenletrendszer megoldásához az LU-módszert célszerű használni. Az algoritmus szembetűnő előnye, hogy nem kell A^{-1}-et meghatározni.

Ha az inverz hatványmódszert az eltolt $A-\mu I$ mátrixra alkalmazzuk, akkor $(A-\mu I)^{-1}$ sajátértékei $(\lambda_i-\mu)^{-1}$. Ha μ közelít, mondjuk λ_t-hez, akkor $\lambda_i - \mu \to \lambda_i - \lambda_t$. Ezért az eltolt mátrix sajátértékeire teljesül, hogy

$$|\lambda_t - \mu|^{-1} > |\lambda_i - \mu|^{-1} \quad (i \neq t).$$

A konvergencia sebességét pedig a

$$q = |\lambda_t - \mu| \,/\, \{\max |\lambda_i - \mu|\}$$

hányados határozza meg. Ha μ elég közel van λ_t-hez, akkor q kicsinysége miatt az inverz hatványiteráció rendkívül sebesen fog konvergálni. Ezt a tulajdonságot használhatjuk ki közelítő sajátvektorok meghatározásánál, ha egy sajátérték μ közelítése ismert. Ekkor feltéve, hogy $\det(A - \mu I) \neq 0$, az $A - \mu I$ mátrixra alkalmazzuk az inverz hatványmódszert. Annak ellenére, hogy $A - \mu I$ közel szinguláris mátrix és ezért $(A - \mu I)\, z^{(k)} = v^{(k)}$ nem oldható meg nagy pontossággal, az eljárás sok esetben igen jó sajátvektor közelítést ad.

Végül megjegyezzük, hogy rangszámcsökkentő eljárásokkal és egyéb módosításokkal a Mieses eljárás alkalmassá tehető az összes sajátérték-sajátvektor meghatározására is.

8.2. Ortogonalizálási eljárások

Szükségünk van a következő definícióra.

10.1 Definició. *Egy Q $n \times n$ mátrix ortogonális, ha $Q^T Q = I$.*

10.1 Tétel. $\|x\|_2 = \|Qx\|_2$.

Bizonyítás. $\|x\|_2^2 = x^T x = x^T Q^T Q x = (Qx)^T Qx = \|Qx\|_2^2$. □

Tehát ortogonális mátrixszal való szorzás nem változtatja meg a vektor euklideszi hosszát.

10.2 Tétel *(QR-felbontás). Minden A $n \times n$ mátrix felbontható $A = QR$ alakban, ahol Q ortogonális mátrix, azaz $Q^T Q = I$ és R felső háromszögmátrix.*

A QR-felbontás az LU-módszerhez hasonló eljárást tesz lehetővé. Ha ismert az A nemszinguláris mátrix QR-felbontása, akkor az $Ax = b$ lineáris egyenletrendszert megoldhatjuk a következőképpen is:

$$Ax = QRx = b \Leftrightarrow Rx = Q^T b.$$

Ekkor tehát elég megoldani a felső háromszögmátrixú

$$Rx = Q^T b$$

egyenletrendszert.

Egy mátrix QR-felbontására több módszer is létezik. Ezek közül a Gauss-elimináción alapuló módszert és a Gram-Schmidt eljárást ismertetjük.

Legyen A nemszinguláris mátrix és tekintsük a $B = A^T A = R^T R$ Cholesky-felbontást, ahol R felső háromszögmátrix. Ekkor a $Q = AR^{-1}$ mátrix ortogonális, az $A = QR$ pedig az A mátrix QR-felbontása. A Q mátrix ortogonális, mert

$Q^TQ = (R^{-1})^T A^T A R^{-1} = (R^{-1})^T R^T R R^{-1} = I$. Az észrevétel alapján a QR-felbontást a következőképpen lehet végrehajtani:

(1) Számítsuk ki az $A^TA = R^TR$ Cholesky-felbontást!

(2) Invertáljuk az R mátrixot az $RX = I$ egyenletrendszer megoldásával!

(3) Számítsuk ki az $Q = AR^{-1}$ mátrixot!

A kapott Q és R mátrixok az A mátrix QR-felbontását adják.

A most bemutatott eljárásnak elsősorban elvi jelentősége van. A gyakorlatban a Givens- és Householder-módszereket, valamint az MGS módszert használják.

Az MGS vagy módosított Gram-Schmidt eljárás az önmagában is rendkívül fontos klasszikus Gram-Schmidt (CGS) ortogonalizációs eljárás numerikusan stabilizált ekvivalens változata. Legyen

$$\mathcal{R}\{a_1,\dots,a_m\} = \left\{ \sum_{j=1}^{m} \lambda_j a_j \mid \lambda_j \in \mathbb{R},\ j = 1,\dots,m \right\}$$

az $a_1,\dots,a_m \in \mathbb{R}^n$ $(m \le n)$ vektorok által kifeszített lineáris altér. Az ortogonalizációs feladat a következő: Legyen $a_1,\dots,a_m$ lineárisan független. Keressük azon lineárisan független $q_1,\dots,q_m$ vektorokat, amelyekre fennáll: $q_i^Tq_j = 0\ (i \ne j)$, $\|q_i\|_2 = 1\ (i = 1,\dots,m)$ és $\mathcal{R}\{a_1,\dots,a_m\} = \mathcal{R}\{q_1,\dots,q_m\}$. A $\{q_i\}_{i=1}^m$ vektorrendszert *ortonormált* rendszernek mondjuk. A Gram-Schmidt eljárás alapgondolata a következő.

Adjuk meg a q_1, q_2 vektorokat a következőképpen: $q_1 = a_1/\|a_1\|_2$ és $\tilde{q}_2 = a_2 - r_{12}q_1$. Az r_{12} vektort pedig válasszuk meg úgy, hogy $\tilde{q}_2 \perp q_1$ legyen:

$$\tilde{q}_2 \perp q_1 \Leftrightarrow (a_2 - r_{12}q_1)^T q_1 = a_2^Tq_1 - r_{12} = 0 \Rightarrow r_{12} = a_2^Tq_1.$$

Legyen $r_{22} = \|\tilde{q}_2\|_2 = \left\|a_2 - \left(a_2^Tq_1\right)q_1\right\|_2$, és $q_2 = \tilde{q}_2/r_{22}$. Eddig kaptuk a q_1 és q_2 vektorokat, amelyek ortonormáltak. Igazolnunk kell, hogy $\mathcal{R}\{a_1,a_2\} = \mathcal{R}\{q_1,q_2\}$. Világos, hogy $q_1, q_2 \in \mathcal{R}\{a_1,a_2\}$ és így $\mathcal{R}\{q_1,q_2\} \subset \mathcal{R}\{a_1,a_2\}$. Minthogy $a_1 \in \mathcal{R}\{q_1\}$ és $a_2 = r_{22}q_2 + r_{12}q_1 \in \mathcal{R}\{q_1,q_2\}$, igazoltuk az $\mathcal{R}\{a_1,a_2\} = \mathcal{R}\{q_1,q_2\}$ állítást. Tegyük most fel, hogy előállítottuk a $q_1,\dots,q_{k-1}$ ortonormált vektorokat, amelyekre teljesül, hogy

$$\mathcal{R}\{a_1,\dots,a_{k-1}\} = \mathcal{R}\{q_1,\dots,q_{k-1}\}.$$

Keressük a $\tilde{q}_k$ vektort

$$\tilde{q}_k = a_k - \sum_{j=1}^{k-1} r_{jk}q_j$$

alakban úgy, hogy $\tilde{q}_k \perp q_j\ (j = 1,\dots,k-1)$. Ez akkor és csak akkor teljesül, ha

$$\tilde{q}_k^Tq_i = a_k^Tq_i - \sum_{j=1}^{k-1} r_{jk}q_j^Tq_i = 0 \quad (i = 1,\dots,k-1).$$

Mivel $i \neq j$ esetén $q_j^T q_i = 0$ és $q_i^T q_i = 1$, kapjuk, hogy

$$r_{ik} = a_k^T q_i \quad (i = 1, \ldots, k-1)\,.$$

A kiinduló vektorok lineárisan függetlenek, ezért $\tilde{q}_k \neq 0$ és $r_{kk} = \|\tilde{q}_k\|_2 \neq 0$. Legyen $q_k = \tilde{q}_k / r_{kk}$. Igazoljuk, hogy $\mathcal{R}\{a_1, \ldots, a_k\} = \mathcal{R}\{q_1, \ldots, q_k\}$. Tekintve, hogy $\mathcal{R}\{a_1, \ldots, a_{k-1}\} = \mathcal{R}\{q_1, \ldots, q_{k-1}\}$, a $q_k \in \mathcal{R}\{a_1, \ldots, a_k\}$ tulajdonság is igaz. Fordítva

$$a_k = \sum_{j=1}^{k-1} r_{jk} q_j + r_{kk} q_k$$

miatt $a_k \in \mathcal{R}\{q_1, \ldots, q_k\}$. Következésképpen $\mathcal{R}\{a_1, \ldots, a_k\} = \mathcal{R}\{q_1, \ldots, q_k\}$. Az eljárást a következő formában algoritmizálhatjuk.

CGS (klasszikus Gram-Schmidt) ELJÁRÁS:

Adottak az $a_1, \ldots, a_m \in \mathbb{R}^n$ lineárisan független vektorok, $m \leq n$.

for $k = 1 : m$
  for $i = 1 : k-1$
    $r_{ik} = a_k^T a_i$
    $a_k = a_k - r_{ik} a_i$
  end
  $r_{kk} = \|a_k\|_2$
  $a_k = a_k / r_{kk}$
end

Az eljárás felülírja az a_i vektorokat az ortonormalizált q_i vektorokkal. A QR-felbontással való kapcsolatot a $a_k = \sum_{j=1}^{k-1} r_{jk} q_j + r_{kk} q_k$ összefüggés adja meg. Ugyanis fennáll, hogy

$$\begin{aligned} a_1 &= q_1 r_{11} \\ a_2 &= q_1 r_{12} + q_2 r_{22} \\ a_3 &= q_1 r_{13} + q_2 r_{23} + q_3 r_{33} \\ &\vdots \\ a_m &= q_1 r_{1m} + q_2 r_{2m} + \ldots + q_m r_{mm} \end{aligned} \quad .$$

Ez felírható a tömörebb

$$[a_1, \ldots, a_m] = \underbrace{[q_1, \ldots, q_m]}_{Q} \underbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} & \ldots & r_{1m} \\ 0 & r_{22} & r_{23} & \ldots & r_{2m} \\ 0 & 0 & r_{33} & \ldots & r_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & r_{mm} \end{bmatrix}}_{R}$$

alakban is. Kaptuk tehát a következő eredményt.

10.3 Tétel *(QR-felbontás). Minden $A \in \mathbb{R}^{n\times m}$ mátrix, amelynek rangja m, előáll $A = QR$ alakban, ahol $Q \in \mathbb{R}^{n\times m}$ oszlopai ortonormáltak, rangja m és $R \in \mathbb{R}^{m\times m}$ nemszinguláris felső háromszögmátrix.*

Ha $m = n$, akkor a QR-felbontásra vonatkozó 10.2 Tételt kapjuk nemszinguláris A esetén. Ekkor ugyanis $Q^TQ = \left[q_i^T q_j\right]_{i,j=1}^n = I$.

A numerikusan stabilizált MGS módszer a következőképpen adható meg.

MGS (módosított Gram-Schmidt) ALGORITMUS:

Adottak az $a_1, \ldots, a_m \in \mathbb{R}^n$ lineárisan független vektorok.

for $k = 1 : m$
  $r_{kk} = \|a_k\|_2$
  $a_k = a_k / r_{kk}$
  for $j = k + 1 : m$
    $r_{kj} = a_j^T a_k$
    $a_j = a_j - r_{kj} a_k$
  end
end

Az eljárás felülírja az a_i vektorokat az ortonormalizált q_i vektorokkal. Az MGS eljárás ekvivalens a CGS eljárással. Ugyanakkor numerikusan lényegesen stabilabb. Björck igazolta, hogy $m = n$ esetén a számított $\hat{Q}$ mátrixra fennáll, hogy

$$\hat{Q}^T\hat{Q} = I + E, \qquad \|E\|_2 \cong \operatorname{cond}(A)\, u, \tag{8.2}$$

ahol u az egységnyi kerekítés mértéke.

8.3. A QR-módszer

A ma használt legfontosabb általános eljárás az összes sajátérték meghatározására. Megmutatható, hogy a hatványmódszer általánosítása. Legyen $A_1 = A$ és $A_1 = Q_1R_1$. Képezzük az $A_2 = R_1Q_1$ mátrixot! Az A_2 mátrix hasonló az A_1 mátrixhoz, ui. $R_1 = Q_1^{-1}A_1$ és $A_2 = R_1Q_1 = Q_1^{-1}AQ_1$. Hasonlóképpen az

$$A_k = R_{k-1}Q_{k-1} = Q_kR_k \quad (k = 2, 3, \ldots)$$

mátrixsorozat minden mátrixa hasonló az A mátrixhoz. A QR-módszer a fenti sorozat képzéséből áll.

A QR-MÓDSZER ALGORITMUSA:

for $i = 1, \ldots$
  Számítsuk ki az $A_i = Q_iR_i$ felbontást!
  $A_{i+1} = R_iQ_i$
end

Igaz a következő tétel.

10.4 Tétel *(Parlett). Ha az A mátrix diagonalizálható, sajátértékeire fennáll*

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0$$

és az $X^{-1}AX = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ hasonlósági transzformáció X mátrixának létezik LU-felbontása, akkor az A_k mátrixok alsó háromszög része konvergál egy diagonális mátrixhoz, amelynek diagonális elemei az A sajátértékei lesznek.

A 10.4 Tételnél lényegesen jobb konvergencia eredmények is ismertek. Az A_k mátrixok felső része nem feltétlenül konvergál egy meghatározott mátrixhoz. Ha az A mátrixnak p számú azonos abszolút értékű sajátértéke van, akkor az A_k mátrixok sorozata egy

$$\begin{bmatrix}
\times & & & & & & & & \times \\
0 & \ddots & & & & & & & \\
 & & \times & & & & & & \\
0 & & 0 & * & \ldots & * & & & \\
 & & & \vdots & & \vdots & & & \\
 & & & * & \ldots & * & & & \\
0 & & & & & 0 & \times & & \\
 & & & & & & & \ddots & \\
0 & & & & & & & 0 & \times
\end{bmatrix}$$

alakú mátrixhoz közelít, ahol a $*$ elemekkel jelölt részmátrixok elemei nem konvergálnak, sajátértékeik viszont igen. Ezt a részmátrixot lehet azonosítani és alkalmas módon kezelni. Ha valós mátrixunk van, akkor a karakterisztikus egyenletnek valós vagy komplex konjugált gyökei lehetnek. Komplex konjugált gyökpárok esetén p legalább kettő. Tehát az A_k sorozat a most vázolt jelenséget fogja mutatni.

A QR-felbontás nagyon számításigényes, költsége $O(n^3)$ régi flop. Ugyanakkor a QR-módszert rendkívül gazdaságosan lehet alkalmazni, ha a kiinduló A mátrix felső Hessenberg-alakú.

10.2 Definíció. *Egy A $n \times n$ mátrix felső Hessenberg-alakú, ha*

$$A = \begin{bmatrix}
a_{11} & & & \ldots & & a_{1n} \\
a_{21} & & & & & \\
0 & a_{32} & & & & \vdots \\
\vdots & 0 & \ddots & & & \\
 & & \ddots & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\
0 & & \ldots & 0 & a_{n,n-1} & a_{nn}
\end{bmatrix}.$$

Ekvivalens megfogalmazásban: A felső Hessenberg-alakú, ha $a_{ij} = 0$, $i \geq j + 2$. Hessenberg-alakú mátrixok QR-felbontása $O(n^2)$ régi flop számítási költséget igényel.

10.5 Tétel. *Ha A felső Hessenberg-alakú és $A = QR$, akkor RQ is felső Hessenberg-alakú.*

Bizonyítás. Felhasználjuk a következő észrevételt. Ha egy B mátrix felső Hessenberg-alakú és R felső háromszög alakú, akkor $F = BR$ és $G = RB$ is felső Hessenberg-alakú. Ha $i \geq j + 2$, akkor

$$f_{ij} = [\underbrace{\quad \geq j \quad}_{0,\ldots,0}, b_{i,i-1}, \ldots, b_{in}] \begin{bmatrix} r_{1j} \\ \vdots \\ r_{jj} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0,$$

illetve

$$g_{ij} = [\underbrace{\geq j+1}_{0,\ldots,0}, r_{ii}, \ldots, r_{in}] \begin{bmatrix} b_{1j} \\ \vdots \\ b_{j+1,j} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0,$$

amivel az észrevételt igazoltuk. Minthogy R^{-1} szintén felső háromszögmátrix alakú, ezért $Q = AR^{-1}$ is felső Hessenberg-alakú. Ebből az RQ Hessenberg-alakja is következik. □

Következésképpen a Hessenberg-alakú mátrixok alakja invariáns a QR-módszerrel szemben. Ha tehát a $QR-$módszer egy nulladik lépéseként az A mátrixot hasonlósági transzformációval egy felső Hessenberg-alakú mátrixba visszük át, jelentős számítási költség megtakarítást érhetünk el.

Egy A mátrixot felső Hessenberg-alakra sokféleképpen hozhatunk. Az egyik legolcsóbb ($\approx 5n^3/6$ régi flop) eljárás a Gauss-eliminációs eljáráson alapul. Legyen

$$N_k = I - \beta^{(k)} e_{k+1}^T,$$

ahol $e_{k+1} \in \mathbb{R}^n$ a $(k+1)$-edik egységvektort jelöli és

$$\beta^{(k)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ a_{k+2,k}/a_{k+1,k} \\ \vdots \\ a_{nk}/a_{k+1,k} \end{bmatrix}.$$

Tehát

$$N_k = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & a_{k+2,k}/a_{k+1,k} & \ddots & & \\ & & & \vdots & & \ddots & \\ & & & a_{n,k}/a_{k+1,k} & & & 1 \end{bmatrix}.$$

Feltéve, hogy az A mátrix soronként van particionálva

$$N_k A = A - \beta^{(k)} a_{k+1}^T = \begin{bmatrix} a_1^T \\ \vdots \\ a_{k+1}^T \\ a_{k+2}^T - (a_{k+2,k}/a_{k+1,k})\, a_{k+1}^T \\ \vdots \\ a_n^T - (a_{nk}/a_{k+1,k})\, a_{k+1}^T \end{bmatrix}.$$

A transzformáció tehát kinullázza a k-adik oszlopban az $a_{k+1,k}$ alatti mátrixelemeket és változatlanul hagyja az első $k+1$ sort. Ha $a_{k+1}^T, \ldots, a_n^T$ első $k-1$ eleme 0, akkor ez változatlanul marad és csak az $A\,(k+2:n, k:n)$ részmátrix elemei változnak. Vegyük észre, hogy $N_k^{-1} = I + \beta^{(k)} e_{k+1}^T$ és tetszőleges $B \in \mathbb{R}^{n\times n}$ esetén

$$BN_k^{-1} = B + \underbrace{c \in \mathbb{R}^n}_{B\beta^{(k)}} e_{k+1}^T = B + c e_{k+1}^T = B + [0, \ldots, 0, \underbrace{k+1}_{c}, 0, \ldots, 0].$$

Ez azt jelenti, hogy a $B \to BN_k^{-1}$ transzformáció csak a B mátrix $(k+1)$-edik oszlopát módosítja. Legyen tehát $A^{(1)} = A$ és képezzük az

$$A^{(k+1)} = N_k A^{(k)} N_k^{-1} \quad (k = 1, \ldots, n-2)$$

mátrixsorozatot. Tegyük fel, hogy $A^{(k)}$ első $k-1$ oszlopában az $a_{i+1,i}$ alatti mátrixelemek már ki vannak nullázva $(1 \le i \le k-1)$. Ekkor $N_k A^{(k)}$ kinullázza az $a_{k+1,k}$

alatti oszlopelemeket. Minthogy az $\left(N_k A^{(k)}\right) N_k^{-1}$ szorzat csak a $(k+1)$-edik oszlopot módosítja, az $A^{(k+1)}$ mátrixban az $a_{k+1,k}$ alatti oszlopelemek zérusok. Végül kapjuk, hogy

$$H = A^{(n-1)} = \underbrace{N}_{N_{n-2}\cdots N_1} \; A \; \underbrace{N^{-1}}_{N_1^{-1}\cdots N_{n-2}^{-1}} = NAN^{-1}$$

felső Hessenberg-alakú, amely hasonlósági transzformációval állt elő. Megmutatható, hogy

$$N^{-1} = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & a_{31}/a_{21} & 1 & & \\ & \vdots & \vdots & & \\ & a_{n-1,1}/a_{21} & a_{n-1,2}/a_{32} & \dots & 1 \\ & a_{n1}/a_{21} & a_{n2}/a_{32} & \dots & a_{n,n-2}/a_{n-1,n-2} \quad 1 \end{bmatrix}.$$

Az $a_{k+1,k}$ elem pivotelemként viselkedik. Ha $a_{k+1,k}$ kis abszolút értékű, vagy zérus, akkor sor- és oszlopcsere alkalmazásával itt is pivotálhatunk. Ekkor az eljárást értelemszerűen módosítani kell.

A QR-módszer konvergenciája a hatványmódszerhez hasonlóan a sajátértékek $|\lambda_{i+1}/\lambda_i|$ hányadosaitól függ. Minthogy az $A - \sigma I$ mátrix sajátértékei $\lambda_1 - \sigma, \lambda_2 - \sigma, \dots, \lambda_n - \sigma$, az ehhez kapcsolódó sajátérték hányadosok:
$|(\lambda_{i+1} - \sigma)\,/\,(\lambda_i - \sigma)|$. A σ ügyes megválasztásával ezek a hányadosok kicsivé tehetők, meggyorsítva ezáltal a konvergenciát. A Hessenberg-alakra hozást és az eltolást is alkalmazó QR-módszer algoritmusa a következő:

AZ ELTOLÁSOS QR-MÓDSZER ALGORITMUSA:

$H_1 = U^{-1}AU$ (H_1 felső Hessenberg-alakú)
for $i = 1, 2, \dots$
Számítsuk ki a $H_i - \sigma_i I = Q_i R_i$ felbontást!
$H_{i+1} = R_i Q_i + \sigma_i I$
end

A H_{i+1} mátrix hasonló H_i-hez, ui.

$$R_i Q_i + \sigma_i I = Q_i^T (H_i - \sigma_i I) Q_i + \sigma_i I = Q_i^T H_i Q_i.$$

A σ_i paraméterek megválasztására különféle stratégiák léteznek. A leggyakrabban javasolt választás a következő:

$$\sigma_k = h_{nn}^{(k)} \quad \left(H_k = \left[h_{ij}^{(k)}\right]_{i,j=1}^{n}\right).$$

A gyakorlatban a QR-módszert csak eltolásos formában használjuk.

Az A sajátvektorai a QR módszer segítségével többféleképpen is könnyen meghatározhatók. Ezek részletezése az irodalomban megtalálható.

Példa. Legyen $A = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ 6 & 0 & 0 \end{bmatrix}$. Oldjuk meg a sajátérték-sajátvektor feladatot! Ellenőrízzük a kapott eredményeket a hatványmódszerrel, QR-módszerrel és az eltolásos QR-módszerrel is!

A karakterisztikus polinom: $p(\lambda) = \det(A - \lambda E) = \lambda^3 - 2\lambda^2 - 5\lambda + 6$. A polinom gyökei, tehát A sajátértékei: $\lambda_1 = 3, \lambda_2 = -2, \lambda_3 = 1$. A megfelelő sajátvektorok: $s_1 = [0.5, 0, 5, 1]^T, \quad s_2 = [-\frac{1}{3}, \frac{2}{9}, 1]^T, \quad s_3 = [0, 1, 0]^T$.

A hatványmódszert a $v^{(0)} = [1, 1, 1]^T$-ből indítva kapott sorozat néhány eleme:

k	z	v
1	$[2.0000, 3.0000, \boxed{6.0000}]^T$	$[0.3333, 0.5000, 1.0000]^T$
5	$[1.5667, 1.6583, \boxed{3.4000}]^T$	$[0.4608, 0.4877, 1.0000]^T$
10	$[1.4919, 1.4812, \boxed{2.9517}]^T$	$[0.5055, 0.5018, 1.0000]^T$
20	$[1.4999, 1.4997, \boxed{2.9992}]^T$	$[0.5001, 0.5000, 1.0000]^T$

A $\lambda_1 = 3$-hoz konvergáló γ_k sorozat elemeit bekereteztük. A következő táblázat a QR-módszerrel kapott A_i és a $\sigma_i \equiv -1$ értékkel történt eltolásos QR-módszerrel kapott $\tilde{A}_i$ sorozat néhány elemét tartalmazza. Többféle eltolást próbáltunk ki, ezek közül a $\sigma_i \equiv -1$ gyorsította meg a legnagyobb mértékben a konvergenciát. A sajátértékek a főátlóbeli, bekeretezett elemek határértékeként adódnak.

$$
\begin{array}{rcc}
 & A_i & \tilde{A}_i \\
i = 4 & \begin{bmatrix} \boxed{3.6062} & -1.2015 & -4.4498 \\ -0.0490 & \boxed{1.0226} & 0.2516 \\ 0.7752 & -0.9574 & \boxed{-2.6288} \end{bmatrix} & \begin{bmatrix} \boxed{3.0237} & -0.8743 & -5.2865 \\ 0.0248 & \boxed{0.9893} & 0.3378 \\ 0.0120 & -0.0520 & \boxed{-2.0130} \end{bmatrix} \\
i = 9 & \begin{bmatrix} \boxed{2.9182} & -0.8528 & 5.3781 \\ 0.0110 & \boxed{0.9951} & -0.4177 \\ 0.0765 & -0.0340 & \boxed{-1.9133} \end{bmatrix} & \begin{bmatrix} \boxed{3.0003} & -0.8936 & 5.2948 \\ 0.0008 & \boxed{0.9996} & -0.4061 \\ 0.0000 & 0.0000 & \boxed{-2.0000} \end{bmatrix} \\
i = 12 & \begin{bmatrix} \boxed{3.0243} & -0.9068 & -5.2688 \\ -0.0032 & \boxed{1.0014} & 0.4052 \\ 0.0237 & -0.0106 & \boxed{-2.0257} \end{bmatrix} & \begin{bmatrix} \boxed{3.0000} & -0.8943 & -5.2947 \\ 0.0001 & \boxed{1.0000} & 0.4080 \\ 0.0000 & 0.0000 & \boxed{-2.0000} \end{bmatrix}.
\end{array}
$$

8.4. A szinguláris érték felbontás

A lineáris algebra számos eljárásában a Jordan-féle normálalak helyett lényegesen fontosabb a szerepe van az ún. *szinguláris érték felbontásnak* (*SVD*).

10.6 Tétel. *Minden $A \in \mathbb{R}^{n\times n}$ mátrix felbontható*

$$A = U\Sigma V \tag{8.3}$$

alakban, ahol U és V ortogonális mátrixok (azaz $U^TU = I$, $V^TV = I$),

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \sigma_n \end{bmatrix} \tag{8.4}$$

diagonálmátrix és $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$.

A $\sigma_1, \ldots, \sigma_n$ számokat *szinguláris értékeknek* nevezzük. A zérusmátrixra $\sigma_1 = 0$, nemszinguláris mátrixra $\sigma_n > 0$, egyébként az A mátrix rangja akkor és csak akkor r, ha $\sigma_r > 0$ és $\sigma_{r+1} = 0$. Könnyen látható, hogy $\sigma_1^2, \ldots, \sigma_n^2$ az AA^T mátrix sajátértékei.

A szinguláris értékek jóval kevésbé érzékenyek mint a sajátértékek.

10.7 Tétel. *Legyen $A, B \in \mathbb{R}^{n\times n}$, amelyek szinguláris értékei rendre $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$, illetve $\tau_1 \geq \tau_2 \geq \ldots \geq \tau_n \geq 0$. Ekkor*

$$\max_{1\leq i\leq n} |\sigma_i - \tau_i| \leq \|A - B\|_2 . \tag{8.5}$$

A tétel a szinguláris értékek numerikus stabilitását mondja ki. Ha $B = A+\delta A$, akkor $\max_{1\leq i\leq n} |\sigma_i - \tau_i| \leq \|\delta A\|_2$, tehát az A kismértékű megváltozása a szinguláris értékek kismértékű megváltozását vonja maga után. Ha figyelembevesszük, hogy σ_i^2 az AA^T szimmetrikus (és pozitív szemidefinit) mátrix sajátértéke, akkor a tételből azt a következtetést is levonhatjuk, hogy szimmetrikus pozitív szemidefinit mátrixok szimmetrikusságot és pozitív szemidefinitséget megőrző perturbációi esetén a sajátértékek csak kicsit változnak.

A szinguláris érték felbontás általánosítható téglalapmátrixok esetére is.

10.8 Tétel. *Legyen $A \in \mathbb{R}^{m\times n}$ és* $\operatorname{rank}(A) = r$. *Ekkor léteznek $U \in \mathbb{R}^{m\times m}$ és $V \in \mathbb{R}^{n\times n}$ ortogonális mátrixok úgy, hogy*

$$A = U\Sigma V^T, \quad \Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m\times n},$$

ahol $\Sigma_r =$ $\operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ *és $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$.*

A következőkben egy robusztus eljárást adunk meg a szinguláris érték felbontás meghatározására. Ehhez szükségünk van a következőkre.

A 10.3 Tételben kimondtuk, hogy minden $A \in \mathbb{R}^{n\times m}$ mátrix, amelyre $\text{rank}(A) = m$ ($m \leq n$), előáll $A = QR$ alakban, ahol $Q \in \mathbb{R}^{n\times m}$ oszlopai ortonormáltak, rangja m és $R \in \mathbb{R}^{m\times m}$ nemszinguláris felső háromszögmátrix. Ez a QR-felbontás nem egyértelmű. Legyen $A = Q_1R_1$ és $A = Q_2R_2$ két különböző QR-felbontás. A $Q_1R_1 = Q_2R_2$ egyenlőségből azonos átalakítással kapjuk, hogy $Q_2^TQ_1 = R_2R_1^{-1}$. Minthogy $Q_2^TQ_1 \in \mathbb{R}^{m\times m}$ ortogonális mátrix, a transzponáltja az inverze. Ugyanakkor $R_2R_1^{-1}$ felső háromszögmátrix, aminek az inverze is felső háromszögmátrix, viszont a transzponáltja alsó háromszögmátrix. Egy alsó és egy felső háromszögmátrix csak akkor egyezhet meg, ha diagonálisak. Tehát $Q_2^TQ_1 = R_2R_1^{-1} = D$, ahol D diagonális mátrix. Innen adódik, hogy $Q_2 = Q_1D$, $R_2 = DR_1$. A $Q_2^TQ_1$ mátrix ortogonalitása miatt D is ortogonális, azaz $D^TD = D^2 = I$. Ezért a D mátrix diagonális elemei csak ± 1-ek lehetnek. Ha kikötjük, hogy R minden diagonális eleme pozitív legyen, azaz $r_{ii} > 0$ ($i = 1, \ldots, m$), akkor a QR-felbontás egyértelmű lesz. Ekkor ugyanis D diagonális elemei csak $(+1)$-ek lehetnek.

Legyen $A = QR$ egy tetszőleges felbontás, ahol R diagonálisában van negatív elem. Definiáljuk a D mátrixot a következőképpen. Legyen $d_{ii} = 1$, ha $r_{ii} > 0$ és $d_{ii} = -1$, ha $r_{ii} < 0$. A $Q_1 = QD$ és $R_1 = DR$ mátrixok A egyetlen olyan QR felbontását definiálják, amelyben a felső háromszögmátrix diagonális elemei pozitívak. Természetesen a Q_1 és R_1 mátrixokat úgy kaphatjuk meg legkönnyebben, hogy $r_{ii} < 0$ esetén Q i-edik oszlopát és R i-edik sorát (-1)-el megszorozzuk.

A következőkben feltesszük, hogy a QR-felbontás minden esetben olyan R felső háromszögmátrixot állít elő, amelynek diagonálisában csupa pozitív elem van. A következő algoritmus A szinguláris érték felbontását állítja elő, ha $A \in \mathbb{R}^{n\times m}$, $\text{rank}(A) = m$ és $m \leq n$.

TRANSZPONÁLT QR-ALGORITMUS:

$A = Q_0R_1$

for $i = 1, \ldots$

Számítsuk ki az $R_i^T = Q_iR_{i+1}$ felbontást!

end

Legyen $U_j = Q_2Q_4 \ldots Q_{2j}$ és $V_j = Q_1Q_3 \ldots Q_{2j-1}$. Ekkor $R_{2j} = V_j^TR_1^TU_{j-1}$ és $R_{2j+1} = U_j^TR_1V_j$. Igazolható, hogy $U_j \to \tilde{U}$, $V_j \to V$ és $R_j \to S$, ahol S diagonális mátrix. Minthogy $R_1 = U_jR_{2j+1}V_j^T$, határátmenettel kapjuk, hogy $R_1 = Q_0^TA = \tilde{U}SV^T$, ahonnan az $U = Q_0\tilde{U}$ bevezetésével az

$$A = USV^T$$

szinguláris érték felbontás adódik.

Az eljárással kapcsolatban a következőket jegyezzük meg. Vegyük észre, hogy

$$R_iR_i^T = R_{i+1}^TR_{i+1}.$$

Megmutatható, hogy a $B_i = R_i R_i^T$ mátrixok egymáshoz, illetve $B_1 = Q_0^T A A^T Q_0$ miatt AA^T-hez hasonlóak. Ha csak a szinguláris értékekre van szükségünk, akkor a Cholesky-felbontás segítségével az eljárást a következőképpen írhatjuk át:

$A = Q_0 R_1$
for $i = 1, \ldots$
Számítsuk ki az $R_i R_i^T = R_{i+1}^T R_{i+1}$ Cholesky-felbontást!
end

Példa. Az előző fejezetben szereplő $A = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ 6 & 0 & 0 \end{bmatrix}$ mátrix szinguláris érték felbontása a transzponált QR-algoritmussal a következő sorozatokat eredményezi:

$$
\begin{array}{lcc}
 & Q_0 U_j & V_j \\
j = 2 & \begin{bmatrix} 0.1596 & -0.1682 & 0.9727 \\ 0.3191 & 0.9412 & 0.1104 \\ 0.9342 & -0.2928 & -0.2039 \end{bmatrix}, & \begin{bmatrix} 0.9985 & -0.0466 & -0.0304 \\ 0.0497 & 0.9924 & 0.1128 \\ 0.0249 & -0.1141 & 0.9932 \end{bmatrix} \\
j = 5 & \begin{bmatrix} 0.1596 & -0.3543 & 0.9214 \\ 0.3192 & 0.9018 & 0.2915 \\ 0.9342 & -0.2476 & -0.2570 \end{bmatrix}, & \begin{bmatrix} 0.9985 & -0.0399 & -0.0388 \\ 0.0498 & 0.9522 & 0.3014 \\ 0.0249 & -0.3028 & 0.9527 \end{bmatrix} \\
j = 10 & \begin{bmatrix} 0.1596 & -0.6161 & 0.7713 \\ 0.3192 & 0.7716 & 0.5503 \\ 0.9342 & -0.1584 & -0.3198 \end{bmatrix}, & \begin{bmatrix} 0.9985 & -0.0260 & -0.0492 \\ 0.0498 & 0.8130 & 0.5801 \\ 0.0249 & -0.5816 & 0.8131 \end{bmatrix} \\
\vdots & \vdots & \vdots \\
j = \infty & \begin{bmatrix} 0.1596 & -0.8944 & 0.4178 \\ 0.3192 & 0.4472 & 0.8355 \\ 0.9342 & 0.0000 & -0.3568 \end{bmatrix}, & \begin{bmatrix} 0.9985 & 0.0000 & -0.0556 \\ 0.0498 & 0.4472 & 0.8930 \\ 0.0249 & -0.8944 & 0.3568 \end{bmatrix}
\end{array}
$$

Az $s = (\sigma_1, \sigma_2, \sigma_3)$ szinguláris értékeket közelítő, az R_j diagonális elemeiből álló s_j sorozat:

$$
\begin{aligned}
s_4 &= (6.4126, 0.9553, 0.9774), \\
s_{10} &= (6.4128, 0, 9653, 0.9693), \\
s_{20} &= (6.4128, 0.9863, 0.9486).
\end{aligned}
$$

A három, 4 tizedesjegyre pontos szinguláris érték: $s = (6.4128, 1.0000, 0.9356)$. Megjegyezzük, hogy a 64-ik iteráció után állt be minden közelítő szinguláris érték 4 tizedesjegy pontossággal, de ilyen pontossággal már az $A \approx (Q_0 U_6) S_6 V_5^T$ is teljesült.

8.5. Feladatok

1. Alkalmazzuk a hatványmódszert az $A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$ mátrixra a $v^{(0)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ kezdőértékkel! Mi a 20-ik lépés eredménye?

2. Alkalmazzuk a hatványmódszert, az inverz hatványmódszert, illetve a QR-módszert a

$$\begin{bmatrix} -4 & -3 & -7 \\ 2 & 3 & 2 \\ 4 & 2 & 7 \end{bmatrix}$$

mátrixra!

3. Alkalmazzuk az eltolásos QR-módszert a 2. feladat mátrixára egy rögzített $\sigma_i = \sigma$ értékkel!

9. fejezet

INTERPOLÁCIÓ

Az interpoláció feladatát a következőképpen fogalmazhatjuk meg. Ismerjük egy $y = f(x)$ ($f : \mathbb{R} \to \mathbb{R}$) függvény

$$a \leq x_1 < x_2 < \ldots < x_n \leq b \tag{9.1}$$

pontokban felvett értékeit, az

$$y_i = f(x_i) \quad (i = 1, \ldots, n) \tag{9.2}$$

függvényértékeket. Az $f(x)$ függvényt, amely lehet ismert, vagy akár ismeretlen is, egy olyan, általában könnyen számítható $h(x)$ függvénnyel közelítjük (vagy helyettesítjük), amelyre fennáll, hogy

$$y_i = h(x_i) \quad (i = 1, \ldots, n). \tag{9.3}$$

Az $\{x_i\}_{i=1}^n$ pontokat *interpolációs alappontoknak*, az (9.3) feltételt *interpolációs feltételnek* nevezzük. Az interpolációs feltétel teljesülése esetén azt reméljük, hogy a

$$h(x) = h(x; \{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n)$$

interpoláló függvény az (x_i, x_{i+1}) intervallumokban jól közelíti az $f(x)$ függvényt. A $h(x)$ függvény megválasztásától függően beszélünk különböző típusú interpolációkról.

Ha a $h(x)$ függvénnyel $f(x)$-et az (x_1, x_n) intervallumon kívül közelítjük, akkor *extrapolációról* beszélünk.

9.1. A lineáris interpoláció

A lineáris interpoláció esetén a $h(x)$ függvény alakja

$$h(x) = a_1\phi_1(x) + a_2\phi_2(x) + \ldots + a_n\phi_n(x) = \sum_{i=1}^{n} a_i\phi_i(x), \tag{9.4}$$

ahol a $\phi_i : [a,b] \to \mathbb{R}$ $(i = 1, \ldots, n)$ bázisfüggvények adottak. Az ismeretlen $a_1, \ldots, a_n$ együtthatókat az interpolációs feltételből határozhatjuk meg. Ekkor teljesülnie kell az alábbi n feltételnek

$$\begin{array}{rcl} a_1\phi_1(x_1) + a_2\phi_2(x_1) + \ldots + a_n\phi_n(x_1) &=& f(x_1), \\ \vdots & \vdots & \vdots \\ a_1\phi_1(x_n) + a_2\phi_2(x_n) + \ldots + a_n\phi_n(x_n) &=& f(x_n), \end{array} \tag{9.5}$$

amely lineáris egyenletrendszer az ismeretlen $a_1, \ldots, a_n$ együtthatókra nézve. Legyen

$$B = [\phi_j(x_i)]_{i,j=1}^n \tag{9.6}$$

és

$$a = [a_1, \ldots, a_n]^T, \quad c = [f(x_1), \ldots, f(x_n)]^T. \tag{9.7}$$

A fenti feltétel tömör alakban

$$Ba = c. \tag{9.8}$$

Ha $\det(B) \neq 0$, akkor az egyenletrendszernek pontosan egy megoldása van: $a = B^{-1}c$.

A gyakorlatban sokféle $\{\phi_i(x)\}_{i=1}^n$ bázisfüggvényt alkalmaznak. Az egyik legfontosabb a

$$\phi_1(x) = 1, \ \phi_2(x) = x, \ \ldots \ , \ \phi_n(x) = x^{n-1} \tag{9.9}$$

függvényrendszer, amely a *Lagrange-féle interpolációs feladatot* definiálja. Ekkor az interpolációs feladat mátrixa

$$B = \begin{bmatrix} 1 & x_1 & \ldots & x_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^{n-1} \end{bmatrix} \tag{9.10}$$

az ún. Vandermonde-féle mátrix, amely a $\det(B) = \prod_{1 \leq i < j \leq n}(x_j - x_i)$ összefüggés miatt nemszinguláris. Tehát a Lagrange-féle interpolációs feladatnak egyértelmű megoldása van.

További fontos esetek a következők. A trigonometrikus interpolációt a

$$\phi_1(x) = 1, \ \phi_{2k}(x) = \sin(kx), \ \phi_{2k+1}(x) = \cos(kx) \quad \left(k = 1, \ldots, \frac{n-1}{2}\right) \tag{9.11}$$

függvényrendszer ($n = 2k+1$, $[a,b] = [-\pi, \pi]$), az exponenciális interpolációt pedig a

$$\phi_i(x) = e^{\lambda_i x} \quad (i = 1, \ldots, n, \ \lambda_1 < \lambda_2 < \ldots < \lambda_n) \tag{9.12}$$

függvényrendszer definiálja. Racionális törtfüggvényeket használ a

$$\phi_i(x) = \frac{1}{a_i + x} \quad (i = 1, \ldots, n, \ 0 < a_1 < \ldots < a_n) \tag{9.13}$$

függvényrendszer. Itt fel kell tennünk, hogy $x + a_1 > 0$. Ez könnyen teljesül, ha $x \in [a, b]$ és $a + a_1 > 0$.

Nem minden $\{\phi_i(x)\}_{i=1}^n$ függvényrendszer és $x_1 < x_2 < \ldots < x_n$ alappontrendszer esetén van megoldása a lineáris interpolációs feladatnak.

Példa. Legyen $\phi_1(x) = 1$, $\phi_2(x) = x^2$, $x_1 = -1$, $x_2 = 1$. Ekkor

$$B = \begin{bmatrix} 1 & (-1)^2 \\ 1 & 1 \end{bmatrix}, \quad \det(B) = 0.$$

A lineáris interpolációs feladat mátrixa sok esetben rosszul kondicionált. Ilyenkor speciális technikákat vagy más típusú interpolációt kell használni.

9.2. A Lagrange-féle interpolációs feladat

A feladat szokásos megfogalmazása a következő. Adottak az $x_1 < x_2 < \ldots < x_n$ alappontok és az $y_i = f(x_i)$ $(i = 1, \ldots, n)$ függvényértékek. Határozzuk meg azt a legfeljebb $(n-1)$-edfokú

$$p(x) = a_0 + a_1 x + \ldots + a_{n-1} x^{n-1} \tag{9.14}$$

polinomot, amelyre teljesül a

$$y_i = p(x_i) \quad (i = 1, \ldots, n) \tag{9.15}$$

interpolációs feltétel.

A *Lagrange-féle* interpolációs polinom létezését és egyértelműségét már beláttuk. A polinom többféle ekvivalens alakban is felírható. Különösen fontos azonban a Lagrange-féle előállítás. Legyen

$$l_i(x) = \prod_{k=1, k\neq i}^{n} \frac{x - x_k}{x_i - x_k} \quad (i = 1, \ldots, n) \tag{9.16}$$

az i-edik *Lagrange-féle alappolinom*. Ekkor az interpolációs polinom előáll

$$p(x) = \sum_{i=1}^{n} y_i l_i(x) \tag{9.17}$$

alakban. Ennek igazolására vegyük észre, hogy

$$l_i(x_j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{9.18}$$

és

$$p(x_j) = \sum_{i=1}^{n} y_i l_i(x_j) = y_j l_j(x_j) = y_j \quad (j = 1, \ldots, n). \tag{9.19}$$

A Lagrange-féle interpolációs polinom hibájára vonatkozik a következő

11.1 Tétel *(Cauchy). Ha* $f \in C^n[a,b]$, $[x_1, x_n] \subset [a,b]$ *és* $x \in [a,b]$, *akkor*

$$f(x) - p(x) = \frac{f^{(n)}(\xi_x)}{n!}(x - x_1)(x - x_2)\ldots(x - x_n),$$

ahol $\xi_x = \xi(x)$ *az* x *és az* x_1, x_n *pontok által kifeszített intervallumban van.*

Bizonyítás. Ha van i, hogy $x = x_i$, akkor állításunk triviális. Egyébként legyen $\omega(x) = (x - x_1)(x - x_2)\ldots(x - x_n)$ és tekintsük a következő segédfüggvényt:

$$W(t) = f(t) - p(t) - [f(x) - p(x)]\frac{\omega(t)}{\omega(x)}. \tag{9.20}$$

A $W(t) \in C^n[a,b]$ függvénynek van $n+1$ gyökhelye: $x, x_1, \ldots, x_n$. A Rolle–tétel miatt $W(t)$ bármely két gyökhelye között a $W'(t)$ deriváltfüggvénynek zérushelye van. Ezért $W'(t)$-nek legalább n zérushelye van. Hasonlóképpen okoskodva belátható, hogy $W''(t)$-nek legalább $n-1$, $W^{(3)}(t)$-nek legalább $n-2$ zérushelye van, és így tovább. Végül $W^{(n)}(t)$-nek is van legalább egy zérushelye, amit jelöljön ξ_x. Minthogy $p^{(n)}(t) \equiv 0$ és $\omega^{(n)}(t) \equiv n!$, azért

$$W^{(n)}(\xi_x) = f^{(n)}(\xi_x) - [f(x) - p(x)]\frac{n!}{\omega(x)} = 0, \tag{9.21}$$

ahonnan átrendezéssel kapjuk a tétel állítását. □

Következmény. *Ha* $\left|f^{(n)}(x)\right| \le M_n$ *(*$x \in [a,b]$*), akkor*

$$|f(x) - p(x)| \le \frac{M_n}{n!}(b-a)^n. \tag{9.22}$$

Konkrét n esetén szélsőértékszámítással élesebb becslés is levezethető.

Példa. Hány ekvidisztáns alappontban kell megadnunk a $\sin x$ függvény táblázatát a $\left[0, \frac{\pi}{2}\right]$ intervallumon ahhoz, hogy a közbülső pontokban lineáris Lagrange-interpolációt használva az elkövetett hiba legfeljebb $\varepsilon = 10^{-4}$ legyen? Vezessük be a $h = x_{i+1} - x_i$ jelölést. A Cauchy-tétel következménye alapján olyan h-t keresünk, melyre $M_2 h^2/2 \le 10^{-4}$. Mivel $(\sin x)'' = -\sin x$, választhatjuk az $M_2 = 1$ értéket. Ezzel $h \le \sqrt{2}/100$, $n \ge \frac{\pi}{2h}$ miatt $n \ge 112$ adódik. Ha viszont a hibakorlátot az

$$|f(x) - p(x)| \le \frac{M_2}{2} \max |(x - x_i)(x - x_{i+1})|$$

becslésből közvetlenül vezetjük le szélsőértékszámítással, akkor az élesebb,

$$|f(x) - p(x)| \leq \frac{M_2 h^2}{8}$$

eredményt kapjuk. Ez alapján kiderül, hogy $n = 28$ pont is elég.

Az interpolációs eljárásoktól elvárjuk, hogy a pontok számának növelése esetén a közelítés hibája csökken. Ez azonban nem minden esetben van így, amint azt a feladatok között szereplő Runge-féle példa is mutatja. Nagy n-ek esetén numerikus instabilitás is felléphet. Ennek illusztrálására tegyük fel, hogy az $y_i = f(x_i)$ függvényértékeket ε_i hibával ismerjük $(i = 1, \ldots, n)$. Ekkor az elméleti

$$p(x) = \sum_{i=1}^{n} f(x_i) l_i(x)$$

Lagrange-interpolációs polinom helyett a perturbált

$$\tilde{p}(x) = \sum_{i=1}^{n} (f(x_i) + \varepsilon_i) l_i(x)$$

polinommal számolunk. A kettő eltérésére teljesül, hogy

$$\delta\left(p\left(x\right)\right) = |\tilde{p}(x) - p(x)| = \left| \sum_{i=1}^{n} \varepsilon_i l_i(x) \right| \leq \sum_{i=1}^{n} |\varepsilon_i| \, |l_i(x)| \leq \left(\max_{1 \leq i \leq n} |\varepsilon_i| \right) \sum_{i=1}^{n} |l_i(x)| \, .$$

Ez a becslés pontos. Igazolható, hogy

$$\sum_{i=1}^{n} |l_i(x)| > \frac{2}{\pi} \log n + c, \tag{9.23}$$

ahol c konstans. Ha n elég nagy, akkor a $\delta\left(p\left(x\right)\right)$ perturbációs hiba is nagy lesz.

A divergencia és numerikus instabilitás miatt sok esetben más típusú interpolációs technikákat használunk.

Példa. Közelítsük másodfokú függvénnyel az $f(x) = \cos(\frac{\pi}{2}x)$ függvényt a$[-1, 1]$-ben az $x_1 = -1, x_2 = 0, x_3 = 1$ pontokra támaszkodva! $f(x) \approx p(x) = A_1 + A_2 x + A_3 x^2$. Az együtthatókra felírható az

$$\begin{aligned} A_1 - A_2 + A_3 &= 0 \\ A_1 &= 1 \\ A_1 + A_2 + A_3 &= 0 \end{aligned}$$

egyenletrendszer. Innen $p(x) = 1 - x^2$. Természetesen ugyanezt kapjuk az $l_i(x)$ Lagrange-függvényekkel is. $f(x_1) = f(x_3) = 0$ miatt elég az $l_2(x)$-t meghatározni, ez $1 - x^2$, ami jelen esetben a $p(x)$ polinommal megegyezik. A közelítés hibáját

$$h \leq \frac{M_3}{3!} \max_{-1 \leq x \leq 1} |(x+1)x(x-1)|$$

becsli, ahol M_3 az $|f'''(x)|$ maximuma, jelen esetben $\pi^3/8$. Szélsőértékszámítással adódik, hogy

$$\max_{-1\le x\le 1} |(x+1)x(x-1)| = \frac{8}{27},$$

azaz $h \le \pi^3/216 \simeq 0.15$.

9.3. Harmadfokú szplájn interpoláció

A szplájn interpoláció is a lineáris interpolációk közé tartozik alkalmasan megválasztott $\{\phi_i\}$ bázisfüggvény-rendszerrel. Bevezetése és tárgyalása viszont más alapokon szokásos. A szplájn interpoláció esetén a $h(x)$ interpoláló függvényt szakaszonként adjuk meg speciális csatlakozási feltételekkel. Az $a = x_1 < x_2 < \ldots < x_n = b$ alappontokhoz, illetve az $y_i = f(x_i)$ $(i = 1, \ldots, n)$ függvényértékekhez olyan $S(x)$ függvényt keresünk, amely kielégíti a következő feltételeket:

(i) $S(x) = S_i(x)$ $(x \in [x_i, x_{i+1}])$, ahol $S_i(x)$ legfeljebb harmadfokú polinom $(i = 1, \ldots, n-1)$,
(ii) $S(x_i) = y_i$ $(i = 1, \ldots, n)$,
(iii) $S_i(x_{i+1}) = S_{i+1}(x_{i+1})$ $(i = 1, \ldots, n-2)$,
(iv) $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$ $(i = 1, \ldots, n-2)$,
(v) $S_i''(x_{i+1}) = S_{i+1}''(x_{i+1})$ $(i = 1, \ldots, n-2)$,
(vi) $S''(x_1) = A_n$, $S''(x_n) = B_n$.

Az (ii)-(iii) feltételek együttesen azt jelentik, hogy az $S(x)$ függvény folytonos az $[a, b]$ intervallumon. Az (iv)-(v) feltételek azt mondják ki, hogy $S'(x)$ és $S''(x)$ folytonos. Ha $A_n = B_n = 0$, akkor az $S(x)$ függvényt *természetes szplájn*nak nevezzük.

Legyen $h_i = x_{i+1} - x_i$ az i-edik részintervallum hossza $(i = 1, \ldots, n-1)$. A $S(x)$ szplájnt az $[x_i, x_{i+1}]$ intervallumon a

$$S(x) = S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{9.24}$$

alakban keressük $(i = 1, \ldots, n-1)$. Az (ii)-(vi) feltételek felhasználásával az ismeretlen a_i, b_i, c_i és d_i együtthatókat a következőképpen határozhatjuk meg.

Az (ii), azaz az $S(x_i) = S_i(x_i) = y_i$ interpolációs feltétel miatt $a_i = y_i$ $(i = 1, \ldots, n-1)$. Az (iii) csatlakozási feltétel alakja

$$S_i(x_{i+1}) = y_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = y_{i+1} \quad (i = 1, \ldots, n-1), \tag{9.25}$$

Az (iv) csatlakozási feltétel alakja

$$S_i'(x_{i+1}) = b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1} = S_{i+1}'(x_{i+1}) \quad (i = 1, \ldots, n-2). \tag{9.26}$$

Hasonlóképpen kapjuk, hogy a (v) feltétel alakja

$$S_i''(x_{i+1}) = 2c_i + 6d_i h_i = 2c_{i+1} = S_{i+1}''(x_{i+1}) \quad (i = 1, \ldots, n-2).$$

Ebből az egyenlőség-láncból a

$$c_{i+1} = c_i + 3d_i h_i \quad (i = 1, \ldots, n-2)$$

összefüggéseket kapjuk. A (vi) végpont-feltétel alakja

$$S''(x_1) = 2c_1 = A_n, \quad S''(x_n) = 2c_{n-1} + 6d_{n-1}h_{n-1} = B_n. \tag{9.27}$$

Így kapjuk, hogy

$$d_i = \frac{c_{i+1} - c_i}{3h_i} \quad (i = 1, \ldots, n-2), \quad d_{n-1} = \frac{B_n - 2c_{n-1}}{6h_{n-1}}. \tag{9.28}$$

Mindent összevetve a $3(n-1)$ ismeretlenre az (9.25)-(9.28) összefüggések összesen $3(n-1)$ egyenletet adnak. Az (9.25) egyenletből b_i-t kifejezhetjük a következőképpen:

$$b_i = \frac{y_{i+1} - y_i}{h_i} - c_i h_i - d_i h_i^2 \quad (i = 1, \ldots, n-1). \tag{9.29}$$

A (9.27)-(9.29) összefüggéseket felhasználva a szplájn előállítható a $c_1, c_2, \ldots, c_{n-1}$ együtthatók ismeretében. A (9.28) összefüggést a (9.29)–ba beírva kapjuk, hogy $i = 1, \ldots, n-2$ esetén

$$b_i = \frac{y_{i+1} - y_i}{h_i} - c_i h_i - \frac{c_{i+1} - c_i}{3h_i} h_i^2 = \frac{y_{i+1} - y_i}{h_i} - \frac{2c_i + c_{i+1}}{3} h_i. \tag{9.30}$$

Hasonlóképpen kapjuk, hogy

$$\begin{aligned} b_{n-1} &= \frac{y_n - y_{n-1}}{h_{n-1}} - c_{n-1}h_{n-1} - \frac{B_n - 2c_{n-1}}{6h_{n-1}} h_{n-1}^2 \\ &= \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{4c_{n-1} - B_n}{6} h_{n-1}. \end{aligned} \tag{9.31}$$

Legyen $\Delta_i = (y_{i+1} - y_i)/h_i$ és helyettesítsük a b_i és d_i együtthatókra vonatkozó összefüggéseket a (9.26) egyenlőségbe:

$$\begin{aligned} \Delta_{i+1} - \frac{2c_{i+1} + c_{i+2}}{3} h_{i+1} &= \Delta_i - \frac{2c_i + c_{i+1}}{3} h_i + 2c_i h_i + 3\frac{c_{i+1} - c_i}{3h_i} h_i^2 \\ (i &= 1, \ldots, n-3), \end{aligned}$$

$$\Delta_{n-1} - \frac{4c_{n-1} - B_n}{6} h_{n-1} = \Delta_{n-2} - \frac{2c_{n-2} + c_{n-1}}{3} h_{n-2} + \\ +2c_{n-2}h_{n-2} + 3\frac{c_{n-1} - c_{n-2}}{3h_{n-2}} h_{n-2}^2.$$

Az összefüggések átrendezésével kapjuk, hogy

$$h_i c_i + 2(h_i + h_{i+1})c_{i+1} + h_{i+1}c_{i+2} = 3\left(\Delta_{i+1} - \Delta_i\right), \tag{9.32}$$

$(i = 1, \ldots, n-3)$ és

$$h_{n-2}c_{n-2} + 2(h_{n-2} + h_{n-1})c_{n-1} = 3\left(\Delta_{n-1} - \Delta_{n-2} + \frac{h_{n-1}}{6} B_n\right). \tag{9.33}$$

Figyelembevéve, hogy $c_1 = A_n/2$, az első egyenlet $(i = 1)$ átmegy a

$$2\left(h_1 + h_2\right) c_2 + h_2 c_3 = 3\left(\Delta_2 - \Delta_1 - \frac{h_1}{2} A_n\right) \tag{9.34}$$

alakba. Az i-edik egyenletet $(h_i + h_{i+1})$-el osztva kapjuk, hogy

$$2c_2 + \lambda_1 c_3 = \frac{3}{h_1 + h_2}\left(\Delta_2 - \Delta_1 - \frac{h_1}{2} A_n\right), \tag{9.35}$$

$$\mu_i c_i + 2c_{i+1} + \lambda_i c_{i+2} = \frac{3}{h_i + h_{i+1}}\left(\Delta_{i+1} - \Delta_i\right) \quad (i = 2, \ldots, n-3), \tag{9.36}$$

$$\mu_{n-2}c_{n-2} + 2c_{n-1} = \frac{3}{h_{n-2} + h_{n-1}}\left(\Delta_{n-1} - \Delta_{n-2} + \frac{h_{n-1}}{6} B_n\right), \tag{9.37}$$

ahol $\lambda_i = h_{i+1}/\left(h_i + h_{i+1}\right)$, $\mu_i = 1 - \lambda_i$ $(i = 1, \ldots, n-2)$. Ez egy $n-2$ ismeretlenes lineáris egyenletrendszer a $c_2, c_3, \ldots, c_{n-1}$ ismeretlenekre. Az egyenletrendszer mátrixa $n > 4$ esetén tehát

$$A = \begin{bmatrix} 2 & \lambda_1 & 0 & \ldots & 0 \\ \mu_2 & 2 & \lambda_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \mu_{n-3} & 2 & \lambda_{n-3} \\ 0 & \ldots & 0 & \mu_{n-2} & 2 \end{bmatrix}$$

egy három átlóból álló sávmátrix, $n = 3$ esetén $A = \begin{bmatrix} \mu_1 & 2 \end{bmatrix}$, $n = 4$ esetén pedig $A = \begin{bmatrix} 2 & \lambda_1 \\ \mu_2 & 2 \end{bmatrix}$. A Gersgorin tétel miatt A összes sajátértéke benne van

a $|z-2| \leq 1$ körlemezben. Következésképpen A nemszinguláris. Minthogy az A mátrix diagonálisan domináns is, az egyenletrendszer főelemkiválasztás nélkül és numerikusan stabilan megoldható a sávos Gauss-eliminációval $O(n)$ régi flop művelettel. Igaz a következő

11.2 Tétel. *Az (i)-(vi) feltételekkel meghatározott $S(x)$ szplájn létezik és egyértelmű. Ha $f \in C^2[a,b]$, akkor létezik $K > 0$ konstans , hogy*

$$|f(x) - S(x)| \leq K \left(\max_{1 \leq i \leq n-1} h_i \right)^2 \quad (x \in [a,b]) \,. \tag{9.38}$$

Ha $f \in C^3[a,b]$, $A_n = f''(x_1)$ és $B_n = f''(x_n)$, akkor létezik $\widetilde{K} > 0$ konstans, hogy

$$|f(x) - S(x)| \leq \widetilde{K} \left(\max_{1 \leq i \leq n-1} h_i \right)^3 \quad (x \in [a,b]) \,. \tag{9.39}$$

Ha az $f''(x_1)$ és $f''(x_n)$ információk nem állnak rendelkezésre, akkor a természetes szplájnt definiáló $A_n = B_n = 0$ választással élünk. A közelítés hibájára ekkor a (9.38) becslés érvényes. A természetes szplájn az

$$\int_a^b [s''(x)]^2 \, dx \to \min \quad (s(x_i) = y_i, \ i = 1, \ldots, n) \tag{9.40}$$

feltételes szélsőértékfeladat megoldása. A természetes szplájn mechanikai tartalma a következő. Legyen adott egy rugalmas rúd (szplájn), amely átmegy az (x_i, y_i) pontokon (tengelyeken). A legkisebb deformációs energia mechanikai elve miatt a szplájn azt az alakot veszi fel, amely a fenti (közelítő) kifejezést minimalizálja.

A (vi) végpont-feltétel helyett más kikötések is lehetségesek. Ilyenek például az

$$S'(x_1) = a_1, \quad S'(x_n) = b_1, \tag{9.41}$$

ill. az

$$S^{(i)}(x_1) = S^{(i)}(x_n) \quad (i = 1, 2) \tag{9.42}$$

végpont-feltételek. Az előbbi feltételek az ún. *teljes szplájn*t, míg az utóbbiak az ún. *periodikus szplájn*t definiálják. Ezek tárgyalása hasonló a most látotthoz és részletesen megtalálható a szakirodalomban.

A szplájn függvények előnyei: gyors és numerikusan stabil kiszámítás, nagyon jó közelítési tulajdonságok. Hátrányuk a bonyolult megadás, amely számítógépek használata esetén nem jelent komoly problémát.

9.4. Feladatok

1. Legyen $x = [x_1, \ldots, x_n]$ és $y = [y_1, \ldots, y_n]$! Oldjuk meg a következő lineáris interpolációs problémákat!

(A) $x = [0, 1, 2, \ldots, 7]$, $y = [0, 1, 0, -1, 0, 1, 0, -1]$, $\phi_i(x) = \sin(ix)$ $(i = 1, \ldots, 8)$

(B) $x = [1, 2, 4, 8]$, $y = [0.5, 0.3, 0.1, 0.05]$, $\phi_i(x) = e^{\lambda_i x}$, $\lambda_i \in \{0, -1, -2, -3\}$

(C) $x_i = \left(\frac{i-1}{10}\right)^2$, $y_i = \frac{1}{1+i^2}$, $\phi_i(x) \in \left\{1, x, x^2, e^x, e^{-x}, \frac{1}{1+x}\right\}$ $(i = 1, \ldots, 6)$.

2. (*Runge példája*). Ábrázoljuk az $f(x) = \frac{1}{1+x^2}$ függvényt a $[-5, 5]$ intervallumon és n különböző értékeire $(n = 11, 17, \ldots)$ az $f(x)$ függvényhez és az $x_i = -5+10\frac{i-1}{n-1}$ $(i = 1, \ldots, n)$ alappontokhoz tartozó Lagrange-féle interpolációs polinomot! Mit tapasztal, ha n értéke nő? Tapasztalja-e ugyanezt a jelenséget, ha a számításokat a $[-3, 3]$ intervallumon végzi?

3. Oldjuk meg a 2. feladatot természetes szplájn függvényekkel is! Hasonlítsuk össze a kapott eredményt a Lagrange-interpolációval kapottal!

10. fejezet

NUMERIKUS DERIVÁLÁS

A numerikus deriválás alapproblémája az $f : \mathbb{R} \to \mathbb{R}$ függvény deriváltjának kiszámítása egy vagy több adott pontban. A probléma kézenfekvő megoldása az, hogy az $f(x)$ függvényt egy $h(x)$ függvénnyel (lineáris interpolációval, szplájn-interpolációval, stb.) közelítjük és az $f'(x)$ közelítésének a közelítő függvény $h'(x)$ deriváltját tekintjük. Sematikusan: ha $f(x) \approx h(x)$, akkor $f'(x) \approx h'(x)$ és általában $f^{(j)}(x) \approx h^{(j)}(x)$.

10.1. A Lagrange-interpoláció esete

Adott $x_1 < x_2 < \ldots < x_n$ alappontok és $y_i = f(x_i)$ $(i = 1, \ldots, n)$ függvényértékek esetén az $f(x)$ függvény Lagrange-féle interpolációs polinomja $p(x) = \sum_{i=1}^{n} y_i l_i(x)$. Az $f\,(x)$ függény x-pontbeli j-edik deriváltjának közelítését az

$$f^{(j)}(x) \approx p^{(j)}(x) = \sum_{i=1}^{n} y_i l_i^{(j)}(x) \tag{10.1}$$

összefüggés adja meg, amelynek hibájára fennáll az

$$\left|f^{(j)}(x) - p^{(j)}(x)\right| \leq \sum_{i=0}^{j} \frac{j!}{(j-i)!(n+i)!} \max_{x\in[a,b]} \left|f^{(n+i)}(x)\right| \left|\omega^{(j-i)}(x)\right| \tag{10.2}$$

egyenlőtlenség, ahol $x, x_1, x_n \in [a, b]$ és $\omega(x) = (x - x_1)(x - x_2)\ldots(x - x_n)$.

Legyen $n = 2k + 1$, az alappontok pedig legyenek

$$t - kh, \ldots, t - h, t, t + h, \ldots, t + kh. \tag{10.3}$$

Ha $p(x)$ az $f(x)$ függvény ezen pontokra támaszkodó Lagrange-féle interpolációs polinomja, akkor igaz, hogy

$$|f'(t) - p'(t)| < \frac{Kh^{2k}}{\binom{2k}{k}(2k+1)}, \tag{10.4}$$

ahol K az $\left|f^{(2k+1)}(x)\right|$ korlátja a $[t-kh, t+kh]$ intervallumon.

Az $n = 3$ esetben az

$$f'(t) \approx \frac{1}{2h}\left[f(t+h) - f(t-h)\right], \tag{10.5}$$

az $n = 5$ esetben pedig az

$$f'(t) \approx \frac{1}{12h}\left[f(t-2h) - 8f(t-h) + 8f(t+h) - f(t+2h)\right] \tag{10.6}$$

közelítő formulát kapjuk.

A közelítés hibája az első esetben $O(h^2)$, a második esetben pedig $O(h^4)$. Végül megjegyezzük, hogy nagy n értékekre Lagrange-interpoláción alapuló numerikus deriválást ritkán alkalmaznak a fellépő numerikus instabilitás miatt.

10.2. Közelítés differencia hányadosokkal

A differencia hányadosok alkalmazása a közelítő deriválás legelterjedtebb módszere. Legcélszerűbben a Taylor-sorfejtés felhasználásával vezethetünk le közelítő formulákat.

Tegyük fel, hogy $f \in C^2$ és írjuk fel $f(x)$ másodfokú Taylor-polinomját:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(\xi) \quad (x < \xi < x+h).$$

Egyszerű számolással adódik, hogy

$$\frac{f(x+h) - f(x)}{h} = f'(x) + \frac{h}{2}f''(\xi),$$

ahonnan a

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} \tag{10.7}$$

közelítést kapjuk, amelynek hibája $O(h)$.

Ennél pontosabb közelítést kaphatunk, ha $f \in C^3$. Legyen

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f^{(3)}(\xi_1), \quad x < \xi_1 < x+h,$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f^{(3)}(\xi_2), \quad x-h < \xi_2 < x.$$

Kivonással és átrendezéssel kapjuk, hogy

$$\frac{f(x+h)-f(x-h)}{2h} = f'(x) + \frac{h^2}{12}[f^{(3)}(\xi_1) + f^{(3)}(\xi_2)] = f'(x) + \frac{h^2}{6}f^{(3)}(\xi_3),$$

ahol $x-h < \xi_3 < x+h$. Az ebből adódó

$$f'(x) \approx \frac{f(x+h)-f(x-h)}{2h} \tag{10.8}$$

közelítés hibája $O(h^2)$ nagyságrendű.

Magasabbrendű deriváltak

$$f^{(j)}(x) \approx \frac{1}{h^j}\sum_{i=-m}^{n} c_i f(x+ih) \tag{10.9}$$

közelítéseit hasonló módon lehet megkonstruálni. Általában (10.9) alakú közelítéseket keresünk, ahol $n+m \geq j$. Az ismeretlen c_i együtthatókat a pontossági követelményből vezethetjük le. Legyen

$$f(x+ih) = \sum_{l=0}^{j} f^{(l)}(x)\frac{i^l h^l}{l!} + O\left(h^{j+1}\right)$$

és

$$\sum_{i=-m}^{n} c_i f(x+ih) = \sum_{i=-m}^{n} c_i \left(\sum_{l=0}^{j} f^{(l)}(x)\frac{i^l h^l}{l!}\right) + O\left(h^{j+1}\right).$$

Átrendezéssel kapjuk, hogy

$$\sum_{i=-m}^{n} c_i f(x+ih) = \sum_{l=0}^{j} f^{(l)}(x)\frac{h^l}{l!}\left(\sum_{i=-m}^{n} c_i i^l\right) + O\left(h^{j+1}\right).$$

Ha most fennáll, hogy

$$\sum_{i=-m}^{n} c_i i^l = 0 \quad (0 \leq l \leq j-1), \quad \sum_{i=-m}^{n} c_i i^j = j!,$$

akkor

$$\sum_{i=-m}^{n} c_i f(x+ih) = f^{(j)}(x)h^j + O\left(h^{j+1}\right).$$

Innen az $O(h)$ pontosságú (10.9) formulát kapjuk.

A második derivált ismert közelítései az

$$f''(x) \approx \frac{f(x) - 2f(x+h) + f(x+2h)}{h^2} \tag{10.10}$$

és az

$$f''(x) \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} \tag{10.11}$$

képletek. Az utóbbi hibája $O(h^2)$. Általában is igaz, hogy az ún. *centrális differencia formulák* ($m = n$ eset) egy nagyságrenddel jobb közelítést adnak.

Közelítő parciális differencia hányadosokat hasonló módon vezethetünk le.

A következőkben megmutatjuk, hogy a deriválás instabil művelet. Legyen $f(x)$ tetszőleges differenciálható függvény. Ha ezt a függvényt a differenciálható $\eta(x)$ függvénynyel perturbáljuk, akkor a deriváltak megváltozása

$$|f'(x) - (f'(x) + \eta'(x))| = |\eta'(x)|.$$

Megmutatjuk, hogy tetszőlegesen kis $\eta(x)$ esetén is lehet $\eta'(x)$ nagy. Legyen $\eta(x) = \epsilon \sin\left(\frac{x}{\epsilon^2}\right)$, amelyre fennáll, hogy $|\eta(x)| \leq \epsilon$. Minthogy $\eta'(x) = \frac{1}{\epsilon}\cos\left(\frac{x}{\epsilon^2}\right)$, az $x = 0$ pontban a derivált megváltozása $|\eta'(0)| = \frac{1}{\epsilon}$. Ha $\epsilon \to 0$, akkor ez tetszőleges nagy lehet.

Vizsgáljuk most a perturbációk hatását a numerikus deriválás esetén. Legyen

$$D_h(f(x)) = (f(x+h) - f(x))/h.$$

Ennek hibája $|D_h(f(x)) - f'(x)| \leq (K/2)h$, ahol $K = \max_{x\in[x,x+h]} |f''(x)|$. Tegyük fel, hogy $f(x)$ helyett az $\widetilde{f}(x)$ perturbált függvénnyel számolunk, amelyre teljesül, hogy $\left|\widetilde{f}(t) - f(t)\right| \leq \epsilon/2$ ($t \in [x, x+h]$). Ekkor $f'(x)$ közelítésének hibája

$$\left|D_h\left(\widetilde{f}(x) - f(x)\right)\right| \leq \epsilon/h$$

miatt

$$\left|D_h\left(\widetilde{f}(x)\right) - f'(x)\right| = \left|D_h(f(x)) - f'(x) + D_h\left(\widetilde{f}(x) - f(x)\right)\right| \leq \frac{K}{2}h + \frac{\epsilon}{h}.$$

Ha rögzített ϵ esetén $h \to 0$, akkor a hibakorlát végtelenhez tart. Tehát h és ϵ megválasztása célszerűen nem független egymástól. A most kapott hibakorlátot a $h = \sqrt{2\epsilon/K}$ választás minimalizálja. Ekkor a korlát értéke $\sqrt{2K\epsilon}$. Ha K értéke, vagy jó becslése nem ismert, akkor a $h = c_1\sqrt{\epsilon}$ választást használjuk egy c_1 tapasztalati konstanssal. Ez választás a becslés $O(\sqrt{\epsilon})$ nagyságrendjét tekintve helyes eredményt szolgáltat. Dupla pontosságú lebegőpontos aritmetikában $\epsilon \geq \epsilon_M \approx 2.2204 \times 10^{-16}$. A $\epsilon = \epsilon_M$ esetben a hiba mértéke 10^{-8} nagyságrendű.

10.3. Numerikus differenciálás szplájnokkal

Határozzuk meg az $x_1 < x_2 < \ldots < x_n$ alappontokhoz és $y_i = f(x_i)$ $(i = 1, \ldots, n)$ függvényértékekhez tartozó természetes szplájnt:

$$S(x) = S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad (x_i \leq x \leq x_{i+1}) .$$

A szplájn deriváltja

$$S'(x) = S_i'(x) = b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2,$$

ami az $f'(x)$ egy közelítését adja, ha $x_i \leq x \leq x_{i+1}$. A közelítés hibájára $f \in C^2[a, b]$ esetén fennáll, hogy

$$|f'(x) - S'(x)| \leq K_1 \left(\max_{1 \leq i \leq n-1} h_i \right), \tag{10.12}$$

ahol $K_1 > 0$ konstans. A közelítés hibája a legnagyobb részintervallum hosszával arányos. A szplájn közelítéseknek van egy simító jellege is, amely mérsékeli a kerekítési, ill. adathibák negatív hatását. A szplájn használata akkor előnyös, ha a derivált sok pontban szükséges. A szplájnnal történő közelítések esetében is megmutatható, hogy az f függvényben bekövetkező perturbációk hatását a $\max_{1 \leq i \leq n-1} h_i \approx c_1 \sqrt{\epsilon}$ választás minimalizálja.

10.4. Feladatok

1. Igazoljuk, hogy $f \in C^4$ esetén az $f''(x) \approx \left(f(x + h) - 2f(x) + f(x - h)\right)/h^2$ közelítés hibája $O(h^2)$.
2. Becsüljük az $f'(0), f'(1)$ és $f'(10)$ értékét az

$$\left(f\left(x + h\right) - f\left(x\right)\right)/h, \quad \left(f\left(x + h\right) - f\left(x - h\right)\right)/\left(2h\right)$$

formulákkal, különböző h értékekre az e^x, $\sin(2x)$, x^4, $1/(1 + x^2)$ és a

$$\frac{\left[\cos\left(x^2 - 1\right) + x^2\right]^{1/2}}{\left(3x^2 + 17\right)^{1/3}}$$

függvények esetén!
3. Becsüljük meg a numerikus deriválás hibáját a következő eljárással! Számítsuk ki a közelítő deriváltakat a h és $h/2$ értékekre és tekintsük ezek különbségét a hibának. Mennyire pontos ez az eljárás a 2. feladat függvényei esetén?

11. fejezet

NUMERIKUS INTEGRÁLÁS

Numerikus integrálást (numerikus kvadraturát) általában akkor végzünk, ha a primitív függvény nem ismert, vagy nem állítható elő könnyen, illetve, ha az $f(x)$ függvénynek csak véges sok értéke ismert. A numerikus eljárások alapötlete sematikusan a következő:

$$f(x) \approx h(x) \Rightarrow \int_a^b f(x)dx \approx \int_a^b h(x)dx. \tag{11.1}$$

11.1. Interpolációs eljárások

Legyen adott $a \leq x_1 < x_2 < \ldots < x_n \leq b$ és $y_i = f(x_i)$ $(i = 1, \ldots, n)$. Az $f(x)$ függvényt a Lagrange-féle interpolációs polinommal közelítve kapjuk, hogy

$$\int_a^b f(x)dx \approx \int_a^b p(x)dx = \int_a^b \left[\sum_{i=1}^n y_i l_i(x)\right] dx = \sum_{i=1}^n y_i \int_a^b l_i(x)dx. \tag{11.2}$$

Az ilyen közelítéseket *interpolációs típusú kvadratura (integráló) formuláknak* nevezzük. A közelítés hibájára $f \in C^n[a,b]$ esetén a 11.1 Tétel alapján fennáll, hogy

$$R_n(f) = \int_a^b f(x)dx - \int_a^b p(x)dx = \frac{1}{n!}\int_a^b f^{(n)}(\xi_x)(x-x_1)(x-x_2)\ldots(x-x_n)dx.$$

Ha $\left|f^{(n)}(x)\right| \leq M_n$ $(x \in [a,b])$, akkor

$$|R_n(f)| = \left|\int_a^b f(x)dx - \int_a^b p(x)dx\right| \leq \frac{M_n}{n!}(b-a)^{n+1}. \tag{11.3}$$

Nagy n értékekre ez a közelítés igen rosszul viselkedhet. Ezért helyette az ún. *összetett kvadratura (integráló) formulákat* használjuk. Ezek lényege az, hogy az

$[a, b]$ intervallumot felosztjuk részintervallumokra, az egyes részintervallumokra alkalmazunk egy előre rögzített kvadraturaformulát, és az így kapott részeredményeket összegezzük. A legismertebb és legegyszerűbb eljárások a következők.

11.1.1. A trapézformula

Legyen $x_1 = a$ és $x_2 = b$. A két pontra támaszkodó elsőfokú Lagrange-féle interpolációs polinom

$$p(x) = f(x_1)\frac{x - x_2}{x_1 - x_2} + f(x_2)\frac{x - x_1}{x_2 - x_1}, \tag{11.4}$$

amelynek határozott integrálja

$$\begin{aligned}\int_{x_1}^{x_2} p(x)dx &= \left[f(x_1)\frac{(x - x_2)^2}{2(x_1 - x_2)} + f(x_2)\frac{(x - x_1)^2}{2(x_2 - x_1)}\right]_{x_1}^{x_2} \\ &= \frac{x_2 - x_1}{2}\left[f(x_1) + f(x_2)\right].\end{aligned}$$

Ha $f(x_1)$ és $f(x_2)$ előjele azonos, akkor ez az eredmény az $(x_1, 0)$, $(x_2, 0)$, $(x_2, f(x_2))$, $(x_1, f(x_1))$ pontok által határolt trapéz területe. A kapott

$$\int_a^b f(x)\,dx \approx \frac{b - a}{2}\left[f(a) + f(b)\right] \tag{11.5}$$

közelítés hibájára fennáll, hogy

$$\left|\int_a^b f(x)dx - \frac{b - a}{2}\left[f(a) + f(b)\right]\right| \leq \frac{M_2}{12}(b - a)^3. \tag{11.6}$$

Ha az $[a, b]$ intervallumot felbontjuk az

$$a = x_1 < x_2 < \ldots < x_{n+1} = b \tag{11.7}$$

pontokkal n részintervallumra, akkor az *összetett trapézformula* a következő:

$$\int_a^b f(x)dx \approx T_n(f) = \sum_{i=1}^{n} \frac{x_{i+1} - x_i}{2}\left[f(x_i) + f(x_{i+1})\right]. \tag{11.8}$$

A képlet hibájára $f \in C^2[a, b]$ esetén fennáll, hogy

$$\left|\int_a^b f(x)dx - \sum_{i=1}^{n} \frac{x_{i+1} - x_i}{2}\left[f(x_i) + f(x_{i+1})\right]\right| \leq \frac{M_2}{12}\sum_{i=1}^{n}(x_{i+1} - x_i)^3. \tag{11.9}$$

Ha az alappontok ekvidisztánsak, azaz $x_i = x_1+(i-1)h$ ($h = \frac{b-a}{n}$, $i = 1, \ldots, n+1$), akkor a képlet alakja egyszerűsödik:

$$\int_a^b f(x)dx \approx T_n(f) = \frac{h}{2}\left[f(x_1) + 2\sum_{i=2}^{n} f(x_i) + f(x_{n+1})\right]. \tag{11.10}$$

A képlet hibájára pedig $nh = b - a$ miatt fennáll, hogy

$$\left|\int_a^b f(x)dx - T_n(f)\right| \leq \frac{M_2(b-a)h^2}{12} = \frac{M_2(b-a)^3}{12n^2}. \tag{11.11}$$

11.1.2. A Simpson formula

Legyen $x_1 = a$, $x_2 = \frac{a+b}{2}$ és $x_3 = b$. Tekintsük a három pontra támaszkodó másodfokú Lagrange-féle interpolációs polinomot:

$$\begin{aligned} p(x) &= f(x_1)\frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_2-x_3)} + f(x_2)\frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)} + \\ &\quad + f(x_3)\frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)}. \end{aligned}$$

Ennek az $[a, b]$ intervallumon vett integrálja adja a következő közelítő formulát

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left[f(a) + 4f(\frac{a+b}{2}) + f(b)\right], \tag{11.12}$$

amelynek hibájára $f \in C^4[a, b]$ esetén fennáll, hogy

$$\left|\int_a^b f(x)dx - \frac{b-a}{6}\left[f(a) + 4f(\frac{a+b}{2}) + f(b)\right]\right| \leq M_4\frac{(b-a)^5}{2880}. \tag{11.13}$$

Ha az $[a, b]$ intervallumot itt is felbontjuk az

$$a = x_1 < x_2 < \ldots < x_{n+1} = b$$

pontokkal n részintervallumra, akkor az *összetett Simpson formula* a következő:

$$\int_a^b f(x)dx \approx S_n(f) = \sum_{i=1}^{n} \frac{x_{i+1}-x_i}{6}\left[f(x_i) + 4f(\frac{x_i+x_{i+1}}{2}) + f(x_{i+1})\right].$$

Ennek hibájára fennáll, hogy

$$\left|\int_a^b f(x)dx - S_n(f)\right| \leq \frac{M_4}{2880}\sum_{i=1}^{n}(x_{i+1}-x_i)^5.$$

Ha az alappontok ekvidisztánsak, azaz $x_i = x_1+(i-1)h$ ($h = \frac{b-a}{n}$, $i = 1, \ldots, n+1$), akkor a képlet alakja

$$S_n(f) = \frac{h}{6}\left[f(x_1) + 2\sum_{i=2}^{n} f(x_i) + 4\sum_{i=1}^{n} f(x_i + \frac{h}{2}) + f(x_{n+1})\right], \tag{11.14}$$

amelynek képlethibájára fennáll, hogy

$$\left|\int_a^b f(x)dx - S_n(f)\right| \leq \frac{M_4(b-a)}{2880}h^4 = \frac{M_4(b-a)^5}{2880n^4}. \tag{11.15}$$

11.2. Kvadraturaformulák hibáinak utólagos becslése

Alapgondolata egyszerű, az extrapoláció (Runge-féle szabály) általános ötletét használja. Technikailag a következőképpen járunk el. Elvégezzük a numerikus integrálást n és $2n$ részintervallum esetén. Ha fennáll, hogy

$$|T_n(f) - T_{2n}(f)| \leq \varepsilon, \tag{11.16}$$

akkor a $T_{2n}(f)$ közelítést ε pontosságúnak fogadjuk el. Ugyanezt csináljuk a Simpson-formula esetén is. Igazolhatók a következő állítások.

13.1 Tétel *(Rownland-Miel). Ha $f''(x)$ előjele állandó, akkor az összetett trapézmódszer hibájára fennáll, hogy*

$$\left|\int_a^b f(x)dx - T_{2n}(f)\right| \leq |T_n(f) - T_{2n}(f)|. \tag{11.17}$$

13.2 Tétel *(Rownland-Miel). Ha $f^{(4)}(x)$ előjele állandó, akkor az összetett Simpson-formula hibájára fennáll, hogy*

$$\left|\int_a^b f(x)dx - S_{2n}(f)\right| \leq |S_n(f) - S_{2n}(f)|. \tag{11.18}$$

11.3. Numerikus integrálás természetes szplájnokkal

Kiszámítjuk az $a = x_1 < x_2 < \ldots < x_{n+1} = b$ alappontokhoz és $y_i = f(x_i)$ $(i = 1, \ldots, n+1)$ függvényértékekhez tartozó természetes szplájnt:

$$S(x) = S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3.$$

Mármost ennek integrálja az $[x_i, x_{i+1}]$ részintervallumon

$$\int_{x_i}^{x_{i+1}} S_i(x)dx = a_i h_i + b_i\frac{h_i^2}{2} + c_i\frac{h_i^3}{3} + d_i\frac{h_i^4}{4}.$$

Ezek összegzése adja a keresett közelítő formulát

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n-1} \left(a_i h_i + b_i \frac{h_i^2}{2} + c_i \frac{h_i^3}{3} + d_i \frac{h_i^4}{4} \right). \tag{11.19}$$

A szplájn közelítés hibája (9.38) alapján

$$\left| \int_a^b f(x)dx - \int_a^b S(x)\,dx \right| \leq \int_a^b |f(x) - S(x)|\,dx = (b-a)\,K \left(\max_{1 \leq i \leq n-1} h_i \right)^2,$$

ahol $K > 0$ az $f(x)$ függvénytől függő konstans. Ha a felosztás ekvidisztáns, azaz $h_i = h = (b-a)/n$, akkor a hibakorlát értéke $K(b-a)^3/n$. Ez nagyságrendben megegyezik az összetett trapézformula hibájával.

11.4. Adaptív kvadratura eljárások

A kvadratura eljárások számítási költsége arányos az alappontok (függvénybehelyettesítések) számával. Sok esetben az ekvidisztans felosztáson alapuló eljárások indokolatlanul sok pontot igényelnek. Ennek a hátránynak a kiküszöbölését szolgálják az ún. adaptív kvadratura eljárások. Alapgondolatukat egy, a trapézformulán alapuló eljárással szemléltetjük. Az egyszerűség kedvéért tegyük fel, hogy az $f(x)$ függvény konkáv az $[a,b]$ intervallumon. Ebben az esetben az $[a,b]$ intervallum egy adott felosztása esetén a trapézformula hibáját felülről becsülhetjük a trapézok feletti háromszögek területeinek összegével, amelyeket a következő ábra mutat.

Legyen az $[x_L, x_R]$ tetszőleges részintervallum. Legyen továbbá $f_L = f(x_L)$ és $f_R = f(x_R)$. Az ábráról leolvasható, hogy a háromszög alapja

$$d = \left((f_R - f_L)^2 + (x_R - x_L)^2 \right)^{1/2},$$

magassága pedig $h = d/(\cot\alpha_L + \cot\alpha_R)$. A háromszög területe, amely egyben a trapézközelítés $B(x_L, x_R)$ hibakorlátja is, $dh/2$. Az intervallumot megfelezve (x_M) a hiba mértéke az ábrán látható módon csökken.

Az adaptív stratégiát a következőképpen fogalmazhatjuk meg. Legyen $\epsilon > 0$ előre megadott hibakorlát. Megbecsüljük egy $[x_L, x_R]$ intervallum $B(x_L, x_R)$ hibakorlátját. Ha ez kielégíti a

$$B(x_L, x_R) \leq \epsilon \frac{x_R - x_L}{b-a} \tag{11.20}$$

egyenlőtlenséget, akkor az intervallumot tovább már nem osztjuk, az intervallumot töröljük és az intervallumhoz tartozó integrálbecslést (trapézterületet) elfogadva,

hozzáadjuk az integrálközelítő összeghez. Ha a hibakorlátra vonatkozó feltétel nem teljesül, akkor az intervallumot tovább felezzük. Ha már minden intervallumot töröltünk, akkor a közelítő összeg hibájára fennáll, hogy

$$\textit{teljes hiba} \leq \sum_{[x_L,x_R] \text{ intervallumok}} B\left(x_L, x_R\right) \leq \sum_{[x_L,x_R] \text{ intervallumok}} \epsilon \frac{x_R - x_L}{b-a} = \epsilon.$$

Helyezzük el az intervallumokat egy verembe úgy, hogy a verem tetején a baloldali intervallum legyen. A következő felezésre kerülő intervallum a verem tetején lévő intervallum. Töröljük azokat az intervallumokat, amelyekre az (11.20) elfogadási feltétel teljesül, a többit pedig tegyük a verem tetejére vissza. Az eljárást akkor fejezzük be, ha a verem kiürül.

11.5. Feladatok

1. Integráljuk az $f(x) = x^{\alpha}(1.2 - x)(1 - e^{\beta|x-1|})$ függvényt a $[0, 1]$ intervallumon az $\alpha = 2$, $\beta = 0.2$ és az $\alpha = 0.1$, $\beta = 20$ paraméter értékekre a trapéz, a Simpson, a szplájn, ill. az adaptív eljárással. A megkövetelt pontosság legyen rendre $\varepsilon = 10^{-4}, 10^{-6}, 10^{-8}$. Táblázatoljuk a szükséges intervallumok és függvénybehelyettesítések (régi flopok) számát!
2. Közelítsük az alábbi integrálokat! A hibaképlettel becsüljük a szükséges alappontszámot!
   (A) $\int_0^1 \sin(x^2)dx$, $\varepsilon = 10^{-4}$.
   (B) $\int_0^1 \sqrt{x}dx$, $\varepsilon = 10^{-3}$.
   (C) $\int_0^1 \frac{e^{-x}}{0.2+4x^4}dx$, $\varepsilon = 10^{-6}$.
   (D) $\int_0^1 \cos(e^{-x} + \log(1 + x^2))dx$, $\varepsilon = 10^{-7}$.

12. fejezet

FÜGGVÉNYEK LEGJOBB EGYENLETES KÖZELÍTÉSE

Legyen az $f \in C[a,b]$ függvény adott, amelyet egy $F = F_A \in C[a,b]$ paraméteres függvénnyel közelítünk, ahol $A = [a_1, \ldots, a_n]^T \in \Omega$, $\Omega \subseteq \mathbb{R}^n$ adott paraméter halmaz. A függvényközelítés jóságát az $e = f - F_A$ hibafüggvény normájával mérjük.

14.1 Definíció. $\|.\| : C[a,b] \to \mathbb{R}$ *függvény a* $C[a,b]$ *függvényhalmazon értelmezett norma, ha minden* $f, g \in C[a,b]$ *esetén fennáll, hogy*

(i) $\|f\| \geq 0$, $\|f\| = 0 \Leftrightarrow f(x) = 0$, $\forall x \in [a,b]$

(ii) $\|\lambda f\| = |\lambda| \, \|f\|$, $\lambda \in \mathbb{R}$

(iii) $\|f + g\| \leq \|f\| + \|g\|$.

A legjobb függvényközelítés (approximáció) problémáját a következőképpen fogalmazhatjuk meg. Adott norma esetén keressük azt az $A^* \in \Omega$ paramétervektort (F_{A^*} közelítő függvényt), amelyre fennáll, hogy

$$\|f - F_{A^*}\| \leq \|f - F_A\|, \quad \forall A \in \Omega. \tag{12.1}$$

Az $F_{A^*}(x)$ megoldást *legjobb approximációnak (közelítésnek)* nevezzük.

A probléma megoldása az f függvénytől, az F_A közelítő függvényektől és az adott normától függ. Ha a norma azonos az ún.

$$\|f\|_C = \max_{x \in [a,b]} |f(x)| \tag{12.2}$$

Csebisev-normával, akkor *legjobb egyenletes közelítésről*, vagy *Csebisev-féle approximációról* beszélünk. Ezt az ún. legjobb egyenletes közelítést két F_A függvényosztályra vizsgáljuk.

Lineáris approximációról beszélünk, ha

$$F_A(x) = \sum_{i=1}^{n} a_i \phi_i(x), \tag{12.3}$$

ahol $\{\phi_i\}_{i=1}^n \subseteq C[a,b]$ adott bázisfüggvények. Feltételezzük, hogy a $\phi_1, \ldots, \phi_n$ bázisfüggvények lineárisan függetlenek. Ez azt jelenti, hogy

$$\sum_{i=1}^{n} c_i \phi_i(x) = 0, \forall x \in [a,b] \tag{12.4}$$

esetén szükségképpen teljesül, hogy $c_1 = \ldots = c_n = 0$. Igaz a következő

14.1 Tétel. *Ha $\{\phi_i\}_{i=1}^n \subset C[a,b]$ lineárisan függetlenek, akkor minden $f \in C[a,b]$ esetén létezik legjobban közelítő $F_{A^*} = \sum_{i=1}^n a_i^* \phi_i$ függvény.*

Számos esetben megmutatható, hogy a legjobban közelítő függvény egyértelmű is.

12.1. Legjobb egyenletes approximáció polinomokkal

Vizsgáljuk most folytonos függvények legjobb polinom approximációját a Csebisev-norma esetén. Legyen P_n a legfeljebb n-edfokú polinomok halmaza. Minden n-ed fokú $p(x) = a_0 + a_1 x + \ldots + a_n x^n \in P_n$ polinom előáll

$$p(x) = \sum_{i=0}^{n} a_i \phi_i(x) \tag{12.5}$$

alakban, ahol $\phi_i(x) = x^i$ $(i = 0, \ldots, n)$. Minthogy a $\{\phi_i(x)\}_{i=0}^n$ függvények lineárisan függetlenek, a $p^*(x) \in P_n$ legjobb egyenletes polinom approximáció létezik, azaz fennáll, hogy

$$E_n(f) = \|f - p^*\|_C \leq \|f - p\|_C, \quad p(x) \in P_n. \tag{12.6}$$

Az $E_n(f)$ mennyiség az f függvény legjobb n-ed fokú polinom közelítésének hibája a Csebisev normában. A legjobb egyenletes közelítés tulajdonképpen azt jelenti, hogy az $f - p$ eltérésfüggvény maximuma a $p^*(x)$ legjobb approximációs polinomra a legkisebb.

Példa. $f \in C[a,b]$ és legyen $n = 0$. Keressük tehát a legjobban közelítő konstans függvényt. Legyen $M = \max_{x\in[a,b]} f(x)$ és $m = \min_{x\in[a,b]} f(x)$. Legyen $a_0 = (m+M)/2$. Ez a keresett legjobban közelítő konstans, amelyre $E_0(f) = (M-m)/2$.

Példa. Legyen $f \in C^2[a,b]$ és tegyük fel, hogy $f(x)$ konvex, azaz $f''(x) \geq 0$ $(x \in [a,b])$. Határozzuk meg az $(a, f(a))$ és $(b, f(b))$ pontokat összekötő szelőt, valamint a görbe ezzel párhuzamos c pontbeli érintőjét $(a < c < b)$. Az $f(x)$ görbe ezen két párhuzamos egyenes között van. Tekintsük azt a velük párhuzamos

egyenest, amely a két egyenestől pontosan egyenlő távolságra van! Ez lesz az $f(x)$ függvény legjobban approximáló elsőfokú polinomja.

A legjobb egyenletes polinom approximációknak a következő fontos tulajdonságai vannak.

14.2 Tétel *(Weierstrass). Legyen $f \in C[a,b]$ adott. Minden $\epsilon > 0$ esetén létezik egy $p(x)$ polinom, hogy $\|f-p\|_C < \epsilon$.*

A tétel következménye, hogy $E_n(f) \to 0$, ha $n \to +\infty$. A konvergencia sebességét, ill. a legjobb approximáció mértékét jellemzi a

14.3 Tétel *(Jackson). Ha $f \in C[a,b]$ k-szor folytonosan differenciálható, akkor*

$$E_n(f) \leq \frac{A_k (b-a)^k}{n^k} c_k(n), \tag{12.7}$$

ahol A_k csak k-tól függő állandó és $c_k(n) \to 0$, ha $n \to \infty$.

14.4 Tétel *(Csebisev, de la Vallée-Poussin). A $p^*(x) \in P_n$ polinom akkor és csak akkor a legjobban közelítő polinom, ha létezik $n+2$ pont: $a \leq x_1 < x_2 < \ldots < x_{n+2} \leq b$, úgy, hogy*

$$|f(x_i) - p^*(x_i)| = \|f-p^*\|_C \quad (i = 1, \ldots, n+2), \tag{12.8}$$

$$f(x_i) - p^*(x_i) = -[f(x_{i+1}) - p^*(x_{i+1})] \quad (i = 1, \ldots, n+1). \tag{12.9}$$

A Csebisev tételből következik a legjobban approximáló polinom unicitása is. A tételbeli $\{x_i\}_{i=1}^{n+2}$ pontokat *alternáló ponthalmaznak* nevezzük. Vegyük észre, hogy az előbbi két példában éppen az alternáló ponthalmazokat határoztuk meg.

A Csebisev tétel lehetővé teszi a legjobban approximáló polinom meghatározását. Számos algoritmus kiindulópontja a legjobb approximációs feladat megoldása egy véges ponthalmazon.

Legyen $X_m = \{x_i\}_{i=1}^m$ véges ponthalmaz az $[a,b]$ intervallumban. A $p^*(x) \in P_n$ polinom az $f(x)$ függvény diszkrét legjobb approximációja, ha

$$E_n(f, X_m) = \max_{x \in X_m} |f(x) - p^*(x)| \leq \max_{x \in X_m} |f(x) - p(x)|, \quad p \in P_n. \tag{12.10}$$

Igazolható, hogy $m \geq n+2$ esetén létezik pontosan egy legjobb approximáció és (legalább) egy $X_{n+2}^* \subseteq X_m$ alternáló ponthalmaz úgy, hogy

$$f(x_j^*) - p^*(x_j^*) = (-1)^j s E_n(f, X_m), \quad x_j^* \in X_{n+2}^*, \ j = 1, \ldots, n+2, \ s = \pm 1. \tag{12.11}$$

Ha $m = n+2$, akkor az alternáló tulajdonság alapján könnyen meghatározhatjuk a legjobban approximáló polinomot, ui. $p^*(x)$ együtthatói és az $sE_n(f, X_{n+2})$ menynyiség kielégítik a

$$\sum_{i=0}^{n} a_i x_j^i + (-1)^j s E_n(f, X_{n+2}) = f(x_j) \quad (j = 1, \ldots, n+2) \tag{12.12}$$

lineáris egyenletrendszert. Az $m > n+2$ esetben a legjobban közelítő polinom meghatározása többféleképpen is lehetséges. Az

$$E_n(f, X_m) = \max_{X_{n+2} \subseteq X_m} E_n(f, X_{n+2}) \tag{12.13}$$

összefüggés alapján kiválasztjuk az összes $X_{n+2} \subseteq X_m$ részhalmazt, ezekre meghatározzuk a legjobban approximáló polinomot, ill. a legjobb approximáció mértékét, az
$E_n(f, X_{n+2})$-t. Az X_m halmazon legjobban approximáló polinomot, ill. X^*_{n+2} alappontrendszert azon X_{n+2} alappontrendszer határozza meg, amelyre fennáll

$$\max_{X_{n+2} \subseteq X_m} E_n(f, X_{n+2}) = E_n\left(f, X^*_{n+2}\right). \tag{12.14}$$

Ha ezt a maximumot több X_{n+2} alappontrendszer is eléri, akkor tetszés szerint lehet közülük választani. Tekintettel arra, hogy $\binom{m}{n+2}$ különböző eset van, a diszkrét approximációs feladat fenti megoldása sok számítást igényelhet. Újabban a lineáris programozási eljárást javasolják a fenti feladat megoldására. Legyen

$$\rho = \max_{x \in X_m} |f(x) - p(x)|.$$

Ez ekvivalens az

$$-\rho \leq f(x_j) - \sum_{i=0}^{n} a_i x_j^i \leq \rho \quad (j = 1, \ldots, m)$$

feltétellel. Az approximációs feladat tehát átmegy a

$$\begin{array}{rcl} \rho & \to & \min \\ \rho + \sum_{i=0}^{n} a_i x_j^i & \geq & f(x_j) \quad (j = 1, \ldots, m) \\ \rho - \sum_{i=0}^{n} a_i x_j^i & \geq & -f(x_j) \quad (j = 1, \ldots, m) \end{array}$$

lineáris programozási feladatba, amelynek megoldására rendkívül hatékony eljárások és programok ismeretesek.

Remez algoritmusával az $f \in C[a,b]$ függvény $[a,b]$ intervallumbeli (nem diszkrét) approximálása úgy történik, hogy az $[a,b]$ intervallumot felosztjuk az $x_k = a + k\frac{b-a}{N}$ $(k = 0, \ldots, N)$ pontokkal és erre meghatározzuk a legjobban közelítő diszkrét polinom approximációt. Megmutatható, hogy ha N elég nagy, akkor a diszkrét polinom approximáció jól közelíti az elméletileg legjobb polinom approximációt. Gyakorlati célokra a Remez eljárás gyorsabb konvergenciájú változatait használják.

12.2. Legjobb egyenletes approximáció racionális törtfüggvényekkel

Legyen $R(m,n)$ mindazon $r(x) = p(x)/q(x)$ alakú racionális törtfüggvények halmaza, ahol $p(x) \in P_m$, $q(x) \in P_n$ és a $p(x)$ és $q(x)$ polinomoknak nincs közös zérushelyük. Az $f \in C[a,b]$ függvények

$$r(x) = \frac{a_0 + a_1 x + \ldots + a_m x^m}{b_0 + b_1 x + \ldots + b_n x^n} \tag{12.15}$$

alakú legjobb approximációját keressük Csebisev normában. Jegyezzük meg, hogy a polinom közelítésektől eltérően $R(m,n)$ nem részhalmaza $C[a,b]$-nek. Igazak a következő eredmények.

14.5 Tétel. *Ha $f \in C[a,b]$, akkor létezik legjobb egyenletes racionális approximáció, azaz olyan $r^*(x) \in R(m,n)$ racionális törtfüggvény, hogy*

$$\|f - r^*\|_C \leq \|f - r\|_C, \quad r(x) \in R(m,n). \tag{12.16}$$

14.2 Definíció. *Az $a \leq x_1 < x_2 < \ldots < x_N \leq b$ pontokat (az $f(x) - r(x)$ hibafüggvényre nézve) alternáló pontoknak nevezzük, ha*

$$\begin{array}{c} |f(x_j) - r(x_j)| = \|f - r\|_C \quad (j = 1, \ldots, N) \\ f(x_j) - r(x_j) = -[f(x_{j+1}) - r(x_{j+1})] \quad (j = 1, \ldots, N-1) \end{array}. \tag{12.17}$$

14.6 Tétel. *Legyen adott $f(x) \in C[a,b]$. Az $r(x) = p(x)/q(x) \in R(m,n)$ racionális törtfüggvény akkor és csak akkor az f függvény legjobb egyenletes approximációja, ha az $f(x) - r(x)$ függvénynek létezik egy $N = 2 + \max\{n + \partial p, m + \partial q\}$ pontból álló alternáló ponthalmaza (∂p a $p(x)$ fokszáma, ∂q a $q(x)$ fokszáma).*

14.7 Tétel. *A legjobb egyenletes racionális approximáció egyértelmű.*

A racionális törtfüggvények sok esetben a polinomoknál lényegesen jobb approximációt szolgáltatnak. Ezért széles körben használják speciális függvények kiszámítására.

Legyen $E_{m,n} = E_{m,n}(f,[a,b]) = \|f - r^*\|_C$ a legjobb racionális approximáció hibája. Newman igazolta, hogy $n > 4$ és $f(x) = |x|$ esetén $E_{n,n}(f,[-1,1]) \leq 3e^{-\sqrt{n}}$, ha n páros és $E_{n+1,n-1}(f,[-1,1]) \leq 3e^{-\sqrt{n}}$, ha n páratlan. Polinomok esetén $E_n(f) > c/n$, ahol c konstans. Minthogy elég nagy n-re $3e^{-\sqrt{n}} \ll c/n$, a racionális approximáció hibája lényegesen kisebb, mint a polinom approximációé.

A legjobb racionális approximációt a Remez-módszer általánosításaival lehet meghatározni. Ezekkel itt nem foglalkozunk.

12.3. A Padé-approximáció

Függvények Padé-approximációja egy racionális közelítő függvény, amelyet az $x = 0$ pont környezetében sorfejtés segítségével definálunk. A Padé-approximáció általában nem azonos a legjobb racionális approximációval. Tegyük fel, hogy $f(x) \in C[a,b]$ Taylor-sora

$$f(x) = \sum_{j=0}^{\infty} c_j x^j \tag{12.18}$$

és legyen

$$R_{mk}(x) = \frac{p(x)}{q(x)}, \tag{12.19}$$

ahol $p(x) = a_0 + a_1 x + \ldots + a_m x^m$, $q(x) = b_0 + b_1 x + \ldots + b_k x^k$ és $b_0 \neq 0$. Az általánosság megszorítása nélkül feltehetjük, hogy $b_0 = 1$. Vizsgáljuk az

$$f(x) - \frac{p(x)}{q(x)} = \frac{\left(\sum_{j=0}^{\infty} c_j x^j\right)\left(\sum_{j=0}^{k} b_j x^j\right) - \sum_{j=0}^{m} a_j x^j}{\sum_{j=0}^{k} b_j x^j} \tag{12.20}$$

különbséget. Előírjuk, hogy az eltérés mértéke

$$f(x) - \frac{p(x)}{q(x)} = O\left(x^{m+k+1}\right) \tag{12.21}$$

legyen. Ezt úgy lehet elérni, hogy teljesül

$$\left(\sum_{j=0}^{\infty} c_j x^j\right)\left(\sum_{j=0}^{k} b_j x^j\right) - \sum_{j=0}^{m} a_j x^j = \sum_{j=m+k+1}^{\infty} d_j x^j. \tag{12.22}$$

Tehát a baloldalon x első $m+k+1$ hatványának együtthatója el kell hogy tűnjön. Minthogy

$$\left(\sum_{j=0}^{\infty} c_j x^j\right)\left(\sum_{j=0}^{k} b_j x^j\right) = \sum_{t=0}^{\infty}\left(\sum_{i=0}^{\min\{t,k\}} c_{t-i} b_i\right) x^t,$$

az első $m+k+1$ együttható akkor lesz zérus, ha

$$\sum_{i=0}^{\min\{t,k\}} c_{t-i} b_i = a_t \quad (t = 0, \ldots, m), \tag{12.23}$$

$$\sum_{i=0}^{\min\{t,k\}} c_{t-i} b_i = 0 \quad (t = m+1, \ldots, m+k). \tag{12.24}$$

Ha ennek az egyenletrendszernek van megoldása, akkor $R_{m,k}(x)$ az f függvény (m,k) indexű Padé-féle approximációja. Noha a Padé-approximáció általában nem azonos a legjobb racionális közelítő függvénnyel, sok esetben igen jó közelítést szolgáltat az $x=0$ pont egy környezetében (tehát lokálisan, mint a Taylor-sor).

Példa. Legyen $f(x) = 1 - \frac{1}{2}x + \frac{1}{3}x^2 + \ldots$. Ekkor $m = k = 1$, $b_0 = 1$, $c_0 = 1$, $c_1 = -\frac{1}{2}$, $c_3 = \frac{1}{3}$ és a vonatkozó egyenletrendszer

$$\begin{aligned} c_0 b_0 &= 1 = a_0, \\ c_1 b_0 + c_0 b_1 &= -\frac{1}{2} + b_1 = a_1, \\ c_2 b_0 + c_1 b_1 &= \frac{1}{3} - \frac{1}{2} b_1 = 0. \end{aligned}$$

Ennek megoldása $a_0 = 1$, $a_1 = \frac{1}{6}$ és $b_1 = \frac{2}{3}$. Tehát a megfelelő Padé közelítés

$$R_{11}(x) = \frac{1 + \frac{1}{6}x}{1 + \frac{2}{3}x},$$

amelynek hibája $O(x^3)$.

Példa. Összehasonlítjuk az e^x függvény néhány, a $[0,1]$ intervallumon vett közelítését. Az e^x másodfokú legjobb egyenletes polinomközelítése $p(x) = 1.008934 + 0.855897x + 0.844519x^2$, ahol az együtthatók pontossága 10^{-6}. A további közelítések: az

$$R_{22}(x) = \frac{12 + 6x + x^2}{12 - 6x + x^2}$$

Padé-approximáció, a $q(x) = 1 + x + \frac{1}{2}x^2$ másodfokú Taylor-polinom és az $x_1 = 0$, $x_2 = \frac{1}{2}$, $x_3 = 1$ alappontokra támaszkodó Lagrange-féle interpolációs polinom $r(x) = 1 + 0.876604x + 0.841679x^2$. Az egyes közelítések előjeles hibafüggvényét (*e^x-közelítés*) mutatja az ábra.

Itt látható, hogy a legjobb közelítést a Padé-approximáció adja ($|e^x - R_{22}(x)| \leq 3.996 \times 10^{-3}$). A Padé-féle közelítés hibája nő, amint x távolodik nullától. A Taylor polinom adja a legrosszabb eredményt ($|e^x - q(x)| \leq 0.2182$ és hibája nő, amint x távolodik az origótól. A Csebisev-féle polinomközelítésre fennáll, hogy $|e^x - p(x)| \leq 8.934 \times 10^{-3}$. A hibafüggvényen a 4 alternáló pont leolvasható. A Lagrange-féle interpolációs polinom közelítésének hibája: $|e^x - r(x)| \leq 1.4421 \times 10^{-2}$. Vegyük észre, hogy $r(x)$ együtthatói igen közel vannak $p(x)$ megfelelő együtthatóihoz. Az $r(x)$ esetén azonban csak 2 alternáló pont van.

12.4. Elemi függvények kiszámítási módjai

Az elemi függvények kiszámítására számos módszer ismeretes. Lényeges szempont az algoritmus megbízhatósága, pontossága, numerikus stabilitása és végrehajtásának gyorsasága, ideje. A leggyakrabban használt technikák között megemlíthetjük

a Taylor-sorfejtést, a Padé-approximációt, a legjobb egyenletes approximációkat, iteratív módszereket és ezek különféle kombinációit. Egyes közelítési technikák kihasználják a függvények különféle speciális (pl. periódikus) tulajdonságait is. A ma használt eljárások már olyan gyorsak, hogy a számítási idők összevethetők a multiplikatív műveletek számítási idejével.

Néhány ismert algoritmust a technikák illusztrálására bemutatunk.

1. $\ln x$ számítása ($x > 0$): Az x szám előáll $x = 2^p X$ alakban, ahol $\frac{1}{2} \leq X \leq 1$. Tehát $\ln x = p \ln 2 + \ln X$. Ezzel redukáltuk a feladatot az $\ln X$ kiszámítására, amelyhez a

$$\ln X = 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \left(\frac{X-1}{X+1} \right)^{2k+1} \tag{12.25}$$

végtelen sort használjuk. Ha $\frac{1}{2} \leq X \leq 1$, akkor a sor tagjai 3^{-k}-nál gyorsabban csökkennek. A konvergencia gyorsítására vezessük be az $X = y/\lambda$ változót! Ekkor

$$\ln y = 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \left(\frac{y-1}{y+1} \right)^{2k+1}. \tag{12.26}$$

A λ tényezőt az $u = \max_{1/2 \leq X \leq 1} \left| \frac{y-1}{y+1} \right| \to \min$ feltételből határozzuk meg. Eszerint $\lambda = \sqrt{2}$, amikor is $u < 0.172$. Tehát a sor konvergencia sebessége növelhető. Az $\ln y = \ln X + \ln \sqrt{2}$ képlet figyelembevételével

$$\ln x = \left(p - \frac{1}{2} \right) \ln 2 + 2v \sum_{k=0}^{\infty} \frac{1}{2k+1} v^{2k}, \quad v = \frac{\sqrt{2}X - 1}{\sqrt{2}X + 1}. \tag{12.27}$$

A $\sum_{k=0}^{\infty} \frac{1}{2k+1} v^{2k}$ sorfejtést helyettesíthetjük az egyenletesen legjobban közelítő, vagy hozzá közeli polinommal a v megfelelő tartományán. A hiba nem nagyobb mint 3×10^{-8}, ha az alábbi képletet használjuk

$$\ln x \approx (p - 0.5) \ln 2 + v \left(2.000000815 + 0.666445069 v^2 + 0.415054254 v^4 \right) \tag{12.28}$$

2. e^x számítása racionális törttel: Ha $|x| \leq \frac{\ln 2}{2}$, akkor a következő $(3,3)$ indexű Padé-közelítés 9×10^{-9} pontosságot biztosít:

$$\exp(x) \approx \frac{120 + 60x + 12x^2 + x^3}{120 - 60x + 12x^2 - x^3}. \tag{12.29}$$

3. $\sqrt{x}$ számítása Newton-iterációval:

$$y_{k+1} = \frac{1}{2} \left(y_k + \frac{x}{y_k} \right), \quad k = 0, 1, \ldots \tag{12.30}$$

ahol $y_0 \geq \sqrt{x}$. Az iteráció hibájára fennáll, hogy

$$\left|\sqrt{x} - y_{k+1}\right| = \frac{\left|\sqrt{x} - y_k\right|^2}{2\left|y_k\right|} \leq \frac{\left(\sqrt{x} - y_0\right)^2}{2\sqrt{x}} \leq 2\sqrt{x}\left(\frac{\sqrt{x} - y_0}{2\sqrt{x}}\right)^{2^{k+1}}. \qquad (12.31)$$

Az eljárás rendkívül gyorsan (kvadratikusan) konvergál, ha $|y_0 - \sqrt{x}| < 2\sqrt{x}$. Ha ez utóbbi feltétel nem teljesül, akkor az eljárás eleinte lineárisan konvergál, majd begyorsul. Az iterációs eljárás részletes elemzését a későbbiekben végezzük el.

Egy tetszőleges $x > 0$ szám négyzetgyökének tényleges kiszámítása a következőképpen történhet. A szám felírható az $x = 2^{2m}X$ alakban, ahol $\frac{1}{4} \leq X \leq 1$. Ezért $\sqrt{x} = 2^m\sqrt{X}$. A $\sqrt{x}$ függvényt az $\left[\frac{1}{4}, 1\right]$ intervallumon közelítjük. Például az egyenletesen legjobban közelítő elsőfokú polinom $p^*(x) = \frac{17}{48} + \frac{2}{3}x$. Ennek maximális eltérése $\sqrt{x}$-től $\frac{1}{48}$. A (12.30) iterációt ezzel az $y_0 = \frac{17}{48} + \frac{2}{3}x$ értékkel kezdjük.

4. Az IEEE lebegőpontos aritmetikai szabványt kielégítő és a matematikai koprocesszort kihasználó eljárás e^x kiszámítására a következő:

i. Redukáljuk az x értéket a $\left[-\log\frac{2}{64}, \log\frac{2}{64}\right]$ intervallumra úgy, hogy fennálljon

$$x = (32m + j)\log\frac{2}{32} + R, \quad |R| \leq \log\frac{2}{64}. \qquad (12.32)$$

ii. Approximáljuk az $\exp(R) - 1$ értéket a $p(R)$ polinommal, ahol

$$p(t) = t + a_1t^2 + \ldots + a_nt^{n+1}. \qquad (12.33)$$

iii. Rekonstruáljuk $\exp(x)$-et az

$$\exp(x) = 2^m\left(2^{j/32} + 2^{j/32}p(R)\right) \qquad (12.34)$$

képlettel.

Az algoritmus gyors végrehajtásának érdekében a $2^{j/32}$ értékek nagy pontossággal táblázatolva vannak.

5. A 2^x $(0 < x < 1)$ kiszámítása CORDIC iterációval. Legyen $c_k = \log_2\left(1 + 2^{-k}\right)$ $(k = 1, 2, \ldots)$. Legyen $x_0 = x$ és képezzük a következő sorozatot:

$$\begin{aligned} x_{i+1} &= x_i \quad \text{és} \quad \alpha_{i+1} = 0, \text{ ha } x_i < c_{i+1}, \\ x_{i+1} &= x_i - c_i \quad \text{és} \quad \alpha_{i+1} = 1, \text{ ha } x_i \geq c_{i+1}. \end{aligned}$$

Ekkor

$$2^x = \prod_{k=1}^{\infty}\left(1 + 2^{-k}\right)^{\alpha_k},$$

és

$$2^x \approx \prod_{k=1}^{n}\left(1 + 2^{-k}\right)^{\alpha_k}, \quad \varepsilon_n = 2^{2-n} - 1 \approx (\ln 2)\,2^{-n}.$$

A CORDIC eljárások a most bemutatotthoz hasonló jellegű iteratív algoritmusok, amelyeket koprocesszorokban, ill. tudományos kalkulátorokban használnak. Az eljárások alapötlete Henry Briggstől (1561-1631) származik.

Végül megjegyezzük, hogy a szűkebb intervallumra való visszavezetések általában pontosságvesztést okoznak. A szabványos eljárások ezt kezelik.

12.5. Feladatok

1. Határozza meg az $f(x) = e^x$ ($x \in [0,1]$) legjobban közelítő elsőfokú polinomját!
2. Határozza meg az $f(x) = |x|$ függvényt legjobban approximáló másodfokú polinomot az $X_5 = \left\{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\right\}$ ponthalmazon ($p^*(x) = x^2 + \frac{1}{8}$).
3. Az $f(x) = e^x$ függvény $(2,2)$ indexű Padé-féle közelítése

$$R_{22}(x) = \frac{12 + 6x + x^2}{12 - 6x + x^2}.$$

Igazoljuk, hogy ez tényleg a $(2,2)$ indexű Padé-approximáció! Mekkora a maximális hibája a $[-1,1]$ és a $[-3,3]$ intervallumon? Ábrázoljuk a hibafüggvényt!

13. fejezet

FÜGGVÉNYEK LEGKISEBB NÉGYZETES KÖZELÍTÉSE

Függvények legkisebb négyzetes közelítéséről, approximációjáról akkor beszélünk, ha a $C[a,b]$ halmazon értelmezett (L_2-) norma a következő

$$\|f\|_2 = \left(\int_a^b f^2(x)\, w(x)\, dx \right)^{\frac{1}{2}}, \tag{13.1}$$

ahol a $w(x) \in C[a,b]$ rögzített súlyfüggvényről kikötjük, hogy $w(x) > 0$, $\forall x \in [a,b]$. Fontos speciális eset: $w(x) \equiv 1$.

Diszkrét legkisebb négyzetes közelítésről beszélünk, ha a (diszkrét L_2-, vagy l_2-) norma a következő

$$\|f\|_2 = \left(\sum_{i=1}^{m} f^2(x_i)\, w(x_i) \right)^{\frac{1}{2}}, \tag{13.2}$$

ahol $a \leq x_1 < x_2 < \ldots < x_m \leq b$ adott pontok, a $w(x)$ súlyfüggvény pedig kielégíti a $w(x_i) > 0$, $x_i \in X_m = \{x_i\}_{i=1}^m$ feltételeket.

Az L_2-, ill. l_2-norma jelentése más mint a Csebisev-normáé. Ennek illusztrálására tekintsük a $[0,1]$ intervallumon adott $f_1(x) = 1$, $f_2(x) = 1 + 0.5\sin(6\pi x)$ és $f_3(x) = 1 + 10e^{-100x}$ függvények Csebisev- és L_2-normáit a $w(x) = 1$ súlyfüggvénnyel:

	$\|.\|_C$	$\|.\|_2$
$f_1(x)$	1	1
$f_2(x)$	1.5	1.0606
$f_3(x)$	11	1.3038

Látható, hogy nagy függvényérték változás erősen megváltoztatja a Csebisev-normát, míg az L_2-normát alig. Ez a tulajdonság motiválja az L_2-approximáció (Fourier-sorfejtés) felhasználását is.

Igaz a következő

15.1 Tétel. *Legyen $f \in C[a,b]$ adott. Legyen $\{\phi_i(x)\}_{i=1}^n \subset C[a,b]$ lineárisan független. Létezik pontosan egy $F_{\hat{A}}(x) = \sum_{i=1}^n \hat{a}_i \phi_i(x)$ legjobban közelítő függvény, amelyre*

$$\left\| f - \sum_{i=1}^n \hat{a}_i \phi_i \right\|_2 \leq \left\| f - \sum_{i=1}^n a_i \phi \right\|_2 . \tag{13.3}$$

A legjobb approximáció létezése következik a lineáris approximáció létezésére vonatkozó tételből. Az egyértelműség a $\|.\|_2$ norma egy tulajdonságából (szigorú konvexitásából) következik. Megjegyezzük, hogy analóg tétel igaz a diszkrét l_2-norma esetén is.

Az L_2- vagy l_2-norma esetén könnyen meghatározhatjuk a legjobb approximáció $\hat{a}_i$ együtthatóit. Az $\|f - \sum_{i=1}^n a_i \phi_i\|_2 = \min$ feladat ekvivalens az

$$\left\| f - \sum_{i=1}^n a_i \phi_i \right\|_2^2 = \min \tag{13.4}$$

feladattal. Eszerint az n-változós valós

$$R(a_1, \ldots, a_n) = \int_a^b \left(f(x) - \sum_{j=1}^n a_j \phi_j(x) \right)^2 w(x)\, dx \tag{13.5}$$

függvény egyetlen minimumát kell meghatároznunk. Az $[\hat{a}_1, \ldots, \hat{a}_n]^T$ minimumhely, ha kielégíti a $\nabla R = 0$, azaz a

$$\frac{\partial R(a_1, \ldots, a_n)}{\partial a_i} = 0 \quad (i = 1, \ldots, n) \tag{13.6}$$

stacionárius egyenletrendszert. Az a_i paraméter szerint deriválva kapjuk:

$$\frac{\partial R(a_1, \ldots, a_n)}{\partial a_i} = -2 \int_a^b \left(f(x) - \sum_{j=1}^n a_j \phi_j(x) \right) \phi_i(x)\, w(x)\, dx = 0. \tag{13.7}$$

Innen átrendezéssel kapjuk, hogy

$$\sum_{j=1}^n a_j \int_a^b \phi_j(x)\, \phi_i(x)\, w(x)\, dx = \int_a^b f(x)\, \phi_i(x)\, w(x)\, dx \quad (i = 1, \ldots, n). \tag{13.8}$$

Vezessük be az

$$\langle f, g \rangle = \int_a^b f(x)\, g(x)\, w(x)\, dx, \quad f, g \in C[a,b] \tag{13.9}$$

(skalárszorzat) jelölést. Ekkor a fenti egyenletrendszer alakja

$$\sum_{j=1}^{n} a_j \underbrace{g_{ij}}_{\langle \phi_j, \phi_i \rangle} = \underbrace{c_i}_{\langle f, \phi_i \rangle} \quad (i = 1, \ldots, n), \tag{13.10}$$

amely felírható a tömörebb

$$Ga = c \tag{13.11}$$

alakban, ahol $a = [a_1, \ldots, a_n]^T$, $G = [g_{ij}]_{i,j=1}^n$, $c = [c_1, \ldots, c_n]^T$. A G mátrixot Gram-mátrixnak nevezzük.

Példa. Legyen $a = 0$, $b = 1$, $w(x) \equiv 1$ és $\phi_i(x) = x^{i-1}$ $(i = 1, \ldots, n)$. Ekkor $\phi_i(x)\,\phi_j(x) = x^{i+j-2}$, $g_{ij} = \int_0^1 x^{i+j-2} dx = 1/(i+j-1)$ és

$$G = \left[\frac{1}{i+j-1}\right]_{i,j=1}^{n}, \tag{13.12}$$

ami nem más mint az ún. Hilbert-mátrix, amely rosszul kondicionált.

Így a legkisebb négyzetes feladat megoldása reménytelen. A járható út a $\{\phi_i\}_{i=1}^n$ függvényrendszer ügyes megválasztásában rejlik. Vezessük be a következő fogalmakat.

15.1 Definíció. *A $\{\phi_i\}_{i=1}^n \subset C[a,b]$ függvényrendszer ortogonális, ha*

$$\langle \phi_i, \phi_j \rangle = 0, \quad i \neq j.$$

A rendszer ortonormált, ha még teljesül a $\langle \phi_i, \phi_i \rangle = 1$ $(i = 1, \ldots, n)$ összefüggés is.

Tegyük fel, hogy a $\{\phi_i(x)\}_{i=1}^n$ rendszer ortonormált. Ekkor a Gram-mátrix azonos az egységmátrixszal és

$$a_i = \langle f, \phi_i \rangle \quad (i = 1, \ldots, n). \tag{13.13}$$

Az egyértelmű legjobb (legkisebb) négyzetes approximáció tehát felírható az

$$\sum_{i=1}^{n} \langle f, \phi_i \rangle \, \phi_i(x) \tag{13.14}$$

alakban. Ennek elnevezése: az $f(x)$ függvény Fourier-sora, ill. annak első n-tagja.

Nevezetes ortonormált függvényrendszerek a következők.

1. $C[-\pi, \pi]$, $w(x) = 1$ esetén

$$\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\cos x, \frac{1}{\sqrt{\pi}}\sin x, \frac{1}{\sqrt{\pi}}\cos 2x, \frac{1}{\sqrt{\pi}}\sin 2x, \ldots \tag{13.15}$$

ortonormált függvényhalmaz.

2. $C[-1,1]$, $w(x)=1/\sqrt{1-x^2}$ esetén

$$\frac{1}{\sqrt{\pi}}T_0, \sqrt{\frac{2}{\pi}}T_n(x) \quad (n=1,2,\ldots) \tag{13.16}$$

ortonormált függvényhalmaz, ahol

$$T_n(x)=\cos(n\arccos x) \tag{13.17}$$

az ún. n-edik Csebisev-polinom. Az első néhány Csebisev-polinom:

$$T_0(x)=1,\ T_1(x)=x,\ T_2(x)=2x^2-1,\ T_3(x)=4x^3-3x. \tag{13.18}$$

3. $C[-1,1]$, $w(x)\equiv 1$ esetén

$$\sqrt{\frac{1}{2}}P_0, \sqrt{\frac{2n+1}{2}}P_n(x) \quad (n=1,2,\ldots) \tag{13.19}$$

ortonormált függvényhalmaz, ahol

$$P_n(x)=\frac{1}{2^n n!}\frac{d^n}{dx^n}\left(x^2-1\right)^n \tag{13.20}$$

az ún. n-edik Legendre-polinom. Az első néhány Legendre-polinom:

$$P_0(x)=1,\ P_1(x)=x,\ P_2(x)=\frac{3}{2}x^2-\frac{1}{2},\ P_3(x)=\frac{5}{2}x^3-\frac{3}{2}x. \tag{13.21}$$

Ha rendelkezésünkre áll egy ortonormált $\{\phi_i(x)\}_{i=1}^n$ függvényrendszer, akkor a legkisebb négyzetes approximáció meghatározásához (elvileg) csak az $\langle f,\phi_i\rangle$ ún. Fourier-együtthatókat kell kiszámolnunk.

Ortogonális polinomokat a QR-felbontásnál látott Gram-Schmidt eljáráshoz hasonló módon lehet meghatározni. Ez azonban numerikusan nem stabil. Helyette az alábbi háromtagú rekurziót használjuk.

Legyen

$$\tilde{\phi}_0(x)\equiv 1, \quad \tilde{\phi}_1(x)\equiv x-\frac{\left\langle x\widetilde{\phi}_0,\widetilde{\phi}_0\right\rangle}{\left\langle\widetilde{\phi}_0,\widetilde{\phi}_0\right\rangle}, \tag{13.22}$$

és

$$\tilde{\phi}_{k+1}(x)=(x-\alpha_k)\widetilde{\phi}_k(x)-\beta_k\widetilde{\phi}_{k-1}(x) \quad (k=1,2,\ldots,n-1), \tag{13.23}$$

ahol

$$\alpha_k=\frac{\left\langle x\widetilde{\phi}_k,\widetilde{\phi}_k\right\rangle}{\left\langle\widetilde{\phi}_k,\widetilde{\phi}_k\right\rangle}, \quad \beta_k=\frac{\left\langle\widetilde{\phi}_k,\widetilde{\phi}_k\right\rangle}{\left\langle\widetilde{\phi}_{k-1},\widetilde{\phi}_{k-1}\right\rangle} \quad (k=1,2,\ldots,n-1). \tag{13.24}$$

Ekkor a $\left\{\widetilde{\phi}_i(x)\right\}_{i=0}^{n}$ polinomrendszer ortogonális, amelyből a

$$\phi_n(x) = \frac{1}{\left\langle \tilde{\phi}_n, \tilde{\phi}_n \right\rangle^{1/2}} \tilde{\phi}_n(x) \quad (n = 0, 1, \ldots) \tag{13.25}$$

normálással kapjuk a $\{\phi_i(x)\}_{i=0}^{n}$ ortonormált polinomrendszert.

Ha az $\{\alpha_i, \beta_i\}_{i=0}^{n}$ együtthatók és az $\hat{a}_i = \langle f, \phi_i \rangle$ $(i = 0, \ldots, n)$ Fourier-együtthatók ismertek, akkor ortonormált polinomok esetén az

$$f(x) \approx \sum_{i=0}^{n} \hat{a}_i \phi_i(x) \tag{13.26}$$

közelítés egy x pontbeli helyettesítési értékét célszerűen a háromtagú rekurzióval számítjuk ki.

Az eddigiekhez hasonló állítások vezethetők le a diszkrét esetben is, ha definiáljuk a

$$\langle f, g \rangle = \sum_{i=1}^{m} f(x_i)\, g(x_i)\, w(x_i) \tag{13.27}$$

skalárszorzatot, ahol $x_i \in X_m = \{x_i\}_{i=1}^{m} \subset [a, b]$ adott véges ponthalmaz.

13.1. Feladatok

1. Ábrázoljuk a $[0, 1]$ intervallumon adott $f_1(x) = 1$, $f_2(x) = 1 + 0.5 \sin(6\pi x)$ és $f_3(x) = 1 + 10e^{-100x}$ függvényeket és ellenőrizzük a normájukra vonatkozó táblázatot!

2. Oldjuk meg a

$$\min_{a_i} \int_{-1}^{1} \left(e^x - a_0 - a_1 x - a_2 x^2\right)^2 dx$$

feladatot a Legendre-polinomok segítségével $(p(x) \approx 0.537x^2 + 1.104x + 0.996)$. Ábrázoljuk a két függvényt!

3. Határozzuk meg az első három Csebisev-, ill. Legendre-függvényt a háromtagú rekurzió és a Derive (vagy Maple) segítségével $(n = 3)$!

14. fejezet

A LINEÁRIS LEGKISEBB NÉGYZETEK MÓDSZERE

A legkisebb négyzetek módszere Gausstól származik. Általános alapelve a következő: Fennáll egy $y = f(x; a_1, \ldots, a_n)$ függvénykapcsolat, amelynek paramétereit nem ismerjük. Az $x_1, \ldots, x_m$ alappontokban méréseket végzünk, amelyek eredményei az $\{y_i\}_{i=1}^m$ megfigyelések ($m > n$). Elvileg fenn kellene állnia az $y_i = f(x_i; a_1, \ldots, a_n)$ $(i = 1, \ldots, m)$ összefüggésnek. Ez azonban nem teljesül, mert a mérésekre hibák rakódnak. Ezért az

$$y_i = f(x_i; a_1, \ldots, a_n) + \epsilon_i \quad (i = 1, \ldots, m) \tag{14.1}$$

feltétel teljesül, ahol az ϵ_i véletlen hiba. Az ismeretlen paramétereket úgy kell meghatározni, hogy a hibák négyzetösszege minimális legyen, azaz

$$\sum_{i=1}^{m} \epsilon_i^2 = \sum_{i=1}^{m} (y_i - f(x_i; a_1, \ldots, a_n))^2 = \min \tag{14.2}$$

teljesüljön. A feladat rokon a diszkrét legkisebb négyzetes függvényközelítéssel.

A lineáris legkisebb négyzetek problémája (LKN) a

$$b = Ax + \varepsilon, \quad A \in \mathbb{R}^{m \times n},\ b, \varepsilon \in \mathbb{R}^m,\ x \in \mathbb{R}^n,\ m > n \tag{14.3}$$

túlhatározott egyenletrendszer megoldásához kapcsolódik. A cél az $\varepsilon = b - Ax$ reziduális hiba euklideszi normájának (normanégyzetének) minimalizálása, azaz

$$\|Ax - b\|_2 \to \min. \tag{LKN}$$

A b vektort megfigyeléseknek, az A mátrixot pedig a magyarázó változók mátrixának szokás nevezni. A $\operatorname{rank}(A) < n$ esetet *ranghiányosnak* nevezzük. Igaz a következő

16.1 Tétel. Az $x \in \mathbb{R}^n$ *akkor és csak akkor az LKN-feladat megoldása, ha* $A^T(b - Ax) = 0$.

Bizonyítás. Tegyük fel, hogy $A^T r(x) = 0$. Ekkor minden $y \in \mathbb{R}^n$ esetén $r(y) = b - Ay = r(x) + A(x - y)$ és

$$\|r(y)\|_2^2 = r^T(x)\, r(x) + 2(x-y)^T \underbrace{A^T r(x)}_{=0} + \|A(x-y)\|_2^2 \geq \|r(x)\|_2^2 .$$

Tehát az $A^T r(x) = 0$ feltételt kielégítő x vektor az LKN feladat megoldása. Fordítva: tegyük fel, hogy $A^T r(x) = z \neq 0$. Igazoljuk, hogy x nem minimalizálhatja az $\|r(x)\|_2^2$ mennyiséget. Legyen $x - y = -\epsilon z$ és használjuk fel az előbbi azonosságot. Ekkor

$$\|r(y)\|_2^2 = \|r(x)\|_2^2 - 2\epsilon \|z\|_2^2 + \epsilon^2 \|Az\|_2^2 < \|r(x)\|_2^2 ,$$

ha ϵ elég kicsi. Tehát az olyan x vektorok, amelyekre $A^T r(x) \neq 0$, nem lehetnek az LKN probléma megoldásai. □

Az $A^T(b - Ax) = 0$ egyenletet az ekvivalens

$$A^T A x = A^T b \tag{14.4}$$

alakban szokás megadni, amelyet *normálegyenletnek* nevezzünk. Igaz a következő

16.2 Tétel. $A^T A$ *akkor és csak akkor pozitív definit, ha* $\operatorname{rank}(A) = n$.

Bizonyítás. Ha $\operatorname{rank}(A) = n$, akkor $x \neq 0$ esetén $Ax \neq 0$. Ebből következik, hogy $x \neq 0$ esetén $x^T A^T A x = \|Ax\|_2^2 > 0$, azaz $A^T A$ pozitív definit. Ha $\operatorname{rank}(A) < n$, akkor létezik $x_0 \neq 0$ vektor úgy, hogy $Ax_0 = 0$. Ekkor $x_0^T A^T A x_0 = 0$ egy $x_0 \neq 0$ vektorra. Tehát $A^T A$ nem pozitív definit. □

Az LKN-feladat megoldása a $\operatorname{rank}(A) = n$ esetben a legegyszerűbb. Ekkor az egyértelmű $x = \left(A^T A\right)^{-1} A^T b$ megoldást a normálegyenlet megoldásával kaphatjuk meg. A leggyakrabban javasolt numerikus eljárás a Cholesky-módszer. Ha a rendszer nagyon rosszul kondicionált, akkor más módszereket kell alkalmazni. Ilyen például a QR-felbontáson alapuló következő eljárás is.

16.3 Tétel *(QR-felbontás). Legyen* $A \in \mathbb{R}^{m\times n}$, $m \geq n$. *Ekkor létezik* $Q \in \mathbb{R}^{m\times m}$ *ortogonális mátrix, hogy*

$$Q^T A = \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{14.5}$$

ahol R *felsőháromszögmátrix és* $r_{ii} \geq 0$ *minden* i *indexre.*

Ez a tétel a 10.2 Tétel általánosítása álló téglalapmátrixokra. Ha $\operatorname{rank}(A) = n$, akkor R nemszinguláris, azaz $r_{ii} > 0$ minden i esetén. Legyen

$$Q^T A = \begin{bmatrix} R \\ 0 \end{bmatrix}, \quad Q^T b = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \quad (c_1 \in \mathbb{R}^n). \tag{14.6}$$

16.4 Tétel. *Legyen* $\operatorname{rank}(A) = n$. *Az* $x \in \mathbb{R}^n$ *vektor akkor és csak akkor az LKN feladat megoldása, ha kielégíti az*

$$Rx = c_1, \quad \|r(x)\|_2 = \|c_2\|_2 \tag{14.7}$$

egyenleteket.

Bizonyítás.

$$\begin{aligned} \|Ax-b\|_2^2 &= \left\|Q^T(Ax-b)\right\|_2^2 = \left\|Q^TAx - Q^Tb\right\|_2^2 = \\ &= \left\| \begin{bmatrix} R \\ 0 \end{bmatrix} x - \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \right\|_2^2 = \underbrace{\geq 0}_{\|Rx-c_1\|_2^2} + \|c_2\|_2^2 \geq \|c_2\|_2^2, \end{aligned}$$

ahonnan az állítás következik. □

Tulajdonképpen azt is megkaptuk, hogy

$$\min_x \|Ax-b\|_2 = \|c_2\|_2. \tag{14.8}$$

A QR-felbontással az LKN-feladat megoldása a következő lehet: előállítjuk a Q^TA és Q^Tb mennyiségeket particionált alakban és megoldjuk az $Rx = c_1$ egyenletrendszert.

A ranghiányos esetben szükségünk van $A \in \mathbb{R}^{m\times n}$ ($\operatorname{rank}(A) = r$) szinguláris érték felbontására (10.8 Tétel):

$$A = U\Sigma V^T \in \mathbb{R}^{m\times n}, \quad \Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m\times n}, \tag{14.9}$$

ahol $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ ortogonális, $\Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ és $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$.

Ha $\operatorname{rank}(A) < n$, akkor az LKN feladatnak több megoldása is van. Ezek közül ki szoktuk tüntetni azt az x megoldás vektort, amelyre $\|x\|_2 = \min$. Igaz a következő

16.5 Tétel. *Legyen* $A \in \mathbb{R}^{m\times n}$ *és* $\operatorname{rank}(A) = r \leq n$. *Ekkor*

$$x = V \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^Tb \qquad \left(\begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{n\times m} \right) \tag{14.10}$$

az LKN feladat egyetlen olyan megoldása, amelyre $\|x\|_2$ *minimális.*

Bizonyítás. Legyen

$$z = V^Tx = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \quad c = U^Tb = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \quad z_1, c_1 \in \mathbb{R}^r. \tag{14.11}$$

Ekkor

$$\begin{aligned}\|b-Ax\|_2^2 &= \left\|U^T\left(b-AVV^Tx\right)\right\|_2^2 = \left\|U^Tb-\left(U^TAV\right)V^Tx\right\|_2^2 = \\ &= \left\|\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} - \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}\right\|_2^2 = \left\|\begin{bmatrix} c_1-\Sigma_r z_1 \\ c_2 \end{bmatrix}\right\|_2^2 = \\ &= \underbrace{\geq 0}_{\|c_1-\Sigma_r z_1\|_2^2} + \|c_2\|_2^2 \geq \|c_2\|_2^2,\end{aligned}$$

ami tetszőleges z_2 esetén a $z_1 = \Sigma_r^{-1}c_1$ vektorra minimális. A $z_2 = 0$ választás minimalizálja a $\|z\|_2^2 = \|z_1\|_2^2 + \|z_2\|_2^2$ normanégyzetet és ennek következtében az $\|x\|_2 = \|Vz\|_2$ normát. □

A tétel alapján a ranghiányos esetben az LKN feladat megoldása az SVD felbontás segítségével történhet. Numerikusan stabilnak tekinthető SVD-felbontást tartalmaz a MATLAB rendszer.

14.1. Feladatok

1. Hasonlítsuk össze az LKN feladat háromféle (normálegyenlet, QR, SVD) megoldási módszerét az

$$A = \begin{bmatrix} 1 & 1 & 1 \\ \epsilon & & \\ & \epsilon & \\ & & \epsilon \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 1-3\epsilon & 1+2\epsilon \\ 1 & 1-\epsilon & 1-2\epsilon \\ 1 & 1+\epsilon & 1-2\epsilon \\ 1 & 1+3\epsilon & 1+2\epsilon \end{bmatrix}$$

mátrixok esetén különböző ϵ értékekre! A b vektor tetszőleges lehet.

2. Hasonlítsuk össze az LKN feladat megoldási módszereit (normálegyenlet, QR, SVD) az

$$A = \begin{bmatrix} 0 & 2 & 1 \\ 10^6 & 10^6 & 0 \\ 10^6 & 0 & 10^6 \\ 0 & 1 & 1 \end{bmatrix}$$

mátrix esetén tetszőlegesen választott b vektorokra.

3. Hasonlítsuk össze az LKN feladat megoldási módszereit (normálegyenlet, QR, SVD) az

$$A = \left[\frac{1}{i+j+1}\right]_{i,j=1}^{8,6} \in \mathbb{R}^{8\times 6}$$

mátrix esetén, ha a pontos megoldás $x = [1/3, 1/4, \ldots, 1/8]^T \in \mathbb{R}^6$ és $b = Ax$!

15. fejezet

NEMLINEÁRIS EGYENLETEK

Nemlineáris egyenletek és egyenletrendszerek általában részproblémaként fordulnak elő optimalizálásnál, nemlineáris differenciál- és integrálegyenletek, stb. megoldásánál. A szakaszban az

$$f(x) = 0 \quad (\ f : \mathbb{R}^n \to \mathbb{R}^n,\ n \geq 1) \tag{15.1}$$

alakú egyenletek (egyenletrendszerek) közelítő megoldási módszereit vizsgáljuk. Az $x^* \in \mathbb{R}^n$ elemet az egyenlet megoldásának nevezzük, ha $f(x^*) = 0$. Ha $n \geq 2$, akkor egyenletrendszerről beszélünk. Minden esetben feltesszük, hogy $f(x)$ folytonos.

Az $f(x) = 0$ egyenlet javasolt megoldási módszerei egy, az x^* megoldáshoz konvergáló $\{x_i\}_{i=0}^{\infty}$ sorozatot képeznek. A konvergencia sebességét az alábbi módon jellemezhetjük.

17.1 Definíció. *Az $\{x_i\}_{i=0}^{\infty} \in \mathbb{R}^n$ $(n \geq 1)$ sorozat lineáris sebességgel konvergál egy $x^* \in \mathbb{R}^n$ határértékhez, ha létezik egy $0 \leq q < 1$ konstans úgy, hogy $\|x_i - x^*\| \leq q \|x_{i-1} - x^*\|$ $(i \geq 1)$.*

A lineáris konvergencia sebesség esetén fennáll, hogy

$$\|x_i - x^*\| \leq q \|x_{i-1} - x^*\| \leq q^2 \|x_{i-2} - x^*\| \leq \ldots \leq q^i \|x_0 - x^*\| . \tag{15.2}$$

Ez azt jelenti, hogy az x_i közelítések hibáit egy nullához tartó mértani sorozat tagjaival tudjuk felülről becsülni.

17.2 Definíció. *Az $\{x_i\}_{i=0}^{\infty} \in \mathbb{R}^n$ $(n \geq 1)$ sorozat p-ed rendű sebességgel $(p > 1)$ konvergál egy $x^* \in \mathbb{R}^n$ határértékhez, ha létezik egy $\gamma > 0$ konstans úgy, hogy $\|x_i - x^*\| \leq \gamma \|x_{i-1} - x^*\|^p$ $(i \geq 1)$.*

A p-edrendű konvergencia sebesség a lineárisnál lényegesen gyorsabb. Teljes indukcióval igazolhatjuk, hogy

$$\|x_i - x^*\| \leq \frac{1}{\sqrt[p-1]{\gamma}} \left(\sqrt[p-1]{\gamma} \|x_0 - x^*\| \right)^{p^i} \quad (i \geq 1) . \tag{15.3}$$

Ha $q = \sqrt[p-1]{\gamma}\,\|x_0 - x^*\| < 1$, akkor az x_i közelítések hibáit a $\left\{cq^{p^i}\right\}$ nullához tartó sorozat becsli felülről ($c = 1/\sqrt[p-1]{\gamma}$). Ez nyilvánvalóan gyorsabban tart 0-hoz mint a cq^i mértani sorozat.

15.1. Egyváltozós egyenletek megoldása

Az $f(x) = 0$ ($f : \mathbb{R} \to \mathbb{R}$) alakú valós egyenletek megoldását vizsgáljuk. Három módszert ismertetünk.

15.1.1. Az intervallumfelező eljárás

Tegyük fel, hogy $f : \mathbb{R} \to \mathbb{R}$ folytonos az $[a, b]$ intervallumon és fennáll, hogy

$$f(a)f(b) < 0. \tag{15.4}$$

Ekkor a Bolzano-tétel miatt az $f(x) = 0$ egyenletnek van legalább egy $x^* \in (a, b)$ gyöke. Ezt a Bolzano–tétel bizonyításából ismert eljárással kaphatjuk meg. Legyen $c = (a + b)/2$ és vizsgáljuk az $f(c)$ értékét. Ha $f(a) f(c) < 0$, akkor az $[a, c]$ intervallumban van gyök. Egyébként a $[c, b]$ intervallum tartalmaz gyököt. Az új intervallumot újra megfelezzük és így tovább. Az egymásba skatulyázott zárt intervallumok ráhúzódnak az egyenlet egy gyökére. Algoritmikus formában: $[a_1, b_1] = [a, b]$, $c_i = (a_i + b_i)/2$, $[a_{i+1}, b_{i+1}] = \begin{cases} [a_i, c_i], & \text{ha } f(a_i)f(c_i) < 0 \\ [c_i, b_i], & \text{egyébként} \end{cases}$, $(i = 1, 2, \ldots)$.

Az x^* gyököt az $[a_i, b_i]$ intervallum tetszőleges y pontjával közelíthetjük. Az y közelítés hibájára fennáll, hogy

$$|x^* - y| \le \max\{y - a_i, b_i - y\}. \tag{15.5}$$

A $\max\{y-a_i, b_i-y\}$ korlát akkor a legkisebb, ha $y = \frac{a_i+b_i}{2}$. Ezért az x^* gyök i-edik közelítéseként általában az $x_i = (a_i + b_i)/2$ felezőpontot használjuk. Nyilván

$$|x^* - x_i| \le \frac{b_i - a_i}{2} = \frac{b - a}{2^i} \quad (i = 1, 2, \ldots). \tag{15.6}$$

Az algoritmust akkor állítjuk le, ha a közelítés hibája kisebb, mint egy előre megadott $\varepsilon > 0$ hibakorlát. Ezért az intervallumfelező eljárás gyakorlati formája a következő.

AZ INTERVALLUMFELEZŐ ALGORITMUS:

input $[a, b]$, $\varepsilon > 0$.
while $b - a > 2\varepsilon$
  $x = (a + b)/2$
  if $f(a)f(x) < 0$

```
      b = x
    else
      a = x
    end
  end
  x = (a + b) /2
```

Megjegyezzük, hogy az $\{x_i\}$ sorozat csak folytonos $f(x)$ esetén konvergál biztosan az x^* gyökhöz. (Egyébként még a gyök létezése sem garantált.)

Példa. Legyen $f(x) = 4(1-x^2) - e^x = 0$ és határozzuk meg a gyököket. Az egyenletnek a $[-1,0]$, ill. $[0,1]$ intervallumokon vannak gyökei, ui.

$$f(-1)\,f(0) = -3e^{-1} < 0, \quad f(0)\,f(1) = -3e < 0.$$

A felező módszer tehát alkalmazható. Az $\varepsilon = 10^{-6}$ pontosságú közelítéshez szükséges lépések számát mindkét intervallum esetén a $|x_i - x^*| \le \frac{b-a}{2^i} = \frac{1}{2^i} \le \varepsilon$ egyenlőtlenség megoldása adja. Eszerint $i \ge -\log\varepsilon/\log 2 \approx 19.93$, azaz $i = 20$ lépés szükséges ($x_1 \approx -0.950455$, $x_2 = 0.703439$).

15.1.2. A fixpont iterációs módszer

A módszert az $f(x) = x - g(x) = 0$ alakú vagy ilyen alakra hozott egyenletek esetén alkalmazzuk. Az $f(x) = 0$ egyenlet ekvivalens az

$$x = g(x) \tag{15.7}$$

egyenlettel. Az x^* pontot a $g(x)$ leképezés fixpontjának nevezzük, ha $x^* = g(x^*)$. Leképezések fixpontjára vonatkozik a következő

17.1 Tétel. *Ha $g \in C[a,b]$ és $a \le g(x) \le b$ minden $x \in [a,b]$ esetén, akkor a $g(x)$ függvénynek az $[a,b]$ intervallumon van fixpontja.*

Bizonyítás. Feltehetjük, hogy $g(a) > a$ és $g(b) < b$. Legyen $h(x) = g(x) - x$. Ekkor $h(x)$ folytonos $[a,b]$-n és $h(a) > 0$ és $h(b) < 0$. Ezért a $h(x)$ függvénynek van $\xi \in (a,b)$ gyöke, azaz $h(\xi) = g(\xi) - \xi = 0$. Tehát ξ fixpont. □

17.3 Definíció. *A $g \in C[a,b]$ függvény kontrakció az $[a,b]$ intervallumon, ha létezik $0 \le q < 1$ úgy, hogy*

$$|g(x) - g(y)| \le q\,|x-y|, \quad x,y \in [a,b]. \tag{15.8}$$

Példa. A $g(x) = x^2$ függvény kontrakció a $\left[0, \frac{1}{4}\right]$ intervallumon, ui.

$$\left|x^2 - y^2\right| = \underbrace{|x+y|}_{\le 1/2} |x-y| \le \frac{1}{2}\,|x-y|, \quad x,y \in \left[0,\frac{1}{4}\right].$$

Példa. A $g(x) = x^2$ függvény nem kontrakció a $[0,1]$ intervallumon, ui. $x, y \in \left[\frac{3}{4}, 1\right]$ esetén

$$\left|x^2 - y^2\right| = \underbrace{|x+y|}_{\geq 3/2} |x-y| \geq \frac{3}{2} |x-y| > |x-y| \quad (x \neq y)$$

17.2. Tétel. *Ha $g \in C[a,b]$, $a \leq g(x) \leq b$ ($x \in [a,b]$) és $g(x)$ kontrakció $[a,b]$-n, akkor pontosan egy fixpont létezik $[a,b]$-ben.*

Bizonyítás. A 17.1 Tétel a fixpont létezését bizonyítja. Tegyük fel, hogy $x^*, y^* \in [a,b]$ fixpontok és $x^* \neq y^*$. Ekkor fennáll, hogy

$$|x^* - y^*| = |g(x^*) - g(y^*)| \leq q |x^* - y^*|,$$

ahonnan osztással az $1 \leq q < 1$ ellentmondást kapjuk. Tehát csak egy fixpont van. □

17.3 Tétel. *A $g \in C^1[a,b]$ függvény kontrakció az $[a,b]$ intervallumon, ha*

$$\max_{x \in [a,b]} |g'(x)| = q < 1. \tag{15.9}$$

Bizonyítás. A Lagrange-tétel alapján

$$|g(x) - g(y)| = |g'(\xi)(x-y)| = |g'(\xi)| |x-y| \leq q |x-y|. \ \square$$

A következő tétel megadja a fixpont iterációs módszert és a konvergenciájára vonatkozó feltételeket.

17.4 Tétel. *Legyen $g \in C[a,b]$ olyan, hogy $a \leq g(x) \leq b$ ($x \in [a,b]$) és tegyük fel, hogy $g(x)$ kontrakció $[a,b]$-n. Ekkor minden $x_0 \in [a,b]$ esetén az*

$$x_{i+1} = g(x_i) \quad (i = 0, 1, \ldots) \tag{15.10}$$

iteráció sorozat lineáris sebességgel konvergál az x^* fixponthoz, azaz

$$|x_i - x^*| \leq q^i |x_0 - x^*| \quad (i = 0, 1, \ldots). \tag{15.11}$$

Bizonyítás. Minthogy $g(x) \in [a,b]$ minden $x \in [a,b]$-re, azért $\{x_i\}_{i=0}^{\infty} \subset [a,b]$. Igaz a következő becslés:

$$|x_{n+1} - x_n| = |g(x_n) - g(x_{n-1})| \leq q |x_n - x_{n-1}| \leq \ldots \leq q^n |x_1 - x_0|.$$

Ennek segítségével belátjuk, hogy $\{x_i\}_{i=0}^{\infty}$ Cauchy-sorozat, azaz $|x_m - x_n| \to 0$, hacsak $m, n \to \infty$. Egyszerű becsléssel kapjuk, hogy $m > n$ esetén

$$\begin{gathered} |x_m - x_n| \leq |x_m - x_{m-1}| + \ldots + |x_{n+1} - x_n| \leq \\ \leq (q^{m-n-1} + \ldots + 1) |x_{n+1} - x_n| \leq \tfrac{1-q^{m-n}}{1-q} |x_{n+1} - x_n| \leq \\ \leq q^n \tfrac{1-q^{m-n}}{1-q} |x_1 - x_0| \leq \tfrac{q^n}{1-q} |x_1 - x_0|, \end{gathered}$$

ahonnan $m, n \to \infty$ esetén $|x_m - x_n| \to 0$ következik. Az $\{x_i\}_{i=0}^{\infty} \subset [a, b]$ Cauchy-sorozat egyúttal konvergens is, tehát létezik $x^* \in [a, b]$ határértéke. A $g(x)$ függvény folytonossága miatt fennáll a

$$\begin{array}{ccc} x_{i+1} & = & g(x_i) \\ \downarrow & & \downarrow \\ x^* & = & g(x^*) \end{array}$$

diagram helyessége. Tehát x^* fixpont. A fenti egyenlőtlenség-láncból az $x_m \to x^*$ határátmenettel kapjuk, hogy

$$|x_n - x^*| \leq \frac{q^n}{1-q} |x_1 - x_0|, \quad n \geq 0.$$

A lineáris konvergencia ebből már következik. A tételben szereplő becslést az

$$|x_n - x^*| = |g(x_{n-1}) - g(x^*)| \leq q |x_{n-1} - x^*| \leq \ldots \leq q^n |x_0 - x^*|$$

egyenlőtlenség láncból kapjuk. □

A FIXPONT ITERÁCIÓS ELJÁRÁS ALGORITMUSA:

Input x_0, $\varepsilon > 0$.
while kilépési feltétel=hamis
$x_{i+1} = g(x_i)$,
$i = i + 1$
end

Ha ismert a q érték, vagy egy jó becslése, akkor az $\varepsilon > 0$ pontosság eléréséhez szükséges iterációk számát a

$$\frac{q^n}{1-q} |x_1 - x_0| \leq \varepsilon \tag{15.12}$$

egyenlőtlenség megoldásával kaphatjuk meg. Alkalmazva az $||a| - |b|| \leq |a - b|$ és

$$|x^* - x_n| - |x_n - x_{n+1}| \leq |x^* - x_{n+1}| \leq q |x^* - x_n|$$

egyenlőtlenségeket kapjuk, hogy

$$|x^* - x_n| \leq \frac{1}{1-q} |x_{n+1} - x_n|. \tag{15.13}$$

Ha teljesül, hogy

$$|x_{n+1} - x_n| \leq (1-q)\varepsilon, \tag{15.14}$$

akkor az x_n közelítés abszolút hibája kisebb mint ε. Ekkor az iterációt leállíthatjuk, abból kiléphetünk.

Ha ismerjük a q értékét és fennáll a konvergencia, akkor a fenti egyenlőtlenségek alapján megállapíthatjuk a szükséges iterációk számát, vagy a megoldáshoz való közelséget. Általában azonban nem ez a helyzet. Vagy nem tudjuk q pontos értékét, vagy azt nem tudjuk, hogy egy adott x_0 pont olyan pont-e, amelyből indulva a konvergencia garantálható. Ennek ellenére általános az alábbi kilépési feltételek használata:

$$(B) \quad |x_{i+1} - x_i| \leq c_2\varepsilon; \qquad (C) \quad i = i_{\max}. \tag{15.15}$$

Minthogy a feltételek egyike sem garantálja az $|x_{i+1} - x^*| \leq \varepsilon$ feltétel teljesülését, célszerű az (B) és (C) feltételt együtt használni.

A 17.4 Tétel feltételeit, a kontraktivitást, de különösen az $a \leq g(x) \leq b$ feltételt általában nem könnyű biztosítani. E feltételek lokális jellegét mutatja a

17.5 Tétel. *Legyen $g \in C^1[a,b]$, a $\xi \in [a,b]$ és $\delta > 0$ olyan, hogy $|g'(x)| \leq q < 1$ teljesül minden $x \in [\xi - \delta, \xi + \delta] \subseteq [a,b]$ esetén. Ha $0 < r \leq \delta$ olyan, hogy $|g(\xi) - \xi| \leq (1-q)\,r$, akkor $g(x)$ kontrakció az $[\xi - r, \xi + r]$ intervallumon és $\xi - r \leq g(x) \leq \xi + r$, ha $x \in [\xi - r, \xi + r]$.*

Bizonyítás. A kontraktivitás a $0 < r \leq \delta$ egyenlőtlenség következménye. Másrészt legyen $x \in [\xi - r, \xi + r]$ tetszőleges. Ekkor fennáll

$$|g(x) - \xi| \leq \underbrace{|g(x) - g(\xi)|}_{\leq q\,|x - \xi|} + |g(\xi) - \xi| \leq qr + (1-q)\,r = r,$$

tehát $g(x) \in [\xi - r, \xi + r]$. □

Következmény. *Az $[\xi - r, \xi + r]$ intervallum a $g(x)$ egy fixpontját tartalmazza, amelyhez a fixpont iteráció bármely $x_0 \in [\xi - r, \xi + r]$ pontból kiindulva konvergál.*

Megjegyezzük, hogy a derivált folytonossága miatt elég feltenni a $|g'(\xi)| < 1$ teljesülését. Ebből már következik, hogy létezik $\delta > 0$, amellyel igaz a 17.5 Tétel deriváltra vonatkozó feltevése.

Példa. Oldjuk meg fixpont iterációs módszerrel az $f(x) = 4(1 - x^2) - e^x = 0$ egyenletet a $[0.68, 0.72]$ intervallumban $\varepsilon = 10^{-5}$ pontossággal! Minthogy az egyenlet pozitív gyökéről van szó, átírhatjuk az $x = g(x) = \sqrt{1 - e^x/4}$ alakba. A kijelölt intervallumon

$$g'(x) = -e^x / \left(4\sqrt{4 - e^x}\right) < 0, \quad g''(x) = e^x\,(e^x - 8) \,/ \left(8\sqrt{(4 - e^x)^3}\right) < 0.$$

Ezért $g'(x)$ monoton csökkenő és $\max_{x \in [0.7, 0.72]} |g'(x)| = |g'(0.72)| \approx 0.3682$. Legyen $\xi = 0.70$. Ekkor az $r = 0.02$ választással

$$|g(0.7) - 0.7| \approx 0.004671 \leq (1 - 0.3682)\,0.02 \approx 0.01263,$$

ami az 17.5 Tétel alapján az iterációs módszer konvergenciáját jelenti.

Az $f(x) = 0$ alakban megadott egyenletek átírása az $x = g(x)$ formára igen könnyű, ui. $x = x - f(x)$ ilyen alak. A kontraktivitás biztosítása azonban korántsem egyszerű feladat. Sok esetben az ekvivalens

$$x = x - \alpha f(x) \tag{15.16}$$

fixpont feladatot vizsgáljuk, ahol az α konstanst vagy $\alpha(x)$ függvényt úgy választjuk meg, hogy a $g(x) = x - \alpha f(x)$ függvény kontrakció legyen. Ilyen tulajdonképpen a következő szakaszban ismertetésre kerülő Newton-módszer is.

A fixpont iteráció és általánosításai fontosak további módszerek kifejlesztésében és az ún. kaotikus jelenségek (fraktálok) vizsgálatában is. Itt némi egyszerűsítéssel azt a kérdést vizsgálják, hogy az iteráció mely pontokból elindulva konvergens, illetve az iterációs sorozat hogyan viselkedik.

Végül megjegyezzük, hogy lebegőpontos aritmetikában előfordulhat, hogy a sorozat nem konvergál a fixponthoz, hanem körülötte "beciklizál".

15.1.3. A Newton-módszer

Tegyük fel, hogy $f : \mathbb{R} \to \mathbb{R}$ folytonosan differenciálható. A módszer lényege, hogy az x_i pontban a függvényhez érintőt húzunk és ennek az érintőnek a zérushelye adja meg a keresett gyök $(i+1)$-edik közelítését, azaz x_{i+1}-et. Az érintő iránytangense $f'(x_i)$ és egyenlete

$$y - f(x_i) = f'(x_i)(x - x_i). \tag{15.17}$$

Az $y = 0$ egyenlet megoldása:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}, \tag{15.18}$$

feltéve, hogy $f'(x_i) \neq 0$. E képlethez eljuthatunk egy kissé más érveléssel is. Nevezetesen $f(x)$-et linearizáljuk az x_i pontban, azaz közelítjük az elsőfokú Taylor-polinomjával:

$$f(x) \approx f(x_i) + f'(x_i)(x - x_i). \tag{15.19}$$

Ezután az $f(x) = 0$ egyenlet helyettesítjük a $f(x_i) + f'(x_i)(x - x_i) = 0$ egyenlettel, amelynek gyöke közelíti az $f(x) = 0$ egyenlet gyökét.

A Newton-módszer tehát a következő. Adott egy $x_0 \in \mathbb{R}$ kezdeti közelítés és képezzük az

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad (i = 0, 1, \ldots) \tag{15.20}$$

sorozatot. Vegyük észre: a Newton-módszer tulajdonképpen az $x = x - f(x)/f'(x)$ fixpont feladatra alkalmazott iterációs eljárás. A Newton-módszer konvergenciájára vonatkozik az alábbi

17.6 Tétel. *Legyen $f : (a,b) \to \mathbb{R}$ kétszer folytonosan differenciálható, $|f''(x)| \leq \gamma$ és $|f'(x)| \geq \rho > 0$ $(x \in (a,b))$. Ha az $f(x) = 0$ egyenletnek van egy x^* gyöke az (a,b) intervallumban, akkor van egy olyan $\eta > 0$ szám, hogy $|x_0 - x^*| < \eta$ esetén $x_i \to x^*$ $(i \to +\infty)$ és*

$$|x_{i+1} - x^*| \leq \frac{\gamma}{2\rho} |x_i - x^*|^2 \quad (i = 0, 1, \ldots). \tag{15.21}$$

Bizonyítás. Legyen $\eta_1 = \min\{x^* - a, b - x^*\} > 0$. Tekintsük a

$$f(x^*) = f(x_i) + f'(x_i)(x^* - x_i) + (1/2) f''(\xi_i)(x^* - x_i)^2$$

Taylor-sort, ahonnan $f(x^*) = 0$ miatt

$$f(x_i) = -f'(x_i)(x^* - x_i) - (1/2) f''(\xi_i)(x^* - x_i)^2$$

következik. Behelyettesítéssel kapjuk, hogy

$$x_{i+1} = x_i + (x^* - x_i) + \frac{1}{2}\frac{f''(\xi_i)}{f'(x_i)}(x_i - x^*)^2,$$

azaz

$$x_{i+1} - x^* = \frac{1}{2}\frac{f''(\xi_i)}{f'(x_i)}(x_i - x^*)^2.$$

Innen a $|f''(x)| \leq \gamma$ és $|f'(x)| \geq \rho > 0$ $(x \in (a,b))$ feltevések miatt

$$|x_{i+1} - x^*| \leq \frac{\gamma}{2\rho} |x_i - x^*|^2$$

következik, feltéve, hogy $x_i \in (a,b)$. Ha $|x_0 - x^*| < \eta = \min\{\eta_1, 2\rho/\gamma\}$, akkor

$$|x_1 - x^*| \leq \left(\frac{\gamma}{2\rho} |x_0 - x^*|\right) |x_0 - x^*| \leq |x_0 - x^*| < \eta$$

miatt $x_1 \in (x^* - \eta, x^* + \eta) \subset (a,b)$. Hasonlóan folytatva könnyen igazolhatjuk, hogy $x_i \in (a,b)$ és

$$|x_{i+1} - x^*| \leq \frac{2\rho}{\gamma} \left(\frac{\gamma}{2\rho} |x_0 - x^*|\right)^{2^{i+1}} \quad (i = 0, 1, \ldots).$$

Tehát $|x_0 - x^*| < \eta$ esetben az $\{x_i\}_{i=0}^{\infty}$ sorozat másodrendben konvergens. □

Azt mondjuk, hogy a Newton-módszer konvergenciája lokális, mert az x_1 kezdeti közelítésnek az x^* gyök "közelében" kell lennie. A Newton-módszer másodrendű

konvergenciáját az alábbi megjegyzéssel jellemezhetjük. Tegyük fel, hogy $x^* \neq 0$. Ekkor fennáll, hogy

$$\frac{|x_{i+1} - x^*|}{|x^*|} \leq \frac{\gamma |x^*|}{2\rho} \left(\frac{|x_i - x^*|}{|x^*|} \right)^2 \leq \frac{\gamma \max\{|a|, |b|\}}{2\rho} \left(\frac{|x_i - x^*|}{|x^*|} \right)^2. \qquad (15.22)$$

Ez azt jelenti, hogy az x_{i+1} közelítés relatív hibája az i-edik közelítés relatív hibájának négyzete. Ha tehát beállt a közelítés első két tizedesjegye, akkor dupla pontosságú aritmetikában 3-4 lépésben beáll az elérhető legnagyobb pontosság.

Kilépési feltételek. A Newton-módszert elvileg akkor kell megállítanunk, amikor elértünk egy adott $\varepsilon > 0$ pontosságú közelítést, azaz fennáll, hogy

$$|x_i - x^*| \leq \varepsilon. \qquad (15.23)$$

Mármost a gyök ismerete nélkül ezt a hibát ténylegesen becsülni nem tudjuk. Ezért különböző heurisztikus kilépési feltételeket használunk. A leggyakoribbak:

$$(A) \quad |f(x_i)| \leq \varepsilon_1; \quad (B) \quad |x_{i+1} - x_i| \leq \varepsilon_2; \quad (C) \quad i = i_{\max}. \qquad (15.24)$$

A feltételek egyikének teljesülése sem garantálja a (15.23) pontossági feltétel teljesülését. Ezért célszerűbb a három feltételt együtt használni.

A függvényközelítéseknél korábban látott

$$x_{k+1} = \frac{1}{2} \left(x_k + \frac{a}{x_k} \right) \quad (k = 0, 1, \ldots) \qquad (15.25)$$

négyzetgyök algoritmus nem más, mint a Newton-módszer az $f(x) = x^2 - a = 0$ egyenletre alkalmazva. Felmerül a kérédés, hogy milyen x_0 értékekre lesz az eljárás konvergens? Igaz a következő, Fourier-től származó

17.7 Tétel. *Legyen $f \in C^2[a,b]$, $f'(x) \neq 0$ és $f''(x) \neq 0$, ha $x \in [a,b]$. Tegyük fel, hogy létezik $x^* \in (a,b)$ gyök. Ha az $x_0 \in [a,b]$ pont olyan, hogy $f(x_0) f''(x_0) > 0$, akkor a Newton-módszer monoton konvergál az x^* megoldáshoz.*

A tételben szereplő $[a,b]$ végtelen intervallum is lehet. Esetünkben $f(x) = x^2 - a$ és $f''(x) = 2$. Ezért $x_0 > \sqrt{a}$ esetén $f(x_0) f''(x_0) > 0$. Tehát ilyen x_0 értékekre a Newton-módszer biztosan konvergál.

15.2. Nemlineáris egyenletrendszerek megoldása

Az $f(x) = 0$ ($f : \mathbb{R}^n \to \mathbb{R}^n$) és az $x = g(x)$ ($g : \mathbb{R}^n \to \mathbb{R}^n$) alakú egyenletrendszereket vizsgáljuk, ahol $f(x)$, ill. $g(x)$ folytonosan differenciálható. A fixpont iterációs eljárás és a Newton-módszer általánosítását tárgyaljuk.

15.2.1. Fixpont iterációs eljárás

Az $\mathbf{x} = g(\mathbf{x})$ egyenletrendszert vizsgáljuk, ahol $g : \mathbb{R}^n \to \mathbb{R}^n$ folytonos. Tegyük fel, hogy $D \subseteq \mathbb{R}^n$ zárt tartomány olyan, hogy teljesülnek rá a

$$g(\mathbf{x}) \in D, \quad \mathbf{x} \in D \tag{15.26}$$

és

$$\|g(\mathbf{x}) - g(\mathbf{y})\| \leq q\|\mathbf{x} - \mathbf{y}\|, \quad \mathbf{x}, \mathbf{y} \in D \quad (0 \leq q < 1) \tag{15.27}$$

feltételek. Ekkor tetszőleges $\mathbf{x}_0 \in D$ esetén az alábbi algoritmus lineáris sebességgel konvergál az $\mathbf{x} = g(\mathbf{x})$ egyenlet egyetlen megoldásához.

A FIXPONT ITERÁCIÓS ELJÁRÁS ALGORITMUSA:

Input $\mathbf{x}_0$, $\varepsilon > 0$.
while kilépési feltétel=hamis
$\mathbf{x}_{i+1} = g(\mathbf{x}_i)$,
$i = i + 1$
end

A használható kilépési feltételek

$$(A) \quad \|f(\mathbf{x}_i)\| \leq \varepsilon_1; \quad (B) \quad \|\mathbf{x}_{i+1} - \mathbf{x}_i\| \leq \varepsilon_2; \quad (C) \quad i = i_{\max} \tag{15.28}$$

és ezek kombinációi.

15.2.2. A Newton-módszer

Az $f(\mathbf{x}) = 0$ ($\mathbf{x} \in \mathbb{R}^n$) egyenletrendszer koordinátás alakja:

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0. \end{aligned}$$

Az $\mathbf{x}_i = [x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}]^T \in \mathbb{R}^n$ pontban linearizáljuk az $f_k(x)$ koordináta függvényt ($k = 1, \ldots, n$):

$$\begin{aligned} f_k(x_1, x_2, \ldots, x_n) \approx\ & f_k(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}) \\ & + \sum_{l=1}^{n} \frac{\partial}{\partial x_l} f_k(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}) \left(x_l - x_l^{(i)}\right). \end{aligned}$$

Tömörebb formában ugyanez

$$\begin{aligned} f_1(\mathbf{x}) &\approx f_1(\mathbf{x}_i) + \nabla f_1(\mathbf{x}_i)^T(\mathbf{x} - \mathbf{x}_i) \\ &\vdots \\ f_n(\mathbf{x}) &\approx f_n(\mathbf{x}_i) + \nabla f_n(\mathbf{x}_i)^T(\mathbf{x} - \mathbf{x}_i). \end{aligned}$$

Az $f(\mathbf{x}) = 0$ egyenletrendszer megoldása helyett keressük a linearizált egyenletrendszer

$$\begin{aligned} f_1(\mathbf{x}_i) + \nabla f_1(\mathbf{x}_i)^T(\mathbf{x} - \mathbf{x}_i) &= 0 \\ &\vdots \\ f_n(\mathbf{x}_i) + \nabla f_n(\mathbf{x}_i)^T(\mathbf{x} - \mathbf{x}_i) &= 0 \end{aligned} \tag{15.29}$$

közös megoldását, amely az $\mathbf{x}_{i+1}$ új közelítést definiálja. Vegyük észre, hogy az $y = f_k(\mathbf{x}_i) + \nabla f_k(\mathbf{x}_i)^T(\mathbf{x} - \mathbf{x}_i)$ egyenlet az $y = f_k(\mathbf{x})$ függvény érintősíkja az $\mathbf{x}_i$ pontban. Az $\mathbf{x}_{i+1}$ közelítés az érintősíkoknak az $y = 0$ síkon lévő közös pontja. Ha bevezetjük az

$$J(\mathbf{x}) = \left[\frac{\partial f_i(\mathbf{x})}{\partial x_j}\right]_{i,j=1}^{n} = \begin{bmatrix} \nabla f_1(\mathbf{x})^T \\ \vdots \\ \nabla f_n(\mathbf{x})^T \end{bmatrix} \tag{15.30}$$

jelölést ($J(\mathbf{x})$ az $f(\mathbf{x})$ függvény Jacobi-mátrixa), akkor a linearizált egyenletrendszer átírható a tömörebb

$$f(\mathbf{x}_i) + J(\mathbf{x}_i)(\mathbf{x} - \mathbf{x}_i) = 0 \tag{15.31}$$

alakba, amelynek megoldása:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - [J(\mathbf{x}_i)]^{-1} f(\mathbf{x}_i) \quad (i = 1, \ldots). \tag{15.32}$$

Az $\{\mathbf{x}_i\}$ sorozat fenti előállítását ($\mathbf{x}_0$ előzetes felvételével) nevezzük Newton-módszernek. A gyakorlatban soha nem invertáljuk a $J(\mathbf{x}_i)$ Jacobi-mátrixot. Helyette a lineáris egyenletrendszer alakot oldjuk meg. Így az eljárás alakja az alábbi.

A NEWTON MÓDSZER EGYENLETRENDSZEREKRE:

Adott $\mathbf{x}_0 \in \mathbb{R}^n$, $\varepsilon > 0$.
for $i = 0, 1, 2, \ldots$
Oldjuk meg a $J(\mathbf{x}_i)\Delta_i = -f(\mathbf{x}_i)$ egyenletrendszert!
$\mathbf{x}_{i+1} = \mathbf{x}_i + \Delta_i$
end

Az alkalmazható kilépési feltételek értelemszerűen itt is az (15.28) feltételek és ezek kombinációi. Az eljárás konvergenciája alkalmas feltételek esetén lokális és másodrendű.

Példa. Tekintsük az $f_1(\mathbf{x}) = x_1^2 + x_2^2 - 1 = 0$, $f_2(\mathbf{x}) = -x_1^2 - x_2 = 0$ kétismeretlenes nemlineáris egyenletrendszert. Az $\mathbf{x}_1 = [1, 1]^T$ pontban a $z = f_1(\mathbf{x})$ függvény S_1 érintősíkja $z = 2x_1 + 2x_2 - 3$, a $z = f_2(\mathbf{x})$ függvény S_2 érintősíkja pedig $z = -2x_1 - x_2 + 1$. A két érintősík egy egyenesben metszi egymást, amely a $z = 0$ síkot az $\mathbf{x}_2 = [-1/2, 2]^T$ pontban metszi. A következő ábra ezen objektumokat mutatja be.

Példa. Oldjuk meg az $f = [(f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})]^T = 0, \quad \mathbf{x} = [x_1, \ldots, x_n]^T$ egyenletrendszert Newton-módszerrel, ha

$$\begin{aligned} f_1(\mathbf{x}) &= x_1 \\ f_i(\mathbf{x}) &= \cos(x_{i-1}) + x_i - 1 \quad (i = 2, \ldots, n)! \end{aligned}$$

Legyen a kezdővektor $\mathbf{x}^{(0)} = [-1, 1, \ldots, -1, 1]^T$. A Jacobi-mátrix:

$$J(\mathbf{x}) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\sin(x_1) & 1 & 0 & & \vdots \\ 0 & -\sin(x_2) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\sin(x_{n-1}) & 1 \end{bmatrix}$$

Az $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - [J(\mathbf{x}^{(k)})]^{-1} f(\mathbf{x}^{(k)})$ sorozat elemei MATLAB-ban írt programmal számolva $n = 4$ esetén:

$$\begin{aligned} \mathbf{x}^{(1)} &= \begin{bmatrix} 0 \\ -0.3818 \\ -0.7030 \\ 0.2098 \end{bmatrix}, \quad \mathbf{x}^{(2)} = \begin{bmatrix} 0 \\ 0 \\ -0.7020 \\ -0.1720 \end{bmatrix}, \\ \mathbf{x}^{(3)} &= \begin{bmatrix} 0 \\ 0 \\ 0 \\ -0.0025 \end{bmatrix}, \quad \mathbf{x}^{(4)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -0.5e-16 \end{bmatrix}. \end{aligned}$$

Megmutatható, hogy az $\mathbf{x}^{(k)}$ sorozat bármely $n \geq 1$ esetén és bármilyen $\mathbf{x}^{(0)}$-ból véges (legfeljebb n) lépésben (a kerekítési hibáktól eltekintve) eléri az egyetlen, $\mathbf{x}^* = [0, \ldots, 0]^T$ megoldást.

15.3. Utólagos hibabecslések

Nemlineáris egyenletrendszerek közelítő megoldásainak utólagos (inverz) hibabecslését elvégezhetjük az Ottli-Prager becslés alábbi általánosításával is.

17.8 Tétel *(Arioli-Duff-Ruiz). Legyen $\hat{E} \geq 0$ és $\hat{f} \geq 0$ ($\hat{E} \in \mathbb{R}^{n \times n}$, $\hat{f} \in \mathbb{R}^n$). Legyen $F : \mathbb{R}^n \to \mathbb{R}^n$ nemlineáris függvény, $b \in \mathbb{R}^n$ és $\tilde{x}$ az*

$$F(x) = b \tag{15.33}$$

egyenletrendszer közelítő megoldása. Akkor és csak akkor van olyan G ($|G| \leq \hat{E}$) és g ($|g| \leq \hat{f}$), amelyre $\tilde{x}$ az

$$F(\tilde{x}) + G\tilde{x} = b + g \tag{15.34}$$

perturbált egyenletrendszer megoldása, ha

$$|r| = |b - F(\tilde{x})| \leq \hat{E}\,|\tilde{x}| + \hat{f}. \tag{15.35}$$

Ha adott E és f mellett $\hat{E} = \omega E$ és $\hat{f} = \omega f$, akkor

$$\omega = \max_i \frac{|r_i|}{(E\,|\tilde{x}| + f)_i} \tag{15.36}$$

a legkisebb ω, amellyel a tétel kielégíthető.

Példa. Tekintsük az előző szakasz példáját. Ha az $E = J(\widetilde{x})$ és $f = [1, \ldots, 1]^T \in \mathbb{R}^n$ választással élünk, akkor $\widetilde{x} = \mathbf{x}^{(1)}$ esetén $\omega = 0.4200$, $\widetilde{x} = \mathbf{x}^{(2)}$ esetén $\omega = 0.1482$, $\widetilde{x} = \mathbf{x}^{(3)}$ esetén $\omega = 0.0025$, $\widetilde{x} = \mathbf{x}^{(4)}$ esetén pedig $\omega = 0$. Ez utóbbit az magyarázza, hogy a MATLAB rendszerben $f\left(\mathbf{x}^{(4)}\right) = 0$.

15.4. Feladatok

1. Oldjuk meg az

$$\int_0^T \sin(s + \sqrt{T^2 + s})ds = 0.25$$

egyenletet T-re az intervallumfelezés módszerével. Használjunk numerikus kvadratúrát a függvény kiértékeléséhez!
2. Alkalmazzuk a Newton-módszert a

$$h\left(x\right) = \sqrt[3]{x}e^{-x^2} = 0, \quad f(x) = \frac{1}{1+x^2} = 0, \quad f(x) = \sqrt{\frac{|1-e^{-x}|}{\log(1/\left|x+4\right|)}} = 0$$

egyenletekre az (A), illetve (B) kilépési feltételekkel! Mit tapasztalunk?
3. Oldjuk meg a Newton-módszerrel a következő egyenletrendszereket!

$$\begin{gathered} 3x + 4y + e^{z+w} = 1.007 \\ 6x - 4y + e^{3z+w} = 11 \\ x^4 - 4y^2 + 6z - 8w = 20 \\ x^2 + 2y^3 + z - w = 4 \end{gathered} \tag{α}$$

$$\begin{gathered} -2x^2 - 3xy + 4\sin y = -6 \\ 3x^2 - 2xy^2 + 3\cos x = 8 \end{gathered} \tag{β}$$

$$\begin{gathered} x_i + \sum_{j=1}^4 x_j - 5 = 0, \ i = 1, 2, 3 \\ x_1 x_2 x_3 x_4 - 1 = 0 \end{gathered} \tag{γ}$$

16. fejezet

DIFFERENCIÁLEGYENLETEK KÖZELÍTŐ MEGOLDÁSA

Közönséges differenciálegyenletek kezdeti- és peremérték feladataival foglalkozunk.

16.1. A kezdetiérték feladat megoldása Runge-Kutta típusú módszerekkel

Az

$$y' = f(x,y), \quad y(x_0) = y_0 \quad \left(f : \mathbb{R} \times \mathbb{R}^\ell \to \mathbb{R}^\ell\right) \tag{16.1}$$

alakú kezdetiérték feladatokat vizsgáljuk, ahol $f(x,y) = [f_1(x,y), \ldots, f_\ell(x,y)]^T$ ($x \in \mathbb{R}$, $y \in \mathbb{R}^\ell$) folytonos a

$$D = \{(x,y) : \|x - x_0\|_\infty < \kappa_x,\ \|y - y_0\|_\infty < \kappa_y\} \subseteq \mathbb{R}^{\ell+1} \tag{16.2}$$

nyílt tartományon, κ_x és κ_y pozitív konstansok és létezik olyan $L > 0$ konstans, hogy

$$\|f(x,y) - f(x,z)\|_\infty \leq L \|y - z\|_\infty \quad ((x,y),(x,z) \in D). \tag{16.3}$$

Ekkor minden $(x_0, y_0) \in D$ esetén a kezdetiérték feladatnak létezik pontosan egy $y(x) = [y_1(x), \ldots, y_\ell(x)]^T$ megoldása valamely $[x_0, B]$ intervallumon, azaz

$$y'(x) = f(x, y(x)), \quad x \in [x_0, B]. \tag{16.4}$$

A feladat numerikus megoldásán a következőt értjük. A megoldást egy $[x_0, b]$ intervallum ($b \leq B$) diszkrét pontjaiban keressük. Ezek a pontok legyenek

$$x_0 = t_0 < t_1 < \ldots < t_j < \ldots < t_N = b. \tag{16.5}$$

A $\{t_i\}_{i=0}^N$ alappont halmazt az $[x_0, b]$ intervallum felosztásának nevezzük. A felosztás ekvidisztans (egyentávolságú), ha

$$t_i = t_0 + ih \quad (i = 0, 1, \ldots, N), \quad h = \frac{b - t_0}{N}. \tag{16.6}$$

Az $y(x)$ elméleti megoldás t_i pontbeli közelítését jelölje y_i. Értelemszerűen $y(t_0) = y_0$. A $h_i = t_{i+1} - t_i > 0$ mennyiséget i-edik lépéshossznak nevezzük.

16.1.1. Az explicit Euler-módszer

A Cauchy-feladat (egyik) jellegzetessége az, hogy ha egy x pontban ismert az $y(x)$ megoldás vektor, akkor ismert az $y'(x) = f(x, y(x))$ derivált vektor is. A vektor komponensekre fennálló $y_i(x + h) \approx y_i(x) + hy_i'(x) = y_i(x) + h_i f(x, y(x))$ $(i = 1, \ldots, \ell)$ elsőrendű közelítéseket (tkp. érintőket) vektor formában felírva kapjuk az

$$y(x+h) = \begin{bmatrix} y_1(x+h) \\ \vdots \\ y_t(x+h) \\ \vdots \\ y_\ell(x+h) \end{bmatrix} \approx y(x) + hy'(x) = y(x) + hf(x, y(x))$$

közelítést. Ha az x pontban az $y(x) \approx \hat{y}$ közelítő érték ismert, akkor a fenti képlet átmegy az

$$y(x+h) \approx \hat{y} + hf(x, \hat{y})$$

közelítésbe. Az Euler-módszer alapgondolata ezek után a következő: A $t_1 = t_0 + h_0$ pontban közelítsük az $y(t_1)$ elméleti megoldást a megoldásgörbe (x_0, y_0) pontbeli "érintőjével", azaz legyen

$$y(t_0 + h_0) \approx y_0 + h_0 y'(t_0) = y_1 = y_0 + h_0 f(t_0, y_0). \tag{16.7}$$

A t_1 pontbeli y_1 közelítést felhasználva kapjuk, hogy

$$y(t_2) \approx y(t_1) + h_1 f(t_1, y(t_1)) \approx y_2 = y_1 + h_1 f(t_1, y_1). \tag{16.8}$$

Az eljárást folytatva kapjuk hogy

$$y_{i+1} = y_i + h_i f(x_i, y_i) \quad (i = 0, 1, \ldots, N-1), \tag{16.9}$$

ahol $y_i \approx y(t_i)$. Ezt a képletet nevezzük explicit Euler-módszernek. Grafikusan ábrázolva az eljárást skalár differenciálegyenlet ($f : \mathbb{R}^2 \to \mathbb{R}$) esetén a következő

ábrát nyerhetjük: A rombusszal jelölt és egyenes szakaszokkal összekötött pontok az y_i $(i = 0, 1, 2)$ közelítő megoldások, amelyeken átmennek az $y'(x) = f(x, y)$, $y(t_i) = y_i$ differenciálegyenletek megoldásai $(t_i = 1 + 0.5(i-1)$, $i = 0, 1, 2)$.

A következőkben az eljárás hibáját elemezzük.

18.1 Definíció. *Az* $T(y(x), h) = y(x+h) - (y(x) + hf(x, y(x)))$ *mennyiséget az* x *pontbeli lokális hibának nevezzük.*

18.2 Definíció. *Az* $e_i = y_i - y(t_i)$ *hibát az* i*-edik pontbeli globális hibának nevezzük* $(i = 0, 1, \ldots, N)$.

A definíciók és a (16.3) feltételek alapján fennáll, hogy

$$\begin{aligned} \|y_{i+1} - y(t_{i+1})\|_\infty &= \|y_i + h_i f(t_i, y_i) - [y(t_i) + h_i f(t_i, y(t_i)) + T(y(t_i), h_i)]\|_\infty \\ &\leq \|y_i - y(t_i) + h_i [f(t_i, y_i) - f(t_i, y(t_i))]\|_\infty + \|T(y(t_i), h_i)\|_\infty \\ &\leq (1 + h_i L) \|y_i - y(t_i)\|_\infty + \|T(y(t_i), h_i)\|_\infty , \end{aligned}$$

azaz

$$\|e_{i+1}\|_\infty \leq (1 + h_i L) \|e_i\|_\infty + \|T(y(t_i), h_i)\|_\infty \quad (i = 0, 1, \ldots, N-1). \tag{16.10}$$

A globális hiba vizsgálatához két további eredményre van szükségünk. Igaz a következő

18.1 Lemma. *A* $\delta_{i+1} \leq \alpha_i \delta_i + \gamma_i$, $(\alpha_i \geq 0$, $i = 0, 1, \ldots, n)$ *egyenlőtlenség megoldása zárt alakban*

$$\delta_{n+1} \leq \left(\prod_{j=0}^{n} \alpha_j \right) \delta_0 + \sum_{i=0}^{n} \left(\prod_{j=i+1}^{n} \alpha_j \right) \gamma_i, \quad n \geq 0. \tag{16.11}$$

Bizonyítás. Az $n = 0$ esetben a rekurzió alapján,

$$\delta_1 \leq \alpha_0 \delta_0 + \gamma_0,$$

a képlet alapján pedig ($\prod_{j=l}^{i} f(j) = 1$, ha $i < l$)

$$\delta_1 \leq \left(\prod_{j=0}^{0} \alpha_j \right) \delta_0 + \sum_{i=0}^{0} \left(\prod_{j=1}^{0} \alpha_j \right) \gamma_i = \alpha_0 \delta_0 + \gamma_0.$$

Feltéve, hogy az állítás valamely $n \geq 0$ értékre igaz, igazoljuk helyességét az $n+1$ értékre is. Ekkor kapjuk, hogy

$$\begin{gathered} \delta_{n+2} \leq \alpha_{n+1} \delta_{n+1} + \gamma_{n+1} \leq \\ \leq \alpha_{n+1} \left(\textstyle\prod_{j=0}^{n} \alpha_j \right) \delta_0 + \alpha_{n+1} \sum_{i=0}^{n} \left(\textstyle\prod_{j=i+1}^{n} \alpha_j \right) \gamma_i + \gamma_{n+1} \leq \\ \leq \left(\textstyle\prod_{j=0}^{n+1} \alpha_j \right) \delta_0 + \sum_{i=0}^{n+1} \left(\textstyle\prod_{j=i+1}^{n+1} \alpha_j \right) \gamma_i, \end{gathered}$$

ami bizonyítandó volt. □

Ha a lemmát az $\alpha_j = 1 + h_j L$, $\gamma_j = \|T(y(t_j), h_j)\|_\infty$ szereposztással alkalmazzuk, akkor az

$$\|e_{n+1}\|_\infty \leq \left[\prod_{j=0}^{n}(1+h_j L)\right]\|e_0\|_\infty + \sum_{i=0}^{n}\left[\prod_{j=i+1}^{n}(1+h_j L)\right]\|T(y(t_i), h_i)\|_\infty \tag{16.12}$$

egyenlőtlenséget kapjuk. Vizsgáljuk most meg az $T(y(x), h)$ lokális hiba nagyságát. Tegyük fel, hogy $y(x) \in C^2[x_0, b]$. Ekkor $y(x+h) = y(x) + hy'(x) + R_1(x, h)$, amelynek maradéktagjára teljesül, hogy

$$\|R_1(x,h)\|_\infty \leq h^2 K_2 \quad \left(2K_2 = \max_{1\leq i\leq \ell}\max_{x\in[x_0,b]}|y_i''(x)|\right). \tag{16.13}$$

A képlethiba ennek megfelelően

$$T(y(x), h) = (y(x) + hy'(x) + R_1(x,h)) - (y(x) + h\overbrace{f(x, y(x))}^{y'(x)}) = R_1(x,h),$$

ahonnan $\|T(y(x), h)\|_\infty \leq K_2 h^2$ $(x, x+h \in [x_0, b])$ és

$$|T(y(t_i), h_i)| \leq K_2 h_i^2 \quad (i = 0, 1, \ldots, N-1) \tag{16.14}$$

következik. Az $1 + x \leq e^x$ $(x \geq 0)$ egyenlőtlenség figyelembevételével kapjuk, hogy $1 + h_j L \leq e^{Lh_j}$ és

$$\prod_{j=i+1}^{n}(1+h_j L) \leq \prod_{j=i+1}^{n} e^{Lh_j} = e^{L\sum_{j=i+1}^{n} h_i} \leq e^{L(b-x_0)}. \tag{16.15}$$

Ha ezt a (16.12) képletbe behelyettesítjük, akkor kapjuk, hogy

$$\|e_{n+1}\|_\infty \leq e^{L(b-x_0)}\|e_0\|_\infty + \sum_{i=0}^{n} e^{L(b-x_0)} K_2 h_i^2. \tag{16.16}$$

A nyilvánvaló

$$\sum_{i=0}^{n} h_i^2 \leq \left(\max_{0\leq i\leq n} h_i\right)\sum_{i=0}^{n} h_i \leq (b-x_0)\max_{0\leq i\leq n} h_i,$$

egyenlőtlenség miatt

$$\|e_{n+1}\|_\infty \leq e^{L(b-x_0)}\left(\|e_0\|_\infty + (b-x_0)K_2 \max_{0\leq i\leq n} h_i\right) \quad (n = 0, 1, \ldots, N-1). \tag{16.17}$$

Mivel $e_0 = y_0 - y(x_0) = 0$, az Euler-módszer globális hibájára igaz a következő

18.1 Tétel. *Az Euler-módszer globális hibájára $y(x) \in C^2[x_0, b]$ esetén fennáll, hogy*

$$\max_{1\le i\le N} \|e_i\|_\infty \le (b - x_0) K_2 e^{L(b-x_0)} \max_{0\le i\le N-1} h_i. \tag{16.18}$$

Ez azt jelenti, hogy az Euler-módszer hibája a legnagyobb lépéshosszal arányos. Ha a $\{t_i\}_{i=0}^N$ felosztás minden határon túl finomodik, azaz $N \to \infty$ és $\max_{0\le i\le N} h_i \to 0$ egyidejűleg teljesül, akkor fennáll, hogy

$$\max_{1\le i\le N} \|e_i\|_\infty \to 0, \quad N \to \infty. \tag{16.19}$$

Tehát az Euler-módszer (elsőrendben) konvergens. A konvergencia jellegét mutatja a következő ábra, amelyen az $y' = 2y/x + 2x^3$, $y(1) = 2$ Cauchy-probléma elméleti és Euler-megoldásai láthatók $h = 1, 1/2, 1/4, 1/8$ esetén.

16.1.2. Explicit egylépéses módszerek

Az Euler-módszernek számtalan hatékonyabb továbbfejlesztése ismeretes. Ezek közül az egyik legfontosabb az explicit egylépéses módszerek osztálya, amelynek alakja

$$y_{i+1} = y_i + h_i \phi(t_i, y_i, h_i) \quad (i = 0, 1, \ldots, N-1). \tag{16.20}$$

Az $f(x, y)$ függvénytől függő $\phi(x, y, h)$ ($\phi : \mathbb{R} \times \mathbb{R}^\ell \times \mathbb{R} \to \mathbb{R}^\ell$) növekményfüggvény minden változójában folytonos, az y változóban kielégíti a

$$\|\phi(x, y, h) - \phi(x, z, h)\|_\infty \le K \|y - z\|_\infty, \left(K \ge 0, (x, y), (x, z) \in D, |h| \le \widehat{h}\right) \tag{16.21}$$

feltételt és

$$\phi(x, y, 0) = f(x, y). \tag{16.22}$$

Az Euler-módszer a $\phi(x, y, h) = f(x, y)$ speciális esetnek felel meg. Az egylépéses módszerek x pontbeli lokális hibáját az

$$T(y(x), h) = y(x + h) - (y(x) + h\phi(x, y(x), h)) \tag{16.23}$$

mennyiség definiálja. Az egylépéses módszer p-edrendű, ha létezik $K_p > 0$ konstans, hogy

$$\|T(y(x), h)\|_\infty \le K_p h^{p+1}, \quad x, x + h \in [x_0, b]. \tag{16.24}$$

Az Euler-módszerhez hasonlóan belátható, hogy az egylépéses módszerek globális hibájára is fennáll az alábbi egyenlőtlenség

$$\|e_{n+1}\|_\infty \le (1 + h_n K) \|e_n\|_\infty + \|T(y(t_n), h_n)\|_\infty \quad (n = 0, 1, \ldots, N-1). \tag{16.25}$$

Ha az egylépéses módszer p-edrendű, akkor az Euler-módszerhez hasonlóan igazolhatjuk, hogy a globális hibára fennáll a

$$\max_{1\leq i\leq N}\|e_i\|_\infty \leq c\left(\max_{0\leq i\leq N-1} h_i\right)^p \tag{16.26}$$

egyenlőtlenség, ahol $c>0$ konstans. Tehát a p-edrendű egylépéses módszer p-edrendben konvergál.

Fontos megjegyezni, hogy egy p-edrendben konvergáló módszer általában csak olyan kezdetiérték feladatok esetében konvergál p-edrendben, amelyek megoldása folytonosan differenciálható legalább $(p+1)$-szer. Ha a differenciálegyenlet elméleti megoldása ennél kevesebbszer differenciálható, akkor a konvergencia rendje is csökken.

A p-edrendű egylépéses módszer ui. az $y(x)$ megoldás függvényt $p+1$-ed rendű hibával közelíti. Ez azt jelenti, hogy az $y(x+h)$ és $y(x)+h\phi(x,y(x),h)$ függvényeket az x pont körül sorbafejtve, a két sorfejtés első $p+1$ tagja megegyezik. A legkézenfekvőbb ilyen tulajdonságú formula az ún. *Taylor-sor módszer*, ahol

$$\phi(x,y(x),h)=\sum_{i=1}^{p} y^{(i)}(x)\,\frac{h^{i-1}}{i!}. \tag{16.27}$$

Az $y^{(i)}(x)$ magasabbrendű deriváltakat az összetett függvény deriválási szabályával és az $y'(x)=f(x,y(x))$ összefüggés segítségével tudjuk meghatározni. Pl.

$$\begin{aligned} y''(x)=(f(x,y(x)))'=f'_x(x,y(x))+f'_y(x,y(x))\,y'(x)= \\ =f'_x(x,y(x))+f'_y(x,y(x))\,f(x,y(x)). \end{aligned} \tag{16.28}$$

Természetesen bonyolult, vagy nehezen számítható deriváltak és többváltozó esetén a Taylor-sor módszer nem előnyös. Előnyösek viszont a *Runge-Kutta módszerek*.

Az explicit Runge-Kutta módszerek elkerülik a Taylor-sor együtthatóinak meghatározását és csak az y és $f(x,y)$ információkat használják. Szokásos alakjuk

$$\begin{aligned} y_{n+1}&=y_n+h_n\textstyle\sum_{i=1}^{m} c_i k_i, \\ k_1&=f(x_n,y_n), \\ k_i&=f\left(x_n+a_i h_n, y_n+h_n\textstyle\sum_{j=1}^{i-1} b_{ij}k_j\right)\quad (i=2,\ldots,m). \end{aligned} \tag{16.29}$$

Igazolható, hogy $\sum_{i=1}^{m} c_i=1$ esetén a $\phi(x,y,0)=f(x,y)$ feltétel teljesül. Általában feltesszük még, hogy

$$a_i=\sum_{j=1}^{i-1} b_{ij}\quad (i=2,\ldots,m). \tag{16.30}$$

Az explicit Runge-Kutta módszereket az alábbi mátrix sémában is meg lehet adni:

$$\begin{array}{c|cccc} 0 & & & & \\ a_2 & b_{21} & & & \\ \vdots & \vdots & \ddots & & \\ a_m & b_{m1} & \dots & b_{m,m-1} & \\ \hline & c_1 & \dots & c_{m-1} & c_m \end{array} \tag{16.31}$$

Legnevezetesebb az alábbi negyedrendű Runge-Kutta módszer:

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

Az Euler-módszer a

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array}$$

sémának felel meg.

A numerikus módszerekkel kapott közelítő megoldásokra általában a

$$\max_{1\le i\le N} \|e_i\|_\infty \le \epsilon \tag{16.32}$$

globális hiba korlátot írjuk elő. Könnyen igazolható, hogy

$$\|T\left(y\left(t_n\right), h_n\right)\|_\infty \le h_n\epsilon \quad (n=0,\dots,N-1) \tag{16.33}$$

esetén alkalmas $\widehat{c}>0$ konstanssal fennáll a

$$\max_{1\le n\le N-1} \|e_i\|_\infty \le \widehat{c}\epsilon \tag{16.34}$$

egyenlőtlenség. Ez a $\widehat{c}$ szám többnyire nem ismert. A (16.32) feltételt úgy próbáljuk meg teljesíteni, hogy a lokális hibákra egy

$$\|T\left(y\left(t_n\right), h_n\right)\|_\infty \le h_n\epsilon/q \quad (i=0,\dots,N-1) \tag{16.35}$$

alakú feltételt írunk elő, ahol $q>0$ egy tapasztalati konstans. Ha a (16.35) feltétel teljesül, akkor a számított közelítő megoldásokat $\epsilon>0$ hibájúnak fogadjuk el.

A lokális hibára vonatkozó (16.35) feltétel felveti a képlethiba becslésének kérdését. Számos módszer esetén analitikus és numerikus becslések is adhatók. Itt a két legjobban bevált, ill. leggyakrabban használt módszert ismertetjük.

A lépésfelezéses hibabecslés (Runge-féle szabály). A t_n pontból kiindulva végezzünk el két lépést a $h_n/2$ lépéshosszal is. Igy a $t_n + h_n$ pontban kapunk egy második $\hat{z}_{n+1}$ közelítést is. Igazolható, hogy

$$T\left(y\left(t_n\right), h_n\right) \approx \frac{\hat{z}_{n+1} - y_{n+1}}{2^p - 1}, \tag{16.36}$$

ahol p a módszer rendje.

Párosított Runge-Kutta formulák. Az ilyen módszereknél egy p-ed és egy $(p+1)$-ed rendű Runge-Kutta formulát úgy választanak meg, hogy az alacsonyabb rendű formulához tartozó k_i értékek egyúttal a magasabb rendű formulában is szerepelnek és a két formulával kapott közelítő megoldások különbsége becsli az alacsonyabbrendű módszer lokális hibáját. Sematikusan ábrázolva

$$\begin{array}{c|cccc} 0 & & & & \\ a_2 & b_{21} & & & \\ \vdots & \vdots & \ddots & & \\ a_m & b_{m1} & \dots & b_{m,m-1} & \\ \hline & c_1 & \dots & c_{m-1} & c_m \\ & d_1 & \dots & d_{m-1} & d_m \end{array} \tag{16.37}$$

Az alacsonyabbrendű formula:

$$y_{n+1} = y_n + h_n \left(c_1 k_1 + \ldots + c_m k_m\right). \tag{16.38}$$

A magasabbrendű formula:

$$\hat{y}_{n+1} = y_n + h_n \left(d_1 k_1 + \ldots + d_m k_m\right). \tag{16.39}$$

A lokális hiba becslése:

$$T\left(y\left(t_n\right), h_n\right) \approx h_n \sum_{i=1}^{m} \left(d_i - c_i\right) k_i. \tag{16.40}$$

Az ilyen formulák előnye a lépésfelezéssel szemben az, hogy csak egy pontra támaszkodó információt használnak fel. Igazolható, hogy $m \leq 5$ esetén ezek a becslések nem lehetnek aszimptotikusan pontosak.

Az egyik legismertebb formula az ún. 2(3)-as Runge-Kutta-Fehlberg képlet (a MATLAB ode23.m eljárása):

$$\begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ 1/2 & 1/4 & 1/4 & \\ \hline & 1/2 & 1/2 & 0 \\ & 1/6 & 1/6 & 4/6 \end{array}$$

A gyakorlatban a magasabbrendű formula közelítésével folytatják az eljárást.

Aszimptotikusan pontos England alábbi 4(5)-ös formulája, ahol $m = 6$.

0						
1/2	1/2					
1/2	1/4	1/4				
1	0	−1	2			
2/3	7/27	10/27	0	1/27		
1/5	28/625	−125/625	546/625	54/625	−378/625	
	1/6	0	4/6	1/6	0	0
	14/336	0	0	35/336	162/336	125/336

Jelölje a továbbiakban EST a lokális hiba becsült értékének normáját. A szokásos adaptív Runge-Kutta sémát az alábbiakban írhatjuk le.

ADAPTÍV EGYLÉPÉSES MÓDSZEREK ALGORITMUSA:

Input t_0, y_0, b, tol; h_0 kiválasztása; $i = 0$.

while $t_i < b$

1. y_{i+1} kiszámítása és a lokális hiba (EST) becslése
2. **if** $EST \leq tol$

   $i = i + 1$

   else

   $h_i = h_{új}$

   goto 1

   end

end

Az új lépéshosszt ($h_{új}$) a $\|T(y(x_i), h_i)\|_\infty \approx c_3 h_i^{p+1} \approx EST$ becslésből és a $c_3 h_{új}^{p+1} \approx tol$ követelményből kaphatjuk meg. Eszerint $c_3 \approx EST/h_i^{p+1}$ és $h_{új}^{p+1} \approx tol/c_3 \approx (tol/EST)\, h_i^{p+1}$, ahonnan

$$h_{új} = h_i \left(\frac{tol}{EST} \right)^{1/(p+1)}. \tag{16.41}$$

Ha a becsült hiba lényegesen kisebb mint az előírt tol hibakorlát, akkor növelni lehet a lépéshosszt az előbbiek figyelembevételével. Bizonyos esetekben ez a stratégia optimális.

Példa. Oldjuk meg az alábbi "orbitális" differenciálegyenlet rendszert a $[0, 20]$ intervallumon a MATLAB adaptív Ode23 programjával:

$$\begin{aligned}
y_1' &= y_3, & y_1(0) &= 1 - \lambda, \\
y_2' &= y_4, & y_2(0) &= 0, \\
y_3' &= \frac{-y_1}{\left(y_1^2 + y_2^2\right)^{3/2}}, & y_3(0) &= 0, \\
y_4' &= \frac{-y_2}{\left(y_1^2 + y_2^2\right)^{3/2}}, & y_4(0) &= \left[(1+\lambda)/(1-\lambda)\right]^{1/2},
\end{aligned}$$

ahol $\lambda = 0.3$. A program alapbeállításaival kapott megoldás komponensek (trajektóriák) képe a következő:

16.2. Peremérték feladatok megoldása differencia módszerekkel

Elsőként vizsgáljuk az

$$\begin{array}{c} y'' + p(x)\, y' + q(x)\, y = r(x) \quad (a < x < b) \\ y(a) = \alpha, \quad y(b) = \beta, \end{array} \tag{16.42}$$

lineáris másodrendű peremérték problémát, ahol $y, p, q, r : \mathbb{R} \to \mathbb{R}$ típusú függvények.

Legyen

$$x_j = a + jh \quad (j = 0, 1, \ldots, N+1), \quad h = \frac{b-a}{N+1}. \tag{16.43}$$

Legyen $u_j \approx y(x_j)$ és közelítsük az x_i pontbeli deriváltakat az

$$y''(x_i) \approx \frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}, \tag{16.44}$$

$$y'(x_i) \approx \frac{y(x_{i+1}) - y(x_{i-1})}{2h} \approx \frac{u_{i+1} - u_{i-1}}{2h} \tag{16.45}$$

differencia hányadosokkal. Ezeket behelyettesítjük a lineáris differenciálegyenletbe az $x = x_i$ pontban. Kapjuk, hogy

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + p(x_i)\frac{u_{i+1} - u_{i-1}}{2h} + q(x_i)\, u_i = r(x_i), \quad 1 \le i \le N,$$

ahol

$$u_0 = \alpha, \quad u_{N+1} = \beta.$$

Ennek átrendezett formája $1 \le i \le N$ esetén

$$-\left(1 - \tfrac{h}{2}p(x_i)\right) u_{i-1} + \left(2 - h^2 q(x_i)\right) u_i - \left(1 + \tfrac{h}{2}p(x_i)\right) u_{i+1} \;=\; -h^2 r(x_i).$$

Ha bevezetjük az $a_i = -\left(1 - \frac{h}{2}p(x_i)\right)$, $b_i = \left(2 - h^2 q(x_i)\right)$, $c_i = -\left(1 + \frac{h}{2}p(x_i)\right)$ jelöléseket, akkor a

$$\begin{array}{rcll} b_1 u_1 + c_1 u_2 & = & -h^2 r(x_1) - a_1\alpha & \\ a_i u_{i-1} + b_i u_i + c_i u_{i+1} & = & -h^2 r(x_i) & (2 \le i \le N-1)\ . \\ a_N u_{N-1} + b_N u_N & = & -h^2 r(x_N) - c_N\beta & \end{array} \tag{16.46}$$

tridiagonális egyenletrendszert kapjuk, amelyet alkalmas feltételek esetén a sávos Gauss-eliminációval stabilan $O(n)$ régi flop művelettel oldhatunk meg. Igaz a következő

18.2 Tétel. *Tegyük fel, hogy* $p, q \in C[a,b]$ *és*

$$|p(x)| \leq P, \quad 0 < Q_1 \leq -q(x) \leq Q_2, \quad a \leq x \leq b. \tag{16.47}$$

Legyen továbbá $h \leq 2/P$. *Ha* $y \in C^4[a,b]$, *akkor fennáll, hogy*

$$|u_j - y(x_j)| \leq M\frac{h^2}{12}(M_4 + 2PM_3), \quad 0 \leq j \leq N+1, \tag{16.48}$$

ahol $M = \max\{1, 1/Q_1\}$ *és* $M_i = \max_{a\leq x\leq b}\left|y^{(i)}(x)\right|$ $(i = 3, 4)$.

Tehát $y \in C^4[a,b]$ esetén a kapott $\{u_j\}_{j=1}^N$ közelítő értékekre fennáll

$$u_j = y(x_j) + O\left(h^2\right), \quad 1 \leq j \leq N.$$

Adott $\epsilon > 0$ pontosság eléréséhez szükséges h intervallum hosszát az extrapoláció elvére alapozva úgy szokás megállapítani, hogy megoldjuk a fenti lineáris egyenletrendszert h és $h/2$ intervallumhosszakkal. Ha a két felosztás közös pontjaiban a közelítő megoldások különbsége kisebb mint $c\epsilon$, ahol c egy tapasztalati szám (pl. $c = 1/100$), akkor h értékét elfogadjuk. Ha nem, akkor a $h/2$ felosztást tovább felezzük, és így tovább. Ha a h értéket elfogadjuk, akkor általában a $h/2$ felosztáshoz tartozó megoldásokat használjuk fel a továbbiakban.

Ha az

$$\begin{gathered} y'' = f(x, y, y') \quad (a < x < b) \\ y(a) = \alpha, \quad y(b) = \beta \end{gathered} \tag{16.49}$$

nemlineáris másodrendű peremérték problémára alkalmazzuk a (16.44)-(16.45) differenciahányadosokat, akkor az

$$\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} - f\left(x_j, u_j, \frac{u_{j+1} - u_{j-1}}{2h}\right) = 0 \quad (j = 1, \ldots, N) \tag{16.50}$$

nemlineáris egyenletrendszert kapjuk, ahol $u_0 = \alpha$ és $u_{N+1} = \beta$.

18.3 Tétel. *Tegyük fel, hogy* $f(x, y, z)$ *folytonosan differenciálható és*

$$\left|\frac{\partial f}{\partial z}\right| \leq P, \quad 0 < Q_1 \leq \left|\frac{\partial f}{\partial y}\right| \leq Q_2. \tag{16.51}$$

Legyen továbbá $h \leq 2/P$. *Ha* $y \in C^4[a,b]$, *akkor fennáll, hogy*

$$|u_j - y(x_j)| \leq M\frac{h^2}{12}(M_4 + 2PM_3) \quad (0 \leq j \leq N+1), \tag{16.52}$$

ahol $M = \max\{1, 1/Q_1\}$ *és* $M_i = \max_{a\le x\le b}\left|y^{(i)}(x)\right|$ $(i = 3, 4)$.

A nemlineáris egyenletrendszer megoldását általában a Newton-módszerrel végezzük. Ekkor az egyenletrendszer Jacobi-mátrixa tridiagonális. Adott $\epsilon > 0$ pontossághoz szükséges h intervallum hosszát a nemlineáris esetben is a felezési eljárással szokás elérni.

Példa. Oldjuk meg differencia módszerrel a $3yy'' + (y')^2 = 0; \quad y(a) = \alpha; y(b) = \beta$ peremérték feladatot! A pontos megoldás:

$$y(x) = \left[(\beta\sqrt[3]{\beta} - \alpha\sqrt[3]{\alpha})x + \alpha\beta(\sqrt[3]{\alpha} - \sqrt[3]{\beta})\right]^{3/4}/(b-a)^{3/4}.$$

Legyen $a = 2$, $y(2) = \alpha = 1$, $b = 6$, $y(6) = \beta = 5$. Ezekkel a konkrét peremértékekkel:

$$y(x) = \left[(5\sqrt[3]{5} - 1)x + 5(1 - \sqrt[3]{5})\right]^{3/4}/\sqrt{8}.$$

A differencia módszert alkalmazva, helyettesítsük be a numerikus deriváltakat és szorozzunk a nevezővel. A következő nemlineáris egyenletrendszert kapjuk:

$$12u_j(u_{j+1} - 2u_j + u_{j-1}) + (u_{j+1} - u_{j-1})^2 = 0 \quad (j = 1, \ldots, N).$$

Egyenletrendszerünket különböző N-ekre, illetve N-ből adódó h-ra megoldva az alábbi eredmények adódnak:

x_j		u_j		$\mathbf{y}(\mathbf{x}_j)$
	$h = 0.5$	$h = 0.25$	$h = 0.0625$	
2.5	1.6432	1.6454	1.6462	**1.6462**
3.0	2.2118	2.2142	2.2152	**2.2151**
3.5	2.7356	2.7377	2.7384	**2.7384**
4.0	3.2278	3.2295	3.2302	**3.2302**
4.5	3.6963	3.6976	3.6980	**3.6980**
5.0	4.1457	4.1466	4.1469	**4.1469**
5.5	4.5795	4.5799	4.5801	**4.5801**

A táblázatban csak a mindegyik felosztásban szereplő pontokat tüntettük fel. Megállapíthatjuk, hogy $h = 0.5$ mellett egy tizedesjegyre (sok helyen két tizedesjegyre is) pontosak a közelítések. Felezve a lépéstávolságot, további egy tizedesjeggyel pontosabb eredményeket kaptunk. A pontosság a h további csökkentésével hasonló javulást mutatott. Egy kivétellel minden pontban 4 tizedesjegy pontosságot kaptunk $h = 0.0625$ lépésköz esetén.

A lépésfelezéses hibabecsléssel pl. $c = 1/100$ választással $h = 0.25$ mellett ϵ-ra 0.24 adódik, ami jócskán felülbecsli a tényleges hibát. Végezzük el a hibaanalízist a 18.3 Tétel alapján is! Rövid számolás után kihozhatjuk a $P = 1$, $Q_1 = 0.078$, $Q_2 = 0.67$, $M = 12.8$, $M_3 = 1.6$, $M_4 = 6.7$ korlátokat. Ezekkel kapjuk, hogy $h \le 2$

esetén $(u_j - y(x_j) \leq 11h^2$, ami pedig még a $h = 0.0625$ mellett is csupán egy tizedesjegy pontosságot garantál, miközben a tényleges helyzet ettől lényegesen jobb. Arra az (óvatos) következtetésre juthatunk, hogy akkor is megkísérelhetjük egy feladatnál a differencia módszer alkalmazását, ha a tétel egyébként nem erősítené meg a megoldásra vonatkozó reményünket.

16.3. Feladatok

1. Oldjuk meg az Euler, a 4-ed rendű Runge-Kutta és a 2(3) RKF-módszerrel az $y' = y$, $y(0) = 1$ feladatot $\epsilon = 10^{-4}, 10^{-6}, 10^{-8}$ pontossággal a $[0, 1]$ intervallumon! Mekkora a becsült és a tényleges hiba? Hogyan alakul a globális hiba a lépéshosszak függvényében?
2. Oldjuk meg a $[0, 20]$ intervallumon a ún. Brüsszelátor problémát, amelynek alakja

$$y_1' = 1 + y_1^2 y_2 - (B+1)\, y_1, \quad y_1(0) = 1.3$$
$$y_2' = B y_1 - y_1^2 y_2, \quad y_2(0) = B$$

A paraméter értéke legyen $2.9 \leq B \leq 3.1$ között. Legyen továbbá $tol = 3 \times 10^{-4}$. Ábrázoljuk is a közelítő megoldást!
3. Oldjuk meg differencia módszerrel a lineáris

$$\epsilon y'' + (1+\epsilon)\, y' + y = 0, \quad y(0) = 0,\ y(1) = 1$$

peremérték feladatot $\epsilon = 10^{-1}, 10^{-2}$ esetén. A megkövetelt pontosság legyen $tol = 10^{-5}$. Hasonlítsuk össze a közelítő megoldást az

$$y(x) = \left(e^{-x} - e^{-x/\epsilon}\right) / \left(e^{-1} - e^{-1/\epsilon}\right)$$

elméleti megoldással!

17. fejezet

IRODALOM

[1] Anderson, E., Bai, Z., et.al: *LAPACK Users' Guide*, SIAM, Philadelphia, 1992
[2] Coleman, T.F., Van Loan, C.: *Handbook for Matrix Computations*, SIAM, Philadelphia, 1988
[3] Chaitin-Chatelin, F., Frayssé, V.: *Lectures on Finite Precision Computations*, SIAM, Philadelphia, 1996
[4] Demmel, J.W.: *Applied Numerical Linear Algebra*, SIAM Philadelphia, 1997
[5] Forsythe, G.E., Malcolm, M.A., Moler, C.B.: *Computer Methods for Mathematical Computations*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1977
[6] Forsythe, G.E., Moler, C.B.: *Lineáris algebrai problémák megoldása számítógéppel*, Műszaki Könyvkiadó, 1976
[7] Golub, G.H., Van Loan, C.F.: Matrix Computations, (second edition), The Johns Hopkins University Press, Baltimore, 1993
[8] Iványi A. (szerk.): Informatikai algoritmusok 1., ELTE Eötvös Kiadó, 2004
[9] Higham, N.J.: *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996
[10] Jennings, A., McKeown, J.J.: *Matrix Computation*, (second edition), John Wiley & Sons, 1992
[11] Móricz F.: *Numerikus módszerek az algebrában és analízisben*, Polygon, 1997
[12] Moler, C.B.: *Numerical Computing with MATLAB*, SIAM, Philadelphia, 2004
[13] Overton, M.L.: *Numerical Computing with IEEE Floating Point Arithmetic*, SIAM, Philadelphia, 2001
[14] Popper Gy., Csizmás F.: *Numerikus módszerek mérnököknek*, Akadémiai Kiadó-Typotex, 1993
[15] Ralston, A.: *Bevezetés a numerikus analízisbe*, Műszaki Könyvkiadó, 1969
[16] Rice, J.E.: *Numerical Methods, Software, and Analysis*, McGraw-Hill, 1983
[17] Rice, J.E.: *Matrix Computations and Mathematical Software*, McGraw-Hill, 1983
[18] Rivlin, T.J.: *An Introduction to the Approximation of Functions*, Dover,1981

[19] Rózsa P.: *Lineáris algebra és alkalmazásai*, Műszaki Könyvkiadó, 1974
[20] Stewart, G.W.: *Matrix Algorithms, Vol. I: Basic Decompositions*, SIAM, Philadelphia, 1998
[21] Stewart, G.W.: *Matrix Algorithms, Vol. II: Eigensystems*, SIAM, Philadelphia, 2001
[22] Stoyan G., Takó G.: *Numerikus módszerek 1-3*, ELTE-Typotex, 1993, 1995, 1997
[23] Stoyan G. (szerk.): *Matlab*, Typotex Kiadó, Budapest 2005
[24] Szamarszkij, A.A.: *Bevezetés a numerikus módszerek elméletébe*, Tankönyvkiadó, 1989
[25] Ueberhuber, C.W.: *Numerical Computation 1-2 (Methods, Software, and Analysis)*, Springer, 1997
[26] Watkins, D.S.: *Fundamentals of Matrix Computations*, John Wiley & Sons, 1991
[27] Wilkinson, J.H.: *Rounding Errors in Algebraic Processes*, Dover, 1994