Numerikus módszerek példatár

Faragó István, Fekete Imre, Horváth Róbert

2013. július 5.

Tartalomjegyzék

Előszó **2**

Feladatok **4**

1. Előismeretek **4**
1.1. Képletek, összefüggések 4
1.2. Feladatok 6
1.2.1. Nevezetes mátrixtípusok 6
1.2.2. Normált és euklideszi terek 7
1.2.3. Banach-féle fixponttétel 8
1.2.4. Vektornormák 8
1.2.5. Mátrixnormák 9

2. Modellalkotás és hibaforrásai **12**
2.1. Képletek, összefüggések 12
2.1.1. Feladatok kondicionáltsága 12
2.1.2. A gépi számábrázolás 12
2.2. Feladatok 13
2.2.1. Feladatok kondicionáltsága 13
2.2.2. A gépi számábrázolás 14

3. Lineáris egyenletrendszerek megoldása **17**
3.1. Képletek, összefüggések 17
3.1.1. Kondicionáltság 17
3.1.2. Direkt módszerek 18
3.1.3. Iterációs módszerek 19
3.1.4. Túlhatározott lineáris egyenletrendszerek megoldása 22
3.2. Feladatok 22
3.2.1. Kondicionáltság 22
3.2.2. Direkt módszerek 24

3.2.3. Iterációs módszerek . . . 28
3.2.4. Túlhatározott lineáris egyenletrendszerek megoldása . . . 31

4. Sajátérték-feladatok numerikus megoldása 33
4.1. Képletek, összefüggések . . . 33
4.2. Feladatok . . . 36
4.2.1. Sajátértékbecslések . . . 36
4.2.2. Hatványmódszer és változatai . . . 37
4.2.3. Jacobi- és QR-iterációk . . . 39

5. Nemlineáris egyenletek és egyenletrendszerek megoldása 42
5.1. Képletek, összefüggések . . . 42
5.2. Feladatok . . . 47
5.2.1. Sorozatok konvergenciarendje, hibabecslése . . . 47
5.2.2. Zérushelyek lokalizációja . . . 48
5.2.3. Intervallumfelezési módszer . . . 48
5.2.4. Newton-módszer . . . 48
5.2.5. Húr- és szelőmódszer . . . 50
5.2.6. Fixpont iterációk . . . 50
5.2.7. Nemlineáris egyenletrendszerek megoldása . . . 51

6. Interpoláció és approximáció 53
6.1. Képletek, összefüggések . . . 53
6.1.1. Polinominterpoláció . . . 53
6.1.2. Trigonometrikus interpoláció . . . 56
6.1.3. Approximáció polinomokkal . . . 57
6.2. Feladatok . . . 58
6.2.1. Polinominterpoláció . . . 58
6.2.2. Trigonometrikus interpoláció . . . 62
6.2.3. Approximáció polinomokkal és trigonometrikus polinomokkal . . . 62

7. Numerikus deriválás és numerikus integrálás 63
7.1. Képletek, összefüggések . . . 63
7.2. Feladatok . . . 64
7.2.1. Numerikus deriválás . . . 64
7.2.2. Numerikus integrálás . . . 66

8. A kezdetiérték-feladatok numerikus módszerei 69
8.1. Képletek, összefüggések . . . 69
8.2. Feladatok . . . 70
8.2.1. Egylépéses módszerek . . . 70

8.2.2. Többlépéses módszerek 75

9. A peremérték-feladatok numerikus módszerei **77**
9.1. Képletek, összefüggések 77
9.2. Feladatok 78
9.2.1. Peremérték-feladatok megoldhatósága 78
9.2.2. Véges differenciák módszere és a belövéses módszer 80

10. Parciális differenciálegyenletek **83**
10.1. Képletek, összefüggések 83
10.2. Feladatok 84
10.2.1. Elméleti feladatok 84
10.2.2. Elliptikus és parabolikus feladatok megoldása véges differenciákkal 85

Útmutatások, végeredmények **88**

Megoldások **117**

Előszó

Ez a példatár a 2011-ben megjelent Numerikus módszerek című elektronikus jegyzetünkhöz készült, így azzal együtt képez egységes oktatási segédanyagot, melyhez jelöléseiben és a fejezetek tagolásában is igazodik.

Minden fejezet elején röviden felsoroljuk a témakör legfontosabb tételeit, és ezekre hivatkozunk is a megoldások során. Több feladat esetén nemcsak a megoldást közöljük, hanem külön helyen megadjuk a feladatok végeredményeit ill. a megoldási útmutatókat, ezzel is segítve a feladatok önálló feldolgozását. A megoldáshoz a $\Longrightarrow$ jelre kattintva lehet eljutni, míg a $\longrightarrow$ jel a végeredményekhez ill. útmutatókhoz visz minket. A megoldások előtti sorszámra kattintva visszajuthatunk a feladathoz.

A példatárban nemcsak elméleti feladatokat közlünk, hanem számítógéppel megoldandó gyakorlati feladatokat is. Ezek segítenek a módszerek alkalmazásának bemutatásában, és elősegítik a módszerek mélyebb megértését. Egyes feladatok számítógépes programok írását követelik meg. Ezt a tényt a feladatok szövege előtti $\boxminus$ szimbólum jelöli. Ha egy feladat megoldásához számítógép szükséges, akkor erre a feladat szövege előtti $\boxplus$ szimbólum hívja fel a figyelmet.

A jegyzet készítése során kihasználtuk azt is, hogy az elektronikus formában fog megjelenni, így több helyen külső linkekkel segítjük a megértést és a szélesebb körű tájékozódást a témakörrel kapcsolatban.

Budapest, 2013. június

A szerzők

Feladatok

1. fejezet

Előismeretek

1.1. Képletek, összefüggések

A vektorokkal és a mátrixokkal kapcsolatos legfontosabb lineáris algebrai fogalmak összefoglalása megtalálható a [4] jegyzet első fejezetében. Kiemelünk azonban néhány fogalmat, melyek különösen fontosak a feladatmegoldások során, és nem tartoznak a standard lineáris algebrai ismeretekhez.

Vektorokhoz és mátrixokhoz normát rendelhetünk, amik segítségével mérhetjük a „hosszukat" és a „távolságukat". A leggyakrabban használt vektornormák az

$$\|\overline{\mathbf{x}}\|_1 = |x_1| + \cdots + |x_n|$$

1-es vagy oktaédernorma, az

$$\|\overline{\mathbf{x}}\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2}$$

2-es vagy euklideszi norma, ill. az

$$\|\overline{\mathbf{x}}\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$$

maximumnorma.

Bizonyos vektornormák származtathatók skaláris szorzatból az $\|x\| = \sqrt{\langle x, x\rangle}$ képlettel. $\mathbb{R}^n$-en a szokásos skaláris szorzat $\langle \overline{\mathbf{x}}, \overline{\mathbf{y}}\rangle = \overline{\mathbf{x}}^T\overline{\mathbf{y}} = x_1y_1 + \ldots + x_ny_n$. Ez a skaláris szorzat az euklideszi normát indukálja.

Vektornormák segítségével ún. mátrixnormákat definiálhatunk az

$$\|\mathbf{A}\| = \sup_{\overline{\mathbf{x}}\neq\overline{\mathbf{0}}} \frac{\|\mathbf{A}\overline{\mathbf{x}}\|}{\|\overline{\mathbf{x}}\|} \tag{1.1}$$

formulával. A nevezetes vektornormák által indukált mátrixnormák az alábbiak: oktaédernorma esetén

$$\|\mathbf{A}\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \text{ (oszlopösszegnorma)},$$

euklideszi norma esetén

$$\|\mathbf{A}\|_2 = \sqrt{\varrho(\mathbf{A}^H\mathbf{A})} \text{ (spektrálnorma)}$$

és maximumnorma esetén

$$\|\mathbf{A}\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}| \text{ (maximum- vagy sorösszeg norma)} \tag{1.2}$$

(ahol $\varrho(\mathbf{A})$ az $\mathbf{A}$ mátrix spektrálsugara, és $\mathbf{A}^H$ az $\mathbf{A}$ mátrix transzponált konjugáltja).

1.1. Tétel ***(Indukált mátrixnorma tulajdonságai.)*** *Ha az $\|\cdot\|$ vektornorma a $\|\cdot\|$ mátrixnormát indukálta, akkor igazak az alábbi tulajdonságok:*

- $\|\boldsymbol{A}\overline{\boldsymbol{x}}\| \leq \|\boldsymbol{A}\| \cdot \|\overline{\boldsymbol{x}}\|$ *(konzisztencia tulajdonság),*
- $\|\boldsymbol{E}\| = 1$ *(az egységmátrix normája 1),*
- $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|$ *(szubmultiplikativitási tulajdonság).*

1.2. Tétel ***(Sajátértékek becslése a normával.)*** *Indukált mátrixnormák esetén*

$$\varrho(\boldsymbol{A}) \leq \|\boldsymbol{A}\|.$$

1.3. Tétel ***(Mátrixhatványok és Neumann-sor konvergenciája.)*** *Egy $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ mátrix esetén pontosan akkor igaz, hogy $\boldsymbol{A}^k \to \boldsymbol{0}$ elemenként, ha $\varrho(\boldsymbol{A}) < 1$. Pontosan ugyanekkor lesz a*

$$\sum_{k=0}^{\infty} \boldsymbol{A}^k$$

sor konvergens, és összege az $(\boldsymbol{E} - \boldsymbol{A})^{-1}$ mátrix.

1.4. Tétel ***(Banach-féle fixponttétel.)*** *Tegyük fel, hogy az $F : H \to H$ függvény egy Banach-tér zárt H részhalmazán értelmezett kontrakció (van olyan $0 \leq q < 1$ szám melyre $\|F(x) - F(y)\| \leq q\|x - y\|$ minden $x, y \in H$ esetén). Ekkor az $x_{k+1} = F(x_k)$ iteráció tetszőleges $x_0 \in H$ elemről indítva olyan egyértelműen meghatározott $x^\star \in H$ elemhez tart, melyre $F(x^\star) = x^\star$. Az $x^\star$ elemet a leképezés fixpontjának nevezzük, továbbá érvényes az*

$$\|x^\star - x_k\| \leq \frac{q^k}{1-q}\|x_1 - x_0\|$$

becslés. A q számot kontrakciós tényezőnek nevezzük.

Az olyan $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrixokat, melyek főátlón kívüli elemei nempozitívak, nemszingulárisak és inverzük nemnegatív, M-mátrixoknak nevezzük. Könnyen látható, hogy az M-mátrixok főátlójában mindig pozitív számok állnak.

1.5. Tétel ***(M-mátrixok karakterizációja.)*** *Legyen az* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *mátrix olyan, hogy a főátlóján kívüli elemek nempozitívak. Ekkor* $\boldsymbol{A}$ *pontosan akkor M-mátrix, ha van olyan* $\overline{\boldsymbol{g}} > \boldsymbol{0}$ *vektor, mellyel* $\boldsymbol{A}\overline{\boldsymbol{g}} > \boldsymbol{0}$*.*

1.6. Tétel ***(Felső becslés M-mátrix inverzének maximumnormájára.)*** *Legyen* $\boldsymbol{A}$ *M-mátrix és* $\overline{\boldsymbol{g}} > \boldsymbol{0}$ *egy olyan vektor, melyre* $\boldsymbol{A}\overline{\boldsymbol{g}} > \boldsymbol{0}$ *teljesül. Ekkor*

$$\|\boldsymbol{A}^{-1}\|_\infty \leq \frac{\|\overline{\boldsymbol{g}}\|_\infty}{\min_i(\boldsymbol{A}\overline{\boldsymbol{g}})_i}.$$

1.2. Feladatok

1.2.1. Nevezetes mátrixtípusok

1.1. Tekintsük az alábbi mátrixot

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ -1 & 3 & 0 \\ 0 & -1 & 3 \end{bmatrix}!$$

Diagonalizálható-e ez a mátrix? Válaszunkat indokoljuk! ⟶ ⟹

1.2. Adjunk példát nem diagonalizálható, ill. nem normális diagonalizálható mátrixra! ⟹

1.3. Adjuk meg az alábbi mátrixok sajátvektorait és sajátértékeit! Ha lehetséges, akkor diagonalizáljuk őket!

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -8 & -12 & -6 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 3 & 2 & 4 \\ 1 & 4 & 4 \\ -1 & -2 & -2 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 5 & 1 & -1 \\ 1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

⟹

1.4. Igazoljuk, hogy ha $\mathbf{A}$ páratlan × páratlan méretű kvadratikus mátrix, melyre $\det \mathbf{A} = 1$ és $\mathbf{A}$ ortogonális, akkor 1 sajátértéke $\mathbf{A}$-nak! ⟶ ⟹

1.5. Határozzuk meg az $\mathbf{A} - \lambda\overline{\mathbf{v}}\overline{\mathbf{v}}^T$ mátrix sajátértékeit és sajátvektorait, ha tudjuk, hogy $\mathbf{A}$ egy szimmetrikus mátrix, melynek λ egy sajátértéke és $\overline{\mathbf{v}}$ a hozzá tartozó sajátvektor! ⟶ ⟹

1.6. Igazoljuk, hogy az $\mathbf{M} = \text{tridiag}[-1, 2, -1]$ alakú mátrixok M-mátrixok! ⟶ ⟹

1.7. Igazoljuk, hogy ha egy szimmetrikus M-mátrixnak szigorúan domináns a főátlója, akkor a mátrix pozitív definit! $\longrightarrow \Longrightarrow$

1.8. Igazoljuk, hogy a szimmetrikus M-mátrixok pozitív definitek! $\longrightarrow \Longrightarrow$

1.9. Igazoljuk, hogy az $\mathbf{M} = \text{tridiag}[-1, 2, -1]$ alakú mátrixok (szimmetrikus) pozitív definitek! $\longrightarrow \Longrightarrow$

1.10. Határozzuk meg az $\mathbf{M} = \text{tridiag}[-1, 2, -1]$ alakú mátrixok sajátértékeit és sajátvektorait! $\longrightarrow \Longrightarrow$

1.11. Igazoljuk, hogy ha $\mathbf{A} \in \mathbb{R}^{n\times n}$ ferdén szimmetrikus, akkor az

$$(\mathbf{E} + \mathbf{A})^{-1}(\mathbf{E} - \mathbf{A})$$

mátrix ($\mathbf{A}$ ún. Cayley-transzformáltja) ortogonális mátrix! $\longrightarrow \Longrightarrow$

1.12. Igazoljuk, hogy felső háromszögmátrixok szorzata és inverze (ha létezik) is felső háromszögmátrix! $\longrightarrow \Longrightarrow$

1.13. Igazoljuk, hogy ha egy $\mathbf{T}$ felső háromszögmátrixra $\mathbf{T}^T\mathbf{T} = \mathbf{T}\mathbf{T}^T$, akkor $\mathbf{T}$ diagonális mátrix! $\longrightarrow \Longrightarrow$

1.2.2. Normált és euklideszi terek

1.14. Igazoljuk, hogy euklideszi térben a skaláris szorzat által indukált normával a skaláris szorzat az

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{4}\left(\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2\right)$$

módon fejezhető ki (polarizációs egyenlőség)! $\Longrightarrow$

1.15. Igazoljuk, hogy ha egy normált tér normáját skaláris szorzatból származtattuk, akkor a normára igaz az ún. parallelogramma-egyenlőség

$$\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2!$$

$\Longrightarrow$

1.16. Igazoljuk, hogy normált tér normája legfeljebb egy skaláris szorzatból származtatható! $\Longrightarrow$

1.17. Igazoljuk az euklideszi terekben érvényes Cauchy–Schwarz–Bunyakovszkij-egyenlőtlenséget:

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|!$$

$\longrightarrow \Longrightarrow$

1.18. Igazoljuk, hogy euklideszi térben $\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 = \|\mathbf{x}+\mathbf{y}\|^2$ pontosan akkor teljesül, ha $\mathbf{x}$ és $\mathbf{y}$ ortogonálisak! $\Longrightarrow$

1.2.3. Banach-féle fixponttétel

1.19. Tegyük fel, hogy az $F : [a,b] \to [a,b]$, $F([a,b]) \subset [a,b]$ függvényre igaz, hogy valamilyen m pozitív egészre a $T := F^m = F \circ F \circ \ldots \circ F$ függvény kontrakció az $[a,b]$ intervallumon. Igazoljuk, hogy az F függvénynek pontosan egy fixpontja van! $\longrightarrow \Longrightarrow$

1.20. Tekintsük az $F : [1,\infty) \to [1,\infty)$, $F(x) = x/2 + 1/x$ függvényt. Igazoljuk, hogy F kontrakció! Határozzuk meg a lehető legkisebb kontrakciós tényezőt! Adjuk meg F fixpontját! $\longrightarrow \Longrightarrow$

1.21. Tegyük fel, hogy a Banach-féle fixponttétel feltételei közül a kontrakciós feltételt $(\exists\ 0 \le q < 1,\ \|F(x) - F(y)\| \le q\|x - y\|,\ \forall\ x, y \in H)$ kicseréljük az

$$\|F(x) - F(y)\| < \|x - y\|,\ \forall\ x, y \in H$$

feltételre! Igazoljuk, hogy ekkor F-nek maximum egy fixpontja lehet, de az is lehet, hogy nincs fixpont! $\longrightarrow \Longrightarrow$

1.22. Tegyük fel, hogy T kontrakció a V Banach-téren! Igazoljuk, hogy tetszőleges $\mathbf{y} \in V$ esetén az $\mathbf{x} = T(\mathbf{x}) + \mathbf{y}$ egyenletnek pontosan egy $\mathbf{x}$ megoldása van, és az $\mathbf{x}$ megoldás folytonosan függ $\mathbf{y}$-tól! $\longrightarrow \Longrightarrow$

1.23. Igazoljuk, hogy ha $f : \mathbb{R} \to \mathbb{R}$ folytonosan differenciálható az $[a,b]$ intervallumon, és $|f'(x)| < 1$ minden $x \in [a,b]$ esetén, akkor f kontrakció $[a,b]$-n! $\longrightarrow \Longrightarrow$

1.2.4. Vektornormák

1.24. Adjuk meg az $\overline{\mathbf{x}} = [1,-2,3]^T$ vektor 1-es, 2-es és maximumnormáját! $\longrightarrow \Longrightarrow$

1.25. Adjuk meg az $\overline{\mathbf{x}} = [1,2,\ldots,100]^T$ vektor 1-es, 2-es és maximumnormáját! $\longrightarrow \Longrightarrow$

1.26. Azonosítsuk $\mathbb{R}^2$ elemeit a sík pontjaival! Adjuk meg a síkon azon pontok halmazát, melyek távolsága az origótól kisebb, mint egy! Használjuk az 1-es, 2-es és maximum-normákat! $\Longrightarrow$

1.27. Igazoljuk közvetlenül az 1-es, 2-es és maximumnormák ekvivalenciáját! $\Longrightarrow$

1.28. Igazoljuk, hogy $\mathbb{R}^n$-en sem a maximum-, sem az 1-es norma nem származtatható skaláris szorzásból! $\longrightarrow \Longrightarrow$

1.29. Igazoljuk, hogy $p \to \infty$ esetén az $\mathbb{R}^n$-en értelmezett

$$\|\overline{\mathbf{x}}\|_p = \sqrt[p]{|x_1|^p + \ldots + |x_n|^p}$$

p-norma $(1 \le p \in \mathbb{R})$ éppen a maximumnormát adja! $\Longrightarrow$

1.30. Igazoljuk az ún. Young-egyenlőtlenséget, azaz, hogy tetszőleges $a, b \geq 0$ és $1 < p, q < \infty$, $1/p + 1/q = 1$ számok esetén

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q},$$

majd ennek segítségével lássuk be az ún. Hölder-egyenlőtlenséget $\mathbb{R}^n$-en: $1 \leq p, q \leq \infty$, $1/p + 1/q = 1$ esetén

$$|\langle \overline{\mathbf{x}}, \overline{\mathbf{y}} \rangle| \leq \|\overline{\mathbf{x}}\|_p \cdot \|\overline{\mathbf{y}}\|_q!$$

⟶ ⟹

1.31. Igazoljuk, hogy a p-norma kifejezése valóban normát ad meg $\mathbb{R}^n$-en! ⟶ ⟹

1.32. Igazoljuk, hogy ha $\mathbf{A}$ nemszinguláris mátrix, akkor az $\|\overline{\mathbf{x}}\|_A := \|\mathbf{A}\overline{\mathbf{x}}\|$ hozzárendelés vektornorma bármilyen $\|\cdot\|$ vektornorma esetén! ⟹

1.2.5. Mátrixnormák

1.33. Tekintsük az $\|\mathbf{A}\| := \max_{i=1,\ldots,n}\{|a_{ij}|\}$ mátrixnormát! Igazoljuk, hogy ez valóban norma! Mutassuk meg, hogy nem lehet vektornormából származtatni. ⟶ ⟹

1.34. Igazoljuk az alábbi becsléseket, melyek nyilvánvalóan az adott mátrixnormák ekvivalenciáját mutatják!

$$\begin{aligned} \frac{1}{n}\|\mathbf{A}\|_1 \leq \|\mathbf{A}\|_\infty \leq n\|\mathbf{A}\|_1 \\ \frac{1}{\sqrt{n}}\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\|_2 \leq \sqrt{n}\|\mathbf{A}\|_\infty \\ \frac{1}{\sqrt{n}}\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_1 \leq \sqrt{n}\|\mathbf{A}\|_2 \end{aligned} \tag{1.3}$$

⟹

1.35. Igazoljuk, hogy az $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2}$ képlettel értelmezett ún. Frobenius-norma valóban norma! Lehet-e ezt a normát vektornormából származtatni? ⟹

1.36. Igazoljuk, hogy $\|\mathbf{A}\|_F^2 = \mathrm{trace}(\mathbf{A}^T\mathbf{A})$, ahol a $\mathrm{trace}(\cdot)$ jelölés az adott mátrix főátlóbeli elemeinek összegét jelenti (amely megegyezik a sajátértékek összegével is)! Igazoljuk továbbá, hogy ha $\mathbf{A}$ és $\mathbf{B}$ ortogonálisan hasonlók, akkor Frobenius-normájuk megegyezik! ⟶ ⟹

1.37. Igazoljuk, hogy a Frobenius-norma konzisztens az euklideszi vektornormával, azaz teljesül, hogy $\|\mathbf{A}\overline{\mathbf{x}}\|_2 \leq \|\mathbf{A}\|_F\|\overline{\mathbf{x}}\|_2$! ⟹

1.38. Igazoljuk, hogy a Frobenius-norma szubmultiplikatív! $\Longrightarrow$

1.39. Igazoljuk, hogy nemcsak indukált mátrixnormákra, hanem tetszőleges szubmultiplikatív mátrixnormára is igaz, hogy $\varrho(\mathbf{A}) \leq \|\mathbf{A}\|$! Az

$$\mathbf{A} = \begin{bmatrix} 0.5 & 0.6 \\ 0.1 & 0.5 \end{bmatrix}$$

mátrix 1-es, maximum- és Frobenius-normáinak értékei közül melyik biztosítja a $\varrho(\mathbf{A}) < 1$ feltételt? $\longrightarrow \Longrightarrow$

1.40. Számítsuk ki a diagonális mátrixok 1-es, 2-es és maximumnormáját az indukált mátrixnorma (1.1) képlete segítségével! $\Longrightarrow$

1.41. Milyen mátrixnormát indukál az 1-es vektornorma? $\Longrightarrow$

1.42. Milyen mátrixnormát indukál a vektorok maximumnormája? $\Longrightarrow$

1.43. Milyen mátrixnormát indukál a vektorok euklideszi-normája (2-es norma)? $\Longrightarrow$

1.44. Igazoljuk, hogy az $\|\mathbf{A}\| := n \max_{i,j=1\ldots,n}\{|a_{ij}|\}$ hozzárendelés szubmultiplikatív normát ad meg $\mathbb{R}^{n\times n}$-en! $\Longrightarrow$

1.45. Igazoljuk, hogy minden szubmultiplikatív mátrixnormához van olyan vektornorma, amivel konzisztens! $\longrightarrow \Longrightarrow$

1.46. Igazoljuk, hogy indukált mátrixnorma esetén

$$\|\mathbf{A}\| = \max\{\|\mathbf{AB}\| \mid \|\mathbf{B}\| \leq 1\}!$$

$\Longrightarrow$

1.47. Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy nemszinguláris mátrix és $\mathbf{B} \in \mathbb{R}^{n\times n}$ egy szinguláris mátrix! Igazoljuk, hogy tetszőleges indukált norma esetén $\|\mathbf{A}^{-1}\| \geq 1/\|\mathbf{A} - \mathbf{B}\|$! $\longrightarrow \Longrightarrow$

1.48. Legyen $\|\cdot\|$ egy tetszőleges indukált mátrixnorma. Igazoljuk, hogy tetszőleges $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix esetén

$$\lim_{k\to\infty} \|\mathbf{A}^k\|^{1/k} = \varrho(\mathbf{A}) \ !$$

$\longrightarrow \Longrightarrow$

1.49. Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy tetszőleges négyzetes mátrix és $\mathbf{A}^{(k)}$ az $\mathbf{A}$ mátrix k-adrendű bal felső főminormátrixa ($A(1:k, 1:k)$)! Igazoljuk, hogy $\|\mathbf{A}^{(k)}\|_2 \leq \|\mathbf{A}\|_2$! $\Longrightarrow$

1.50. Legyen

$$\mathbf{C} = \begin{bmatrix} 1 & -0.1 & -0.2 \\ -0.1 & 1 & -0.1 \\ -0.2 & -0.1 & 1 \end{bmatrix}.$$

Igazoljuk, hogy $\mathbf{C}$ invertálható és adjunk felső becslést az inverz mátrix 1-es normájára az inverz mátrix kiszámítása nélkül! $\longrightarrow \Longrightarrow$

1.51. (⊞) Adjuk meg az $n \times n$-es Hilbert-mátrix inverzének maximumnormáját n függvényében $n = 1, \ldots, 10$ esetén! $\Longrightarrow$

1.52. Adjuk meg az 5×5-ös Hilbert-mátrix 1-es, 2-es és maximumnormáját ill. spektrálsugarát! $\Longrightarrow$

2. fejezet

Modellalkotás és hibaforrásai

2.1. Képletek, összefüggések

2.1.1. Feladatok kondicionáltsága

A matematikai egyenletek egy része $d = F(x)$ alakban írható, ahol d és x valamilyen normált terek elemei. Itt a d elem ismeretében kell meghatározni az x ismeretlen elemet. Amennyiben x a d adat segítségével egyértelműen írható fel $x = G(d)$ alakban, és G deriválható d-ben, akkor a $d = F(x)$ feladat d-beli kondíciószáma

$$\kappa(d) = \frac{\|G'(d)\| \cdot \|d\|}{\|G(d)\|}. \tag{2.1}$$

Ez az érték azt adja meg, hogy x relatív megváltozása hányszorosa d relatív megváltozásának.

2.1.2. A gépi számábrázolás

A számítógépek általában ún. lebegőpontos számrendszert használnak a számok ábrázolására. Ebben a számrendszerben a számokat kerekítés után a

$$\pm b^k \left(\frac{a_0}{b^0} + \frac{a_1}{b^1} + \frac{a_2}{b^2} + \cdots + \frac{a_{p-1}}{b^{p-1}} \right) \equiv a_0.a_1a_2 \ldots a_{p-1} \times b^k$$

alakban írjuk fel, ahol b a számábrázolás alapja, p a szereplő számjegyek (mantissza) száma és k a kitevő (karakterisztika). Az a_i $(i = 0, \ldots, p-1)$ számjegyekről feltesszük, hogy azok az alapnál kisebb nemnegatív egész számok. Ha $a_0 \neq 0$, akkor azt mondjuk, hogy a felírt szám normálalakban van.

Több feladat esetén az egyszerűség kedvéért a tízes számrendszert használjuk ($b = 10$), mert ehhez vagyunk hozzászokva, és ez is mutatja a lebegőpontos számábrázolás tulajdonságait és korlátait.

$F(p, k_{\min}, k_{\max})$ fogja jelölni azt a tízes alapú lebegőpontos számrendszert, amiben a mantissza hossza p, a minimális karakterisztika $k_{\min}$, a maximális pedig $k_{\max}$. Ugyanezt a számrendszert kettes alap esetén az $F_2(p, k_{\min}, k_{\max})$ módon fogjuk jelölni azzal a megkötéssel, hogy ilyenkor p a mantissza kettedespont utáni jegyeinek számát jelenti, hiszen előtte a normálalakban csak 1-es jegy állhat. A MATLAB szokásos dupla pontosságú lebegőpontos számai és normálalak esetén $p = 52$, $k_{\min} = -2^{10} - 2$ és $k_{\max} = 2^{10} - 1$.

A számítógépen való műveletek végrehajtását az alábbi módon fogjuk modellezni.

Egy x valós szám lebegőpontos képét úgy kapjuk meg, hogy normálalakra hozzuk, és a mantisszát a számrendszerben adott mantisszahosszra kerekítjük. Jelölése: $fl(x)$. Ha $|x|$ nagyobb, mint az ábrázolható legnagyobb szám, akkor $fl(x) = Inf$, ha pedig $|x| < \varepsilon_0$, azaz a legkisebb pozitív ábrázolható szám, akkor $fl(x) = 0$. Ezzel a jelöléssel írhatjuk, hogy a számítógép az $x \diamond y$ művelet eredménye helyett az $x \boxed{\diamond} y := fl(fl(x) \diamond fl(y))$ értéket adja vissza.

2.1. Tétel *Legyen $x \in \mathbb{R}$ olyan szám, melyre $|x| \leq M$ (M a legnagyobb ábrázolható lebegőpontos szám). Ekkor érvényes, hogy*

$$fl(x) = (1 + \delta)x, \ |\delta| \leq \mathbf{u},$$

ahol $\mathbf{u}$ a gépi pontosság értéke, azaz az 1 után következő lebegőpontos szám 1-től mért távolságának fele (MATLAB-ban kb. 10^{-16}).

2.2. Feladatok

2.2.1. Feladatok kondicionáltsága

2.1. Vizsgáljuk meg az $x = -d + \sqrt{d^2 - 4}$ kifejezés kondicionáltságát a d változó függvényében! Milyen d értékek esetén lesz korrekt kitűzésű a feladat? Adjunk meg olyan d értéket, melyre a (relatív) kondíciószám 100-nál nagyobb! $\longrightarrow \Longrightarrow$

2.2. Tekintsük az $x+dy = 1$, $dx+y = 0$ egyenletrendszert. Jól vagy rosszul kondicionált az x megoldás, ill. a megoldások $x+y$ összegének kiszámítása a d paraméter függvényében, ha $d \approx 1$? Adjuk meg mindkét esetben a relatív kondíciószám értékét a $d = 0.99$ esetre! $\longrightarrow \Longrightarrow$

2.3. Számítsuk ki az $x - \sqrt{d+1} + \sqrt{d} = 0$ feladat (d a bemenő adat, x pedig a kimenő adat) relatív kondíciószámát! Mikor lesz rosszul és mikor jól kondicionált a feladat? $\Longrightarrow$

2.4. Legyenek f és g differenciálható valós-valós függvények! Hogyan becsülhető az $x = f(d)$ és $x = g(d)$ feladatok kondíciószámával az $x = (f \cdot g)(d)$ feladat kondíciószáma azon d pontokban, melyekben a kondíciószám értelmezhető? (A szokásos módon d a feladat bemenő adata és x a kimenő adat.) $\Longrightarrow$

2.5. Vizsgáljuk meg, hogy korrekt kitűzésű-e az $x+dy = 1$, $dx+y = 0$ egyenletrendszer a d valós paraméter függvényében! Adjuk meg a kondíciószámot maximumnormában! $\Longrightarrow$

2.2.2. A gépi számábrázolás

2.6. Adjuk meg az $F(1,-2,2)$ lebegőpontos számrendszerben pontosan ábrázolható számokat! $\Longrightarrow$

2.7. Adjuk meg az $F(1,-2,2)$ rendszerben az $1/3$, $1/900$, $20\cdot 200$, $(((2+0.1)+0.1)+\cdots+0.1)$, $(((0.1+0.1)+0.1)+\cdots+0.1)+2$ (10 összeadás) értékeket! $\Longrightarrow$

2.8. Adjunk meg olyan lebegőpontos számrendszert $F(p,k_{\min},k_{\max})$ alakban, melyben az alábbi számok ábrázolhatók!

a) 5,50,500,5000;
b) 5,5.5,5.55;
c) 5,0.5,0.05,0.005;
d) 5,55,555,5555! $\Longrightarrow$

2.9. Milyen lebegőpontos számrendszerben számolható kerekítés nélkül

a) $2.2\cdot 3.45$,
b) $1/80$,
c) $2\times 10^2\cdot 7\times 10^2$? $\Longrightarrow$

2.10. Két pozitív számot, x és y, elosztunk egymással egy olyan számítógépen, melynek gépi pontossága $\mathtt{u}$. Jelölje z a hányados pontos értékét és $\hat{z}$ a számított értéket. Adjunk becslést a $|z-\hat{z}|$ és $|z-\hat{z}|/|z|$ abszolút és relatív hibákra! $\Longrightarrow$

2.11. Az $a=0.001$ választás mellett $A=1-1/(1-2a)$ értéke -0.002004008016. Határozzuk meg mi is A értékét egy tízes számrendszerű, hatjegyű mantisszás lebegőpontos számokat használó számítógépen! Javasoljunk numerikus szempontból jobb számolást A-ra és végezzük el úgy is a számolásokat! $\Longrightarrow$

2.12. A $\sum_{i=1}^{\infty}1/i$ harmonikus sor összege $+\infty$. Megkapnánk-e ezt az eredményt úgy, hogy egyre több tagot adunk össze a sorból a MATLAB segítségével? Mekkora összeget kapnánk egy $F(2,-1,1)$ lebegőpontos számokat használó számítógépen, ha a gép csak normálalakban lévő számokat tud ábrázolni? $\longrightarrow\Longrightarrow$

2.13. Egy 10-es számrendszeren alapuló számítógép a $\sin x$, $\cos x$, x^2 függvények értékeit pontosan számolja, majd az eredmények ábrázolásánál hatjegyű mantisszára kerekít. Határozzuk meg ezen a számítógépen az $f(x)=\cos^2 x-\sin^2 x$ függvény értékét az $x=0.7854$ helyen! Mekkora a számított eredmény relatív hibája? Indokoljuk az eredményt! Javasoljunk jobb képletet az $f(x)$ érték kiszámítására! $\longrightarrow\Longrightarrow$

2.14. Az $x^2 + ax + b = 0$ egyenletet szeretnénk megoldani az

$$x_{1,2} = (-a \pm \sqrt{a^2 - 4b})/2$$

megoldóképlettel. Milyen végeredményt adna a MATLAB az $a = -500000000$ és $b = 1$ paraméterekkel? Becsüljük meg, hogy melyik eredmény elfogadható és melyik nem! Hogyan számolhatnánk ki MATLAB-ban a zérushelyeket pontosabban? $\Longrightarrow$

2.15. Szimpla pontosságú lebegőpontos számokat használva (32 biten tároljuk a számokat: 1 előjelbit, 8 bit a karakterisztika és 23 bit a mantissza tárolására) szeretnénk közelíteni számítógépen a $\sum_{i=1}^{\infty} 1/i^2$ sor összegét $(\pi^2/6)$! Az

$$\frac{1}{1^2} + \frac{1}{2^2} + \ldots + \frac{1}{4096^2}$$

összegre 1.6447253 adódott. Mennyivel tér el az

$$s_k = \sum_{i=1}^{k} \frac{1}{i^2}$$

sorozat számítógépen számolt határértéke a tényleges sorösszegtől? Javasoljunk jobb módszert az összeg számítógépes közelítésére! $\longrightarrow \Longrightarrow$

2.16. Az $x = 0.1$ tízes számrendszerbeli szám kettes számrendszerbeli alakja a $0.000\overline{1100}$ szakaszos tizedes tört, ahol az utolsó négy számjegy ismétlődik. Fejezzük ki az $(x - fl(x))/x$ relatív hiba értékét az $\mathtt{u}$ gépi pontosság segítségével, ha $fl(x)$ az x szám szimpla pontosságú lebegőpontos képe (32 biten tároljuk a számokat: 1 előjelbit, 8 bit a karakterisztika és 23 bit a mantissza tárolására)! $\longrightarrow \Longrightarrow$

2.17. (⊞) Írjunk MATLAB programot az

$$y_{k+1} = 2^{k+1}\sqrt{\frac{1}{2}\left(1 - \sqrt{1 - (2^{-k}y_k)^2}\right)}$$

iteráció vizsgálatára! Ismert, hogy ebben az iterációban $y_k \to \pi$, mert a y_k az egységkörbe írt szabályos 2^k szög félkerületét adja meg. Hasonlítsuk össze az eredményt az

$$y_{k+1} = y_k\sqrt{\frac{2}{1 + \sqrt{1 - (2^{-k}y_k)^2}}}$$

iterációval! $\Longrightarrow$

2.18. (⊞) Írjunk MATLAB programot az

$$e^x = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{x^i}{i!}$$

sor részletösszegeinek kiszámítására! Futtassuk negatív értékek esetén (pl. $x = -25$)! Mit tapasztalunk? $\Longrightarrow$

2.19. Az $x^2 - 1634x + 2 = 0$ egyenletet szeretnénk megoldani olyan számítógépen, amely a számok ábrázolásához tízes számrendszerbeli lebegőpontos számokat használ 4-jegyű mantisszával (a karakterisztikára nincs megkötés). Az x_2 megoldásra nulla adódik. Mi ennek az oka? Számítsuk ki x_1-et, és javasoljunk hatékonyabb módszert x_2 kiszámítására az $x_1 x_2$ szorzat értékét felhasználva! Számítsuk ki ezzel a módszerrel x_2 értékét! $\Longrightarrow$

3. fejezet

Lineáris egyenletrendszerek megoldása

3.1. Képletek, összefüggések

3.1.1. Kondicionáltság

A megoldás együtthatóktól való függését adja meg az alábbi tétel, ahol $\kappa(\mathbf{A})$ az $\mathbf{A}$ mátrix adott normabeli kondíciószámát jelenti.

3.1. Tétel ***(Lineáris egyenletrendszerek kondicionáltsága.)*** *Tegyük fel, hogy az* $\boldsymbol{A}\overline{\boldsymbol{x}} = \overline{\boldsymbol{b}}$ *(*$\boldsymbol{A} \in \mathbb{R}^{n\times n}, \det(A) \neq 0$*) egyenletrendszer helyett az* $(\boldsymbol{A} + \boldsymbol{\delta A})\overline{\boldsymbol{y}} = \overline{\boldsymbol{b}} + \boldsymbol{\delta}\overline{\boldsymbol{b}}$ *perturbált egyenletrendszert oldjuk meg, és az együtthatómátrix perturbációjára teljesül a* $\|\boldsymbol{\delta A}\| < 1/\|\boldsymbol{A}^{-1}\|$ *feltétel valamilyen indukált normában. Ekkor a perturbált egyenletrendszernek is egyértelmű megoldása van. Ezt a megoldást* $\overline{\boldsymbol{y}} = \overline{\boldsymbol{x}} + \boldsymbol{\delta}\overline{\boldsymbol{x}}$ *alakban írva érvényes az alábbi becslés:*

$$\frac{\|\boldsymbol{\delta}\overline{\boldsymbol{x}}\|}{\|\overline{\boldsymbol{x}}\|} \leq \frac{\kappa(\boldsymbol{A})}{1 - \kappa(\boldsymbol{A})\|\boldsymbol{\delta A}\|/\|\boldsymbol{A}\|} \cdot \left(\frac{\|\boldsymbol{\delta}\overline{\boldsymbol{b}}\|}{\|\overline{\boldsymbol{b}}\|} + \frac{\|\boldsymbol{\delta A}\|}{\|\boldsymbol{A}\|} \right).$$

A tétel bizonyításához használtuk az alábbi, önmagában is hasznos állítást.

3.2. Tétel ***(Becslés perturbált mátrix inverzének normájára.)*** *Legyen* $\boldsymbol{S} = \boldsymbol{E} + \boldsymbol{R} \in \mathbb{R}^{n\times n}$*, ahol* $\|\boldsymbol{R}\| =: q < 1$ *valamilyen indukált normában. Ekkor* $\boldsymbol{S}$ *reguláris, és*

$$\|\boldsymbol{S}^{-1}\| \leq \frac{1}{1-q}.$$

3.1.2. Direkt módszerek

Direkt módszernek nevezzük azokat a megoldási módszereket, melyekkel véges sok alapművelet segítségével meghatározható a megoldás. A direkt módszereknél fontos szerepe van az együtthatómátrixok szorzatfelbontásainak, melyeket előre elkészítve, az újabb, az eredetivel megegyező mátrixú egyenletrendszerek megoldása már egy nagyságrenddel kevesebb művelettel megvalósítható.

3.3. Tétel ***(LU-felbontás.)*** *Tegyük fel, hogy az* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *mátrixra* $\det(\boldsymbol{A}(1:k,1:k)) \neq 0$ $(k = 1,\ldots,n-1)$. *Ekkor létezik egy olyan* $\boldsymbol{L}$ *normált alsó háromszögmátrix és egy* $\boldsymbol{U}$ *felső háromszögmátrix, melyekkel* $\boldsymbol{A} = \boldsymbol{LU}$. *Ha egy reguláris mátrixnak létezik LU-felbontása, akkor az LU-felbontása egyértelmű.*

Az LU-felbontást a Gauss-módszer segítségével határozhatjuk meg.

3.4. Tétel ***(Általános LU-felbontás.)*** *Azok a mátrixok, melyeknek van LU-felbontása, felírhatók* $\boldsymbol{A} = \boldsymbol{LDM}^T$ *alakban is, ahol* $\boldsymbol{L}$ *és* $\boldsymbol{M}$ *is normált alsó háromszögmátrix,* $\boldsymbol{D}$ *pedig diagonális mátrix. Itt* $\boldsymbol{D}$ *az* $\boldsymbol{U}$ *mátrix diagonálisa,* $\boldsymbol{M}$ *pedig* $\boldsymbol{D}^{-1}\boldsymbol{U}$. *Ha* $\boldsymbol{A}$ *szimmetrikus, akkor* $\boldsymbol{M} = \boldsymbol{L}$.

3.5. Tétel ***(Cholesky-felbontás.)*** *Tegyük fel, hogy* $\boldsymbol{A}$ *egy szimmetrikus, pozitív definit mátrix. Ekkor létezik pontosan egy olyan pozitív diagonálisú* $\boldsymbol{G}$ *alsó háromszögmátrix, mellyel* $\boldsymbol{A} = \boldsymbol{GG}^T$.

A Cholesky-felbontásban szereplő $\mathbf{G}$ mátrix elemeit direkt módon, fentről lefelé és balról jobbra haladva az $\mathbf{A} = \mathbf{GG}^T$ egyenlőséget felhasználva határozhatjuk meg.

3.6. Tétel ***(Általános LU-felbontás.)*** *Legyen* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *egy tetszőleges mátrix. Ekkor létezik egy olyan* $\boldsymbol{L}$ *alsó normált háromszögmátrix, melynek elemei egynél nem nagyobb abszolút értékűek, egy* $\boldsymbol{U}$ *felső háromszögmátrix, és egy* $\boldsymbol{P}$ *permutációs mátrix, melyekkel* $\boldsymbol{PA} = \boldsymbol{LU}$.

Az általános LU-felbontás a részleges főelemkiválasztással kombinált Gauss-módszerrel határozható meg.

3.7. Tétel ***(QR-felbontás.)*** *Legyen* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ $(m \geq n)$ *egy teljes oszloprangú mátrix. Ekkor léteznek olyan* $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ *ortogonális és* $\boldsymbol{R} \in \mathbb{R}^{m\times n}$ *felső háromszögmátrixok, melyekkel* $\boldsymbol{A} = \boldsymbol{QR}$.

A QR-felbontás egymásutáni megfelelő Householder-tükrözésekkel vagy Givens-forgatásokkal előállítható elő.

3.8. Tétel ***(Householder-tükrözés.)*** *Legyen* $\boldsymbol{0} \neq \overline{\boldsymbol{x}} \in \mathbb{R}^n$ *vektor. Vezessük be a* $\overline{\boldsymbol{v}} = \overline{\boldsymbol{x}} \pm \|\overline{\boldsymbol{x}}\|_2 \overline{\boldsymbol{e}}_1$ *és*

$$\boldsymbol{H} = \boldsymbol{E} - \frac{2\overline{\boldsymbol{v}}\overline{\boldsymbol{v}}^T}{\overline{\boldsymbol{v}}^T\overline{\boldsymbol{v}}}$$

jelöléseket ($\boldsymbol{H}$ *szimmetrikus és ortogonális mátrix). Ekkor igaz a*

$$\boldsymbol{H}\overline{\boldsymbol{x}} = \mp\|\overline{\boldsymbol{x}}\|_2 \overline{\boldsymbol{e}}_1$$

egyenlőség. A $\boldsymbol{H}$ *mátrixot az* $\overline{\boldsymbol{x}}$ *vektorhoz tartozó Householder-tükrözésnek nevezzük.*

3.9. Tétel ***(Givens-forgatás.)*** *Legyen* $\overline{\boldsymbol{x}} \in \mathbb{R}^n$ *olyan vektor, melyben két* $i < j$ *indexre* $x_i^2 + x_j^2 \neq 0$*. Legyen továbbá*

$$s = \frac{\pm x_j}{\sqrt{x_i^2 + x_j^2}}, \quad c = \frac{\mp x_i}{\sqrt{x_i^2 + x_j^2}}$$

és

$$\boldsymbol{G}(i,j,\theta) = \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & c & & & -s & & \\ & & & 1 & & & & \\ & & & & \ddots & & & \\ & & & & & 1 & & \\ & & s & & & & c & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

(c *és* s *az* i*-edik és* j*-edik sorban és oszlopban szerepel,* $\boldsymbol{G}(i,j,\theta)$ *ortogonális mátrix). Ekkor a* $\boldsymbol{G}(i,j,\theta)\overline{\boldsymbol{x}}$ *vektor* j*-edik eleme 0 lesz.*

3.1.3. Iterációs módszerek

Az iterációs módszerek esetén az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszer megoldását egy alkalmasan választott iterációs sorozat határértékeként állítjuk elő.

Klasszikus iterációk

Az iterációt az $\mathbf{A} = \mathbf{S} - \mathbf{T}$ felbontással ($\mathbf{S}$ invertálható mátrix) az

$$\overline{\mathbf{x}}^{(k+1)} = \mathbf{S}^{-1}\mathbf{T}\overline{\mathbf{x}}^{(k)} + \mathbf{S}^{-1}\overline{\mathbf{b}} =: \mathbf{B}\overline{\mathbf{x}}^{(k)} + \overline{\mathbf{f}}$$

módon állítjuk elő. A nevezetes módszerek esetén az alábbi $\mathbf{B}$ iterációs mátrixokat és $\overline{\mathbf{f}}$ vektorokat választjuk ($\mathbf{D}$ $\mathbf{A}$ diagonálisa, $\mathbf{L}$ és $\mathbf{U}$ rendre az $\mathbf{A}$ mátrix főátló alatti és feletti részének (-1)-szerese, ω pedig egy megfelelő valós paraméter).

Módszer neve	$\mathbf{B}$	$\overline{\mathbf{f}}$
Jacobi	$\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})$	$\mathbf{D}^{-1}\overline{\mathbf{b}}$
Relaxált Jacobi (JOR)	$\mathbf{E}-\omega\mathbf{D}^{-1}\mathbf{A}$	$\omega\mathbf{D}^{-1}\overline{\mathbf{b}}$
Gauss–Seidel	$(\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}$	$(\mathbf{D}-\mathbf{L})^{-1}\overline{\mathbf{b}}$
Relaxált Gauss–Seidel (SOR)	$(\mathbf{D}-\omega\mathbf{L})^{-1}((1-\omega)\mathbf{D}+\omega\mathbf{U})$	$\omega(\mathbf{D}-\omega\mathbf{L})^{-1}\overline{\mathbf{b}}$

3.10. Tétel ***(Iterációs módszerek konvergenciája.)*** *A fenti módokon konstruált*

$$\overline{\boldsymbol{x}}^{(k+1)} = \boldsymbol{B}\overline{\boldsymbol{x}}^{(k)} + \overline{\boldsymbol{f}}$$

iteráció pontosan akkor tart az $\boldsymbol{A}\overline{\boldsymbol{x}} = \overline{\boldsymbol{b}}$ egyenletrendszer megoldásához tetszőleges kezdővektor esetén, ha $\varrho(\boldsymbol{B}) < 1$.

3.11. Tétel ***(Nevezetes mátrixú egyenletrendszerek konvergenciája.)*** *Szigorúan diagonálisan domináns mátrixokra a Gauss–Seidel és a Jacobi-módszer is konvergál. Szimmetrikus, pozitív definit mátrixokra a Gauss–Seidel-módszer konvergál. M-mátrixokra a Jacobi-, a Gauss–Seidel- és ezek relaxált változatai is konvergálnak ($\omega \in (0,1)$ mellett). A relaxált Gauss–Seidel-iteráció csak $\omega \in (0,2)$ esetén lehet konvergens. Ez a feltétel szimmetrikus, pozitív definit mátrixok esetén elégséges is.*

Variációs módszerek

A variációs módszerek esetén szimmetrikus, pozitív definit mátrixú egyenletrendszerek megoldását keressük. Ezek megoldása ekvivalens a

$$\phi(\overline{\mathbf{x}}) = \frac{1}{2}\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{x}}^T\overline{\mathbf{b}}$$

többváltozós függvény abszolút minimumának megkeresésével, ugyanis az abszolút minimum a $\overline{\mathbf{x}}^\star = \mathbf{A}^{-1}\overline{\mathbf{b}}$ pontban van és értéke $-\overline{\mathbf{b}}^T\mathbf{A}^{-1}\overline{\mathbf{b}}/2$.

Az abszolút minimum keresésének alapja az ún. egyenes menti keresés, amikor egy pontból egy adott irányban keressük meg az iránymenti minimumot.

3.12. Tétel ***(Iránymenti minimumok megkeresése.)*** *Legyenek $\overline{\boldsymbol{x}}$ és $\overline{\boldsymbol{p}} \neq \boldsymbol{0}$ adott vektorok. A $g(\alpha) = \phi(\overline{\boldsymbol{x}} + \alpha\overline{\boldsymbol{p}})$ egyváltozós függvény egyértelmű minimumát az $\alpha = \overline{\boldsymbol{p}}^T\overline{\boldsymbol{r}}/(\overline{\boldsymbol{p}}^T\boldsymbol{A}\overline{\boldsymbol{p}})$ választás esetén veszi fel, ahol $\overline{\boldsymbol{r}}$ a $\overline{\boldsymbol{b}} - \boldsymbol{A}\overline{\boldsymbol{x}}$ maradékvektor.*

A gradiens módszer esetén a maradékvektorokat (gradiens vektorral ellentétes vektor) választjuk keresési iránynak, és sorozatos egyenes menti keresésekkel jutunk el az abszolút minimumhoz.

A gradiens módszer algoritmusa a következő:

$$
\begin{aligned}
&k := 0, \overline{\mathbf{r}}_0 := \overline{\mathbf{b}}, \overline{\mathbf{x}}_0 := \mathbf{0} \\
&\textbf{while } \overline{\mathbf{r}}_k \neq \mathbf{0} \\
&\quad k := k + 1 \\
&\quad \alpha_k := \overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1} / (\overline{\mathbf{r}}_{k-1}^T \mathbf{A} \overline{\mathbf{r}}_{k-1}) \\
&\quad \overline{\mathbf{x}}_k := \overline{\mathbf{x}}_{k-1} + \alpha_k \overline{\mathbf{r}}_{k-1} \\
&\quad \overline{\mathbf{r}}_k := \overline{\mathbf{b}} - \mathbf{A} \overline{\mathbf{x}}_k \\
&\textbf{end while}
\end{aligned}
$$

3.13. Tétel ***(A gradiens-módszer konvergenciája.)*** *A gradiens-módszer során érvényes a*

$$\frac{\phi(\overline{\boldsymbol{x}}_{k+1}) + (1/2)\overline{\boldsymbol{b}}^T \boldsymbol{A}^{-1} \overline{\boldsymbol{b}}}{\phi(\overline{\boldsymbol{x}}_k) + (1/2)\overline{\boldsymbol{b}}^T \boldsymbol{A}^{-1} \overline{\boldsymbol{b}}} \leq 1 - \frac{1}{\kappa_2(\boldsymbol{A})}$$

becslés $(k = 0, 1, \ldots)$.

A konjugált gradiens-módszer esetén a keresési irányokat mindig úgy választjuk, hogy azok legyenek $\mathbf{A}$-ortogonálisak (ortogonálisak az $(\overline{\mathbf{x}}, \overline{\mathbf{y}}) = \overline{\mathbf{x}}^T \mathbf{A} \overline{\mathbf{y}}$ skaláris szorzatban) az előző keresési irányokra. Ezt az alábbi algoritmus valósítja meg.

$$
\begin{aligned}
&k := 0, \overline{\mathbf{r}}_0 := \overline{\mathbf{b}}, \overline{\mathbf{x}}_0 := \mathbf{0}, \overline{\mathbf{p}}_1 = \overline{\mathbf{r}}_0 \\
&\textbf{while } \overline{\mathbf{r}}_k \neq \mathbf{0} \\
&\quad k := k + 1 \\
&\quad \alpha_k := \overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1} / (\overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{p}}_k) \\
&\quad \overline{\mathbf{x}}_k := \overline{\mathbf{x}}_{k-1} + \alpha_k \overline{\mathbf{p}}_k \\
&\quad \overline{\mathbf{r}}_k := \overline{\mathbf{r}}_{k-1} - \alpha_k \mathbf{A} \overline{\mathbf{p}}_k \\
&\quad \beta'_k := \overline{\mathbf{r}}_k^T \overline{\mathbf{r}}_k / (\overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1}) \\
&\quad \overline{\mathbf{p}}_{k+1} := \overline{\mathbf{r}}_k + \beta'_k \overline{\mathbf{p}}_k \\
&\textbf{end while}
\end{aligned}
$$

Az első k maradékvektor, az első k irányvektor, és az első k iterációs vektor ugyanazt az alterét feszíti ki $\mathbb{R}^n$-nek. Ezt az alteret V_k-val jelöljük. Legyen $\|\overline{\mathbf{x}}\|_{\mathbf{A}} = \sqrt{\overline{\mathbf{x}}^T \mathbf{A} \overline{\mathbf{x}}}$ és $\overline{\mathbf{e}}^{(k)} = \overline{\mathbf{x}}^\star - \overline{\mathbf{x}}_k$.

3.14. Tétel ***(A konjugált gradiens-módszer lépésenkénti optimális tulajdonsága.)*** *Ha* $\overline{\boldsymbol{r}}_{k-1} \neq \boldsymbol{0}$*, akkor* $\overline{\boldsymbol{x}}_k$ *az egyetlen pont* V_k*-ban, melyre* $\|\overline{\boldsymbol{e}}^{(k)}\|_{\boldsymbol{A}}$ *minimális,*

$$\|\overline{\boldsymbol{e}}^{(1)}\|_{\boldsymbol{A}} \geq \|\overline{\boldsymbol{e}}^{(2)}\|_{\boldsymbol{A}} \geq \cdots \geq \|\overline{\boldsymbol{e}}^{(k)}\|_{\boldsymbol{A}},$$

továbbá $\overline{\boldsymbol{e}}^{(k)} = \boldsymbol{0}$ *valamilyen* $k \leq n$ *esetén.*

3.15. Tétel ***(A konjugált gradiens-módszer konvergenciasebessége.)*** *Legyen* $\boldsymbol{A}$ *szimmetrikus, pozitív definit mátrix, melynek kondíciószáma* $\kappa_2(\boldsymbol{A})$*. Ekkor a konjugált gradiens-módszer hibavektorára az alábbi becslés érvényes*

$$\|\overline{\boldsymbol{e}}^{(k)}\|_{\boldsymbol{A}} \leq 2\left(\frac{\sqrt{\kappa_2(\boldsymbol{A})}-1}{\sqrt{\kappa_2(\boldsymbol{A})}+1}\right)^k \|\overline{\boldsymbol{e}}^{(0)}\|_{\boldsymbol{A}}.$$

3.1.4. Túlhatározott lineáris egyenletrendszerek megoldása

Az

$$\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}, \quad \mathbf{A} \in \mathbb{R}^{m\times n}, \ m \geq n, \ r(\mathbf{A}) = n$$

alakú egyenletrendszereket vizsgáljuk. Ezek $\overline{\mathbf{x}}_{LS}$ legkisebb négyzetek értelemben legjobb megoldásán azt az egyértelműen meghatározott vektor értjük, melyre $\|\overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}\|_2^2$ minimális.

3.16. Tétel *Az* $\overline{\boldsymbol{x}}_{LS}$ *megoldást meghatározhatjuk az*

$$\boldsymbol{A}^T\boldsymbol{A}\overline{\boldsymbol{x}} = \boldsymbol{A}^T\overline{\boldsymbol{b}}$$

normálegyenlet megoldásával vagy pedig az

$$\boldsymbol{R}_1\overline{\boldsymbol{x}} = \overline{\boldsymbol{c}}$$

egyenlet megoldásával, ahol $\boldsymbol{R}_1$ *az* $\boldsymbol{A}$ *mátrix QR-felbontásában szereplő* $\boldsymbol{R}$ *mátrix felső* $n\times n$*-es része, míg* $\overline{\boldsymbol{c}}$ *a* $\boldsymbol{Q}^T\overline{\boldsymbol{b}}$ *(*$\boldsymbol{Q}$ *a QR-felbontás* $\boldsymbol{Q}$ *mátrixa) vektor felső* n *elemét tartalmazó oszlopvektor.*

3.2. Feladatok

3.2.1. Kondicionáltság

3.1. Adjunk becslést az 1.47. feladat eredményét felhasználva az

$$\mathbf{A} = \begin{bmatrix} 1.01 & 1 \\ 1 & 1 \end{bmatrix}$$

mátrix maximumnormabeli kondíciószámára! Ezek után számítsuk is ki pontosan a kondíciószámot! $\longrightarrow \Longrightarrow$

3.2. Határozzuk meg az alábbi $\mathbf{A}$ mátrix kondíciószámát 1-es, 2-es és maximumnormában! Tekintsük az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszert, ahol $\overline{\mathbf{b}}$ egy adott pozitív vektor! Adjunk felső becslést maximumnormában a $\overline{\mathbf{b}}$ vektor maximumnormájának segítségével arra, hogy ezen egyenletrendszer megoldásától mennyire térhet el azon egyenletrendszer megoldása, melyben a $\overline{\mathbf{b}}$ vektor minden elemét 1%-kal megnöveljük ($\mathbf{A}$ változatlan marad)!

$$\mathbf{A} = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{bmatrix}$$

$\longrightarrow \Longrightarrow$

3.3. Igaz-e az az állítás, hogy egy invertálható valós mátrix pontosan akkor ortogonális, ha 2-es normabeli kondíciószáma 1? Válaszunkat részletesen indokoljuk! $\longrightarrow \Longrightarrow$

3.4. Tegyük fel, hogy az $\mathbf{A}\overline{\mathbf{v}} = \overline{\mathbf{b}}$ egyenletrendszer helyett ($\mathbf{A}$ invertálható mátrix) az $(1+c)\mathbf{A}\overline{\mathbf{u}} = \overline{\mathbf{b}}$ egyenletrendszert oldjuk meg, ahol c valamilyen valós, -1-től különböző paraméter! Számítsuk ki tetszőleges indukált normában a második egyenlet megoldásának az első egyenlet megoldásához viszonyított relatív hibáját a c paraméter függvényében! $\Longrightarrow$

3.5. Igazoljuk, hogy ha egy $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszer jobb oldalához hozzáadunk egy $\delta\overline{\mathbf{b}}$ vektort, akkor az új egyenletrendszer $\overline{\mathbf{x}}^\star$ megoldásával igaz lesz az

$$\|\overline{\mathbf{x}}^\star - \overline{\mathbf{x}}\| \leq \|\mathbf{A}^{-1}\| \cdot \|\delta\overline{\mathbf{b}}\|$$

becslés, ahol a szereplő mátrixnormát a szereplő vektornorma indukálja! Ez alapján adjunk becslést arra, hogy ha az

$$\begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix} \overline{\mathbf{x}} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$$

lineáris egyenletrendszer jobb oldalán álló vektor elemeihez rendre olyan $\varepsilon_1, \varepsilon_2$ számokat adunk, melyekre $|\varepsilon_1|, |\varepsilon_2| \leq 10^{-4}$, akkor maximum mekkorát változhat az egyenletrendszer megoldása 2-es normában! $\Longrightarrow$

3.6. Tekintsük az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszert, ahol

$$\mathbf{A} = \begin{bmatrix} 34 & 55 \\ 55 & 89 \end{bmatrix}, \quad \overline{\mathbf{b}} = \begin{bmatrix} 21 \\ 34 \end{bmatrix}.$$

Az $\overline{\mathbf{r}} = \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}$ maradékvektort az $\overline{\mathbf{x}} = [-0.11, 0.45]^T$ vektorral kiszámítva $\overline{\mathbf{r}} = [-0.01, 0]^T$, míg az $\overline{\mathbf{x}} = [-0.99, 1.01]^T$ vektorral $\overline{\mathbf{r}} = [-0.89, -1.44]^T$. A megoldás melyik $\overline{\mathbf{x}}$ közelítése pontosabb? Adjunk alsó és felső becslést egy $\overline{\mathbf{x}}$ közelítés megoldástól való eltérésére a maradékvektor segítségével! Ellenőrizzük a becslést az adott egyenletrendszeren! $\Longrightarrow$

3.7. Ismert, hogy egy mátrix spektrálsugara becsülhető a mátrix tetszőleges indukált normájával. Igazoljuk ennek segítségével, hogy tetszőleges $\mathbf{A}$ mátrixra $\|\mathbf{A}\|_2^2 \leq \|\mathbf{A}\|_1\|\mathbf{A}\|_\infty$ és hogy tetszőleges invertálható $\mathbf{A}$ mátrix esetén

$$\kappa_2(\mathbf{A}) \leq \sqrt{\kappa_1(\mathbf{A})\kappa_\infty(\mathbf{A})}\ !$$

$\Longrightarrow$

3.8. Igazoljuk, hogy tetszőleges reguláris $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix esetén

$$\frac{1}{n}\kappa_2(\mathbf{A}) \leq \kappa_1(\mathbf{A}) \leq n\kappa_2(\mathbf{A}), \quad \frac{1}{n}\kappa_\infty(\mathbf{A}) \leq \kappa_2(\mathbf{A}) \leq n\kappa_\infty(\mathbf{A}),$$

$$\frac{1}{n^2}\kappa_1(\mathbf{A}) \leq \kappa_\infty(\mathbf{A}) \leq n^2\kappa_1(\mathbf{A})!$$

$\Longrightarrow$

3.9. Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy olyan kvadratikus mátrix, melyben a főátló „felett" -1-esek, „alatta" nullák és a főátlóban 1-esek állnak! Számítsuk ki a mátrix determinánsát és a kondíciószámát maximumnormában! $\Longrightarrow$

3.10. Igazoljuk, hogy reguláris $\mathbf{A}$ mátrixra $\kappa_2(\mathbf{A}^T\mathbf{A}) = \kappa_2^2(\mathbf{A}) \geq \kappa_2(\mathbf{A})$! $\longrightarrow$ $\Longrightarrow$

3.11. Igazoljuk, hogy ha $\mathbf{A}$ és $\mathbf{B}$ ortogonálisan hasonló reguláris mátrixok, akkor $\|\mathbf{A}\|_2 = \|\mathbf{B}\|_2$ és $\kappa_2(\mathbf{A}) = \kappa_2(\mathbf{B})$! $\Longrightarrow$

3.2.2. Direkt módszerek

3.12. Az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszer $\mathbf{A}$ mátrixának és $\overline{\mathbf{b}}$ jobb oldali vektorának elemei mért mennyiségek, melyek relatív hibája 0.01%. Adjunk felső becslést a megoldásvektor relatív hibájára maximumnormában!

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 1.5 & 0 & -0.5 \\ -1 & 0 & -1.7 & -0.2 \\ 0 & -0.5 & -0.2 & 1.7 \end{bmatrix}, \quad \overline{\mathbf{b}} = \begin{bmatrix} 0 \\ 0 \\ 3 \\ 0 \end{bmatrix} \quad \mathbf{A}^{-1} \begin{bmatrix} 0.53 & 0.38 & -0.32 & 0.07 \\ 0.38 & 1.01 & -0.25 & 0.27 \\ -0.32 & -0.25 & -0.39 & -0.12 \\ 0.07 & 0.27 & -0.12 & 0.65 \end{bmatrix}$$

$\Longrightarrow$

3.13. A

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

lineáris egyenletrendszer jobb oldali vektora mérési eredményeket tartalmaz. Mekkora az egyenletrendszer megoldásának relatív hibája maximumnormában, ha tudjuk, hogy a pontos értékek a szereplő értékek 0.1 sugarú környezetében vannak valahol és az együtthatómátrix inverze

$$\begin{bmatrix} 0.0294 & 0.2176 & -0.0353 \\ 0.1765 & -0.0941 & -0.0118 \\ 0.0588 & 0.0353 & 0.1294 \end{bmatrix}?$$

$\longrightarrow \Longrightarrow$

3.14. Oldjuk meg az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszert a Gauss-módszer segítségével a lenti adatokkal! Adjuk meg az együtthatómátrix determinánsát és LU-felbontását is!

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{bmatrix}, \quad \overline{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}, \quad \overline{\mathbf{b}} = \begin{bmatrix} 2 \\ 10 \\ 44 \\ 190 \end{bmatrix}$$

$\Longrightarrow$

3.15. Tekintsük az alábbi mátrixot

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ -1 & 3 & 0 \\ 0 & -1 & 3 \end{bmatrix}!$$

Határozzuk meg a mátrix LU-felbontását! $\Longrightarrow$

3.16. Tekintsük azt az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrixot, melyre $a_{ij} = 1$ ha $i = j$ vagy $j = n$, $a_{ij} = -1$, ha $i > j$, különben nulla. Mutassuk meg, hogy $\mathbf{A}$-nak van LU-felbontása, $|l_{ij}| \leq 1$ és $u_{nn} = 2^{n-1}$. Számítsuk ki a növekedési faktort! $\Longrightarrow$

3.17. Tekintsünk egy olyan lineáris egyenletrendszert, melynek mátrixában csak az első oszlopban, az első sorban ill. a főátlóban vannak nemnulla elemek. Mi történik a mátrixszal a Gauss-módszer alkalmazása során? Adjunk javaslatot a jelenség elkerülésére! $\Longrightarrow$

3.18. Az alábbi mátrix egy $\mathbf{A} \in \mathbb{R}^{4\times 4}$ szimmetrikus mátrix LU-felbontását tartalmazza úgy, hogy a főátló „alatti" rész az $\mathbf{L}$ mátrix megfelelő főátló alatti részét tartalmazza, a többi elem pedig az $\mathbf{U}$ mátrix megfelelő eleme. Létezik-e az $\mathbf{A}$ mátrixnak Cholesky-felbontása? Ha igen, akkor adjuk meg a $\mathbf{G}$ mátrixot! Adjuk meg azt az $\overline{\mathbf{x}} \in \mathbb{R}^4$ vektort, melyre $\mathbf{A}\overline{\mathbf{x}} = [1, 0, 0, 0]^T$!

$$\begin{bmatrix} 2 & 3 & 2 & 4 \\ 3/2 & 3/2 & 2 & 3 \\ 1 & 4/3 & 7/3 & 3 \\ 2 & 2 & 9/7 & 1/7 \end{bmatrix}$$

$\Longrightarrow$

3.19. $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszer megoldására a Gauss–Jordan-eliminációs módszert alkalmazzuk (nemcsak "lefelé", hanem "felfelé" is eliminálunk az oszlopokat). Adjuk meg, hogy pontosan hány lebegőpontos műveletet igényel a megoldás! $\longrightarrow$ $\Longrightarrow$

3.20. Gondoljuk végig, hogy milyen módszerekkel lehetne egy mátrix inverzét kiszámolni, és adjuk meg a módszerek műveletszámát! $\Longrightarrow$

3.21. Határozzuk meg az alábbi $\mathbf{B}$ mátrix $\mathbf{LDM}^T$ felbontását, ahol $\mathbf{L}$ és $\mathbf{M}$ normált alsó háromszögmátrixok és $\mathbf{D}$ diagonális mátrix!

$$\mathbf{B} = \begin{bmatrix} 1 & -2 & 1 \\ 2 & -2 & -4 \\ 2 & 2 & -13 \end{bmatrix}$$

$\Longrightarrow$

3.22. Határozzuk meg az alábbi $\mathbf{B}$ mátrix LDL^T és Cholesky-felbontásait!

$$\mathbf{B} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$\Longrightarrow$

3.23. Adjuk meg az alábbi mátrixok Cholesky-felbontását!

$$\mathbf{B}_1 = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 3 \end{bmatrix}, \quad \mathbf{B}_2 = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}$$

$\Longrightarrow$

3.24. Adjuk meg az alábbi mátrix LU- és Cholesky-felbontásait!

$$\begin{bmatrix} 6 & 4 & 4 \\ 4 & 12 & 8 \\ 4 & 8 & 6 \end{bmatrix}$$

$\Longrightarrow$

3.25. Oldjuk meg az egyenletrendszert a Gauss-módszerrel teljes főelemkiválasztással és anélkül négyjegyű mantisszát használva! Mekkora a két megoldás eltérése maximum-normában?

$$\begin{aligned} 0.003x_1 + 59.14x_2 &= 59.17 \\ 5.291x_1 - 6.13x_2 &= 46.78 \end{aligned}$$

$\Longrightarrow$

3.26. Oldjuk meg az alábbi egyenletrendszert a Gauss-módszerrel részleges főelemkiválasztást alkalmazva egy olyan számítógépen, amely a lebegőpontos ábrázolás során tízes számrendszerben hatjegyű mantisszát használ és a karakterisztikára nincs megkötés!

$$\begin{bmatrix} 0.00001 & 2 & 3 \\ 1 & 2 & 3 \\ 10 & 3 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 5.00001 \\ 6 \\ 17 \end{bmatrix}$$

$\Longrightarrow$

3.27. Adjunk meg egy olyan Householder-féle tükrözési mátrixot, amellyel az $[2, 1, 2]^T$ vektort az $\overline{\mathbf{e}}_1$ vektor számszorosába lehet transzformálni! $\Longrightarrow$

3.28. Adjuk meg Householder-tükrözések segítségével az alábbi mátrix QR-felbontását!

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$$

$\Longrightarrow$

3.29. Adjuk meg a

$$\begin{bmatrix} 4 & 2 & 1 \\ 0 & 3 & 0 \\ 0 & 4 & 3 \end{bmatrix}$$

mátrix (egy) QR-felbontását! $\Longrightarrow$

3.30. Alakítsuk két Givens-forgatás segítségével felső háromszögmátrixszá az

$$\mathbf{A} = \begin{bmatrix} 1/\sqrt{2} & 0 \\ 0 & \sqrt{2} \\ 1/\sqrt{2} & \sqrt{2} \end{bmatrix}$$

mátrixot! $\Longrightarrow$

3.31. Az $n \times n$-es Householder-féle tükrözési mátrixot egyértelműen meghatározza a tükrözési sík $\overline{\mathbf{v}} \in \mathbb{R}^n$ normálvektora. Ha osztjuk ezt a vektort az első elemével (első elemre normáljuk), akkor a vektor egy $n-1$ elemű vektor helyén eltárolható, hiszen az 1-es első elemet nem kell tárolni. Az alábbi mátrix egy $\mathbf{A}$ mátrix QR-felbontását tartalmazza. A főátló és a felette lévő rész az $\mathbf{R}$ mátrix megfelelő elemeit tartalmazza, a főátló alatt az oszlopokban elhelyezkedő elemek a QR-felbontáshoz használt Householder-féle tükrözési mátrixok első elemre normált $\overline{\mathbf{v}}$ vektorainak maradék elemei. Adjuk meg az $\mathbf{A}$ mátrixot!

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

$\Longrightarrow$

3.32. Igazoljuk, hogy ha $\mathbf{A}$ egy nemszinguláris négyzetes mátrix és $\mathbf{Q}_1\mathbf{R}_1$ és $\mathbf{Q}_2\mathbf{R}_2$ két különböző QR-felbontása $\mathbf{A}$-nak, akkor van olyan $\mathbf{D}$ diagonális mátrix, melyre $\mathbf{D}^2 = \mathbf{E}$ és $\mathbf{R}_2 = \mathbf{D}\mathbf{R}_1$ és $\mathbf{Q}_2 = \mathbf{Q}_1\mathbf{D}$! Igazoljuk, hogy ha $\mathbf{R}$-ről feltesszük, hogy a főátlójában pozitív elemek állnak, akkor a mátrix QR-felbontása egyértelmű! $\longrightarrow \Longrightarrow$

3.33. Ha egy felső Hessenberg-mátrixra alkalmazzuk a Gauss-módszert, akkor figyelembe vehetjük, hogy a főátló „alatt" csak a közvetlenül a főátló alatti elemek különböznek nullától. Mekkora lesz az ilyen mátrixok LU-felbontásának műveletszáma? Mit mondhatunk az $\mathbf{L}$ és $\mathbf{U}$ mátrixok szerkezetéről? Ha már a mátrix LU-felbontása elkészült, akkor mennyi műveletbe kerül egy egyenletrendszer megoldása? $\Longrightarrow$

3.2.3. Iterációs módszerek

Klasszikus iterációs módszerek

3.34. Egy olyan lineáris egyenletrendszert szeretnénk megoldani a relaxált Gauss–Seidel-módszerrel, melynek együtthatómátrixa

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ -1 & 3 & 0 \\ 0 & -1 & 3 \end{bmatrix}.$$

Hogyan válasszuk ω értékét, hogy a leggyorsabban konvergáljon az eljárás? Mekkorának választhatjuk ω értékét egyáltalán, hogy konvergáljon a módszer? $\longrightarrow \Longrightarrow$

3.35. A

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

lineáris egyenletrendszert szeretnénk megoldani a Jacobi-iterációval. Végezzünk el két iterációs lépést a nullvektorról indulva, és becsüljük meg, hogy hány iterációs lépés lenne szükséges ahhoz, hogy a kapott közelítésnek a maximumnormabeli eltérése a pontos megoldástól 10^{-6}-nál kisebb legyen! $\longrightarrow \Longrightarrow$

3.36. Legyen $\mathbf{A} = \operatorname{tridiag}[-1, 2, -1] \in \mathbb{R}^{n\times n}$, azaz $\mathbf{A}$ egy olyan négyzetes mátrix, melynek főátlójában 2-esek, a szub- és szuperdiagonálisban -1-esek állnak. A többi elem nulla. Tegyük fel, hogy az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszert Jacobi-módszerrel szeretnénk megoldani. Határozzuk meg a Jacobi-módszer iterációs mátrixának spektrálsugarát, ha tudjuk, hogy az $\mathbf{A}$ mátrix sajátértékei

$$\lambda_k = 2\left(1 - \cos\frac{k\pi}{n+1}\right), \quad k = 1, \ldots, n\ !$$

Mit mondhatunk a módszer konvergenciájáról? $\Longrightarrow$

3.37. A Jacobi- vagy a Gauss–Seidel-iteráció konvergál gyorsabban az alábbi egyenletrendszerre?

$$\begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix} \overline{\mathbf{x}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Adjunk felső becslést arra, hogy hány iterációs lépést kellene elvégeznünk a gyorsabb módszerrel a $[0,0]^T$ kezdővektorral indulva, hogy a megoldást 10^{-6}-nál jobban megközelítse a sorozat határértékét 2-es normában! $\Longrightarrow$

3.38. Döntsük el, hogy az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszer megoldására használt Jacobi- ill. Gauss–Seidel-módszerek közül melyik lesz konvergens, ha

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix}!$$

$\Longrightarrow$

3.39. Döntsük el, hogy az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszer megoldására használt Jacobi- ill. Gauss–Seidel-módszerek közül melyik lesz konvergens, ha

$$\mathbf{A} = \begin{bmatrix} 1 & 1/2 & 1 \\ 1/2 & 1 & 1 \\ -2 & 2 & 1 \end{bmatrix}!$$

$\Longrightarrow$

3.40. Igazoljuk, hogy a $-4x_1 + 5x_2 = 1$, $x_1 + 2x_2 = 3$ lineáris egyenletrendszerre a Gauss–Seidel-módszer konvergálni fog (a megoldáshoz) tetszőleges kezdeti vektor esetén! Végezzünk el egy iterációs lépést a nullvektort választva kezdővektornak! $\Longrightarrow$

3.41. (⊞) Legyen $x_0 = 1$ és $x_{20} = 0$ és

$$x_k = \frac{3}{4}x_{k-1} + \frac{1}{4}x_{k+1},\ k = 1, \ldots, 19.$$

Igazoljuk, hogy az egyenletrendszer megoldása $x_k = 1 - (3^k - 1)/(3^{20} - 1)$! Oldjuk meg az egyenletrendszert Gauss–Seidel-módszerrel! Mit tapasztalunk, javítja-e a konvergenciát az alul- vagy a túlrelaxálás? $\Longrightarrow$

3.42. Az alábbi egyenletrendszert szeretnénk megoldani a Jacobi-módszer relaxálásával. Hogyan válasszuk meg ω értékét, hogy a leggyorsabban konvergáljon az eljárás? Számítsuk ki, hogy a nullvektorról indulva a leggyorsabb módszerrel mennyit kellene iterálni, hogy a megoldást 10^{-6}-nál jobban megközelítsük maximumnormában!

$$\begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$\Longrightarrow$

3.43. Javasoljunk az alábbi egyenletrendszer iterációs megoldására egy alkalmas eljárást! Igazoljuk is a módszer konvergenciáját! Hajtsunk végre egy iterációs lépést vele az $\overline{\mathbf{x}}^{(0)} = [1, 0, 0]^T$ kezdővektorról indulva! Mennyit kellene lépni a módszerrel, ha a megoldásvektort maximumnormában 10^{-6}-nál jobban meg szeretnénk közelíteni?

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

$\longrightarrow \Longrightarrow$

3.44. Az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszert az

$$\overline{\mathbf{x}}^{(k+1)} = (\mathbf{E} - \omega\mathbf{A})\overline{\mathbf{x}}^{(k)} + \omega\overline{\mathbf{b}}$$

iterációval szeretnénk megoldani tetszőleges $\overline{\mathbf{x}}^{(0)}$ vektorról indulva (ω tetszőleges pozitív konstans). Tegyük fel, hogy $\mathbf{A}$ összes sajátértéke valós és az $[\alpha, \beta]$ intervallumba esik, ahol $0 < \alpha \leq \beta$! Adjunk javaslatot ω megválasztására! $\Longrightarrow$

3.45. Az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszert szeretnénk megoldani az $\overline{\mathbf{x}}^{(k+1)} = \overline{\mathbf{x}}^{(k)} + \alpha(\mathbf{A}\overline{\mathbf{x}}^{(k)} - \overline{\mathbf{b}})$ iterációval, ahol

$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix}, \quad \overline{\mathbf{b}} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ és } \overline{\mathbf{x}}^{(0)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Adjuk meg α optimális értékét! $\Longrightarrow$

Variációs módszerek

3.46. Végezzünk el egy lépést a gradiens módszerrel a nullvektorról indulva a

$$\begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

egyenletrendszerből származtatott normálegyenletre! $\Longrightarrow$

3.47. Oldjuk meg a

$$\begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} \overline{\mathbf{x}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

egyenletrendszert a gradiens módszer segítségével! Végezzünk el két lépést a módszerrel a nullvektorról indulva! $\Longrightarrow$

3.48. A konjugált gradiens módszert alkalmazzuk a $\operatorname{tridiag}[-1, 2, -1]\overline{\mathbf{x}} = [1, 0, 1]^T$ egyenletrendszer megoldására. Számítsuk ki az $\overline{\mathbf{x}}_2$ vektort, majd számítsuk ki a hozzá tartozó maradékvektort! Mit tapasztalunk? $\Longrightarrow$

3.49. Oldjuk meg a

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

lineáris egyenletrendszert a konjugált gradiens módszer segítségével! $\Longrightarrow$

3.50. Oldjuk meg a

$$\begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} \overline{\mathbf{x}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

egyenletrendszert a konjugált gradiens módszer segítségével! $\Longrightarrow$

3.51. Tegyük fel, hogy a $(konj)grad(A, b, toll, nmax)$ program a (konjugált) gradiens módszert hajtja végre a nullvektorról indulva az A mátrixú, b jobboldalú egyenletrendszerre. A program akkor áll le, ha a maradékvektor normája kisebb, mint a *toll* toleranciaszint, vagy akkor, ha az iterációszám elérte az $nmax$ értéket. A program kimeneti értéke az aktuális lépés x vektora. Hogyan alkalmazzuk a programot, ha az iterációt egy adott y vektortól szeretnénk indítani? $\Longrightarrow$

3.52. (⊞) Oldjuk meg a

$$\text{tridiag}\,(-1, 2, -1)\overline{\mathbf{x}} = \overline{\mathbf{e}}$$

egyenletrendszert a konjugált gradiens-módszer segítségével, ahol a mátrix 20×20-as méretű! Hány iteráció kellett a megoldáshoz? $\Longrightarrow$

3.53. (⊞) Oldjuk meg a gradiens és a konjugált gradiens módszerrel is a

$$\begin{aligned} 10x - 2y + 3z + u &= 3 \\ -2x + 10y - 2z - u &= -4 \\ 3x - 2y + 10z + 5u &= 7 \\ x - y + 5z + 30u &= 8 \end{aligned}$$

egyenletrendszert! $\Longrightarrow$

3.2.4. Túlhatározott lineáris egyenletrendszerek megoldása

3.54. Adjuk meg az alábbi mátrix QR-felbontását Householder-tükrözések segítségével, majd adjuk meg a QR-felbontást alkalmazva az $\mathbf{A}\overline{\mathbf{x}} = [1, 1, 1]^T$ túlhatározott egyenletrendszer $\overline{\mathbf{x}}_{LS}$ megoldását!

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 1 & 3 \\ 0 & 2 \end{bmatrix}$$

$\Longrightarrow$

3.55. Adjuk meg a 3.54. feladatban szereplő túlhatározott egyenletrendszer $\overline{\mathbf{x}}_{LS}$ megoldását a normálegyenlet segítségével! $\Longrightarrow$

3.56. Adjuk meg az alábbi túlhatározott lineáris egyenletrendszer $\overline{\mathbf{x}}_{LS}$ megoldását!

$$\begin{bmatrix} 0 & 0 & 2 \\ 1 & 3 & 1 \\ 0 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$\Longrightarrow$

3.57. Adjuk meg az alábbi túlhatározott lineáris egyenletrendszer $\overline{\mathbf{a}}_{LS}$ megoldását! Mit ad meg a kapott $\overline{\mathbf{a}}_{LS}$ vektor?

$$\begin{bmatrix} 1 & 1 & 1^2 \\ 1 & 2 & 2^2 \\ 1 & 3 & 3^2 \\ 1 & 4 & 4^2 \\ 1 & 5 & 5^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ -1 \\ 4 \\ 3 \end{bmatrix}$$

$\Longrightarrow$

3.58. Két mennyiséget (x és y) mértünk ill. ezek különbségét és összegét. Az eredmények: $x = a$, $y = b$, $x - y = c$ és $x + y = d$. Oldjuk meg ezt a túlhatározott egyenletrendszert! $\Longrightarrow$

3.59. (⊞) Oldjuk meg az

$$\begin{bmatrix} 1 & 1 \\ 10^{-k} & 0 \\ 0 & 10^{-k} \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 10^{-k} \\ 1 + 10^{-k} \\ 1 - 10^{-k} \end{bmatrix}$$

túlhatározott egyenletrendszert $k = 6, 7, 8$ esetén először papíron számolva, majd az $A\backslash b$ (QR-felbontást használja) és az $(A' * A)\backslash(A' * b)$ MATLAB-beli utasításokkal (Cholesky-felbontásos megoldás)! Hasonlítsuk össze az eredményt! $\Longrightarrow$

4. fejezet

Sajátérték-feladatok numerikus megoldása

4.1. Képletek, összefüggések

Sajátértékfeladatok esetén négyzetes mátrixok sajátértékeit és a hozzájuk tartozó sajátvektorokat határozzuk meg. A mátrixok sajátértékeinek lokalizációját segíti az alábbi tétel.

4.1. Tétel ***(Gersgorin-tétel a sajátértékek elhelyezkedéséről.)*** *Tekintsük az* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *mátrixot. Legyen* K_i *a komplex számsíkon az a zárt körlap, melynek középpontja* a_{ii}*, és sugara* $\sum_{j=1,j\neq i}^{n} |a_{ij}|$ *(*$i = 1, \ldots, n$*). Ekkor a mátrix sajátértékei az* $\cup_{i=1,\ldots,n} K_i$ *halmazban találhatók. Ha* s *darab körlap diszjunkt a többitől, akkor uniójukban pontosan* s *darab sajátérték található.*

A Bauer–Fike-tétel ad becslést arra, hogy egy mátrix sajátértékei mennyit változnak akkor, ha elemeit egy kicsit megváltoztatjuk.

4.2. Tétel ***(Bauer–Fike-tétel a sajátérték-feladatok kondicionáltságáról.)*** *Legyen* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *egy diagonalizálható mátrix (*$\boldsymbol{A}$ *felírható* $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1}$ *alakban), továbbá* $\delta\boldsymbol{A}$ *egy tetszőleges mátrix, és legyen* μ *az* $\boldsymbol{A}+\boldsymbol{\delta A}$ *mátrix egy sajátértéke. Ekkor tetszőleges* p*-normában igaz, hogy*

$$\min_{\lambda \ \boldsymbol{A} \text{ sajátértéke}} |\lambda - \mu| \leq \kappa_p(\boldsymbol{V}) \|\boldsymbol{\delta A}\|_p.$$

A sajátérték-feladatokat, néhány speciális esettől eltekintve, mindig iterációs módszerrel oldjuk meg. Az iterációs módszereket két nagy csoportra oszthatjuk: a sajátértékeket egyenként ill. egyszerre közelítő módszerekre.

A sajátértékeket egyenként közelítő módszerek alapmódszere a hatványmódszer:

4.3. Tétel ***(A hatványmódszer konvergenciája.)*** *Legyen* $\boldsymbol{A}$ *normális mátrix* λ_1 *egyszeresen domináns sajátértékkel és a hozzá tartozó* $\overline{\boldsymbol{v}}_1$ *normált sajátvektorral, és legyen* $\overline{\boldsymbol{y}}^{(0)}$ *olyan kezdővektor, melyre* $\overline{\boldsymbol{v}}_1^T\overline{\boldsymbol{y}}^{(0)} \neq 0$, $\|\overline{\boldsymbol{y}}^{(0)}\|_2 = 1$. *Ekkor az*

$$
\begin{aligned}
&\textbf{\textit{for}}\ k := 1 : k_{\max}\\
&\overline{\boldsymbol{x}}^{(k)} := \boldsymbol{A}\overline{\boldsymbol{y}}^{(k-1)}\\
&\overline{\boldsymbol{y}}^{(k)} := \overline{\boldsymbol{x}}^{(k)}/\|\overline{\boldsymbol{x}}^{(k)}\|_2\\
&\nu^{(k)} := (\overline{\boldsymbol{y}}^{(k)})^T\boldsymbol{A}\overline{\boldsymbol{y}}^{(k)}\\
&\textbf{\textit{end for}}
\end{aligned}
$$

algoritmussal meghatározott $\overline{\boldsymbol{y}}^{(k)}$ *vektorokra és* $\nu^{(k)}$ *számokra igaz, hogy*

$$\overline{\boldsymbol{y}}^{(k)} = \frac{\boldsymbol{A}^k\overline{\boldsymbol{y}}^{(0)}}{\|\boldsymbol{A}^k\overline{\boldsymbol{y}}^{(0)}\|_2},$$

$$\nu^{(k)} \to \lambda_1 \ (k \to \infty),$$

továbbá létezik olyan $\{\gamma_k\} \subset \mathbb{R}$ *sorozat, hogy* $|\gamma_k| = 1$ $(k = 1, \dots)$ *és*

$$\gamma_k\overline{\boldsymbol{y}}^{(k)} \to \overline{\boldsymbol{v}}_1.$$

Egy tetszőleges μ számhoz egyetlen legközelebbi sajátérték és a hozzá tartozó sajátvektor is meghatározható a hatványmódszer megfelelő módosításával.

4.4. Tétel ***(Az inverz iteráció konvergenciája.)*** *Ha* $\boldsymbol{A}$ *olyan normális mátrix, melynek pontosan egy legközelebbi sajátértéke* $(\lambda^\star)$ *van a* $\mu \in \mathbb{R}$ *számhoz, akkor a*

$$
\begin{aligned}
&\textbf{\textit{for}}\ k := 1 : k_{\max}\\
&(\boldsymbol{A} - \mu\boldsymbol{E})\overline{\boldsymbol{x}}^{(k)} = \overline{\boldsymbol{y}}^{(k-1)} \to \overline{\boldsymbol{x}}^{(k)}\\
&\overline{\boldsymbol{y}}^{(k)} := \overline{\boldsymbol{x}}^{(k)}/\|\overline{\boldsymbol{x}}^{(k)}\|_2\\
&\nu^{(k)} := (\overline{\boldsymbol{y}}^{(k)})^T\boldsymbol{A}\overline{\boldsymbol{y}}^{(k)}\\
&\textbf{\textit{end for}}
\end{aligned}
$$

iteráció $(\|\overline{\boldsymbol{y}}^{(0)}\|_2 = 1,\ (\overline{\boldsymbol{y}}^{(0)})^T\overline{\boldsymbol{v}}^\star \neq 0)$ *a hatványmódszernél ismertetett értelemben a* $\lambda^\star$ *sajátértéket és a hozzá tartozó* $\overline{\boldsymbol{v}}^\star$ *sajátvektort adja.*

Ha van egy közelítésünk egy sajátértékre, akkor annak értékére az inverz iterációval mondhatunk pontosabb közelítést, és a hozzá tartozó sajátvektort is megadhatjuk. Ha egy sajátvektorra van közelítésünk $(\overline{\mathbf{v}})$, akkor a sajátértékközelítést a

$$\frac{\overline{\mathbf{v}}^T\mathbf{A}\overline{\mathbf{v}}}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}$$

Rayleigh-hányadossal adhatjuk meg.

Most térjünk át a sajátértékeket egyszerre közelítő módszerekre!

4.5. Tétel ***(A Jacobi-módszer konvergenciája.)*** *Tekintsük az alábbi iterációt egy szimmetrikus $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ mátrix esetén (legyen $\boldsymbol{A}^{(0)} = \boldsymbol{A}$)!*

1. *Válasszuk ki a mátrix főátlója felett a legnagyobb abszolút értékű elemet! Legyen ez az a_{ij} elem ($i < j$).*

2. *Határozzuk meg az $s = \sin\theta$ és $c = \cos\theta$ értékeket az alábbi módon:*

   Ha $a_{ii} = a_{jj}$, akkor a $\cos(2\theta) = 0$, $s^2 + c^2 = 1$ egyenletrendszer megoldását kell meghatározni.

   Ha $a_{ii} \neq a_{jj}$, akkor $tg(2\theta) = -2a_{ij}/(a_{ii} - a_{jj})$. Ebből és az $s^2 + c^2 = 1$ összefüggésből s és c ismét meghatározható.

3. *Definiáljuk az alábbi $n \times n$-es ortogonális mátrixot:*

$$\boldsymbol{S}_{ij} = [\overline{\boldsymbol{e}}_1, \ldots, \overline{\boldsymbol{e}}_{i-1}, c\overline{\boldsymbol{e}}_i - s\overline{\boldsymbol{e}}_j, \overline{\boldsymbol{e}}_{i+1}, \ldots, \overline{\boldsymbol{e}}_{j-1}, s\overline{\boldsymbol{e}}_i + c\overline{\boldsymbol{e}}_j, \overline{\boldsymbol{e}}_{j+1}, \ldots, \overline{\boldsymbol{e}}_n],$$

   azaz

$$\boldsymbol{S}_{ij} = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & c & & & & s & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & -s & & & & c & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix}.$$

4. *Legyen $\boldsymbol{A}^{(k+1)} = \boldsymbol{S}_{ij}^T \boldsymbol{A}^{(k)} \boldsymbol{S}_{ij}$, majd lépjünk vissza a 1. ponthoz!*

Ebben az iterációban minden lépésben az $\boldsymbol{A}^{(k)}$ mátrix főátlón kívüli elemeinek négyzetösszege $2a_{ij}^2$-tel csökken, és így az $\boldsymbol{A}^{(k)}$ mátrixsorozat olyan diagonális mátrixhoz, melynek főátlójában az $\boldsymbol{A}$ mátrix sajátértékei vannak.

4.6. Tétel ***(QR-iteráció konvergenciája.)*** *Ha az $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ mátrixnak minden sajátértéke valós és abszolút értékben különböző ($|\lambda_1| > |\lambda_2| > |\lambda_3| > \cdots > |\lambda_n|$), akkor az*

$\boldsymbol{A}^{(0)} := \boldsymbol{A}$
for $k := 1 : k_{\max}$
$\boldsymbol{A}^{(k-1)} = \boldsymbol{Q}^{(k-1)} \boldsymbol{R}^{(k-1)}$ *(QR-felb.)*
$\boldsymbol{A}^{(k)} := \boldsymbol{R}^{(k-1)} \boldsymbol{Q}^{(k-1)}$
end for

iterációval előállított mátrixsorozatra

$$\lim_{k\to\infty} \boldsymbol{A}^{(k)} = \begin{bmatrix} \lambda_1 & \tilde{a}_{12} & \dots & \tilde{a}_{1n} \\ 0 & \lambda_2 & \tilde{a}_{23} & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_n \end{bmatrix}$$

valamilyen megfelelő $\tilde{a}_{ij}$ konstansokkal, azaz a határértékmátrix egy felső háromszögmátrix.

Ha $\boldsymbol{A}$ szimmetrikus, akkor $\{\boldsymbol{A}^{(k)}\}$ diagonális mátrixhoz tart.

4.2. Feladatok

4.2.1. Sajátértékbecslések

4.1. Adjunk becslést az

$$\mathbf{A} = \begin{bmatrix} 1 & 0.2 & -0.1 \\ 0.3 & 3 & 0.1 \\ 0.1 & 0.1 & -2 \end{bmatrix}$$

mátrix sajátértékeire! Igazoljuk, hogy minden sajátértéke valós a mátrixnak! ⟶ ⟹

4.2. Legyenek

$$\mathbf{A} = \begin{bmatrix} -2 & -1 & 2 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} -1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & -1 \end{bmatrix}.$$

Jelölje $\lambda_j(\varepsilon)$ $(j = 1, 2, 3)$ az $\mathbf{A}+\varepsilon\mathbf{B}$ mátrix sajátértékét! Adjunk becslést a $|\lambda_j(0)-\lambda_j(\varepsilon)|$ eltérésre! ($\mathbf{A}$-nak $[0, 2, 1]^T$ és $[1, -2, 0]^T$ sajátvektora rendre 1, 0 sajátértékkel.) ⟶ ⟹

4.3. Igazoljuk, hogy az

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & -1 & 0 \\ 0 & -2 & 0 & 1 \\ 0 & 0 & 3 & 1 \\ 1 & 1 & 2 & 5 \end{bmatrix}$$

mátrixnak pontosan egy negatív valós sajátértéke van! ⟶ ⟹

4.4. Igazoljuk, hogy az

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 2 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{bmatrix}$$

mátrixnak minden sajátértéke valós! ⟶ ⟹

4.5. (⊞) Adjunk becslést arra, hogy mennyit változnak az alábbi mátrixok sajátértékei, ha az első oszlopuk minden eleméhez 0.1-et adunk! Ellenőrizzük a becslés helyességét a MATLAB-ban számolva! Használjuk a Bauer–Fike-tételt (4.2. tétel)!

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 2 & 4 \\ 2 & 4 & 2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 2 & 4 \\ 0 & 0 & 1 \end{bmatrix}$$

$\Longrightarrow$

4.6. Adjuk meg, hogy az $\overline{\mathbf{x}} = [1, 1, 1]^T$ vektor hányszorosa van legközelebb euklideszi-normában az $\mathbf{A}\overline{\mathbf{x}}$ vektorhoz, ha

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 2 & 4 \\ 2 & 4 & 1 \end{bmatrix}!$$

$\longrightarrow \Longrightarrow$

4.7. A 4.6. feladat mátrixának egyik sajátvektora kb. $\overline{\mathbf{v}} = [-5, -6, -6]^T$. Adjunk becslést a vektorhoz tartozó sajátértékre! $\Longrightarrow$

4.2.2. Hatványmódszer és változatai

4.8. Alkalmazzuk a hatványmódszert az

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 2 & 2 \end{bmatrix}$$

mátrixra! Legyen a kezdővektor $\overline{\mathbf{x}}^{(0)} = [1, 0]^T$, és az iterációt a 4. lépés után leállítva adjunk becslést a domináns sajátértékre és egy hozzá tartozó sajátvektorra! $\Longrightarrow$

4.9. Jelölje $\lambda_1, \lambda_2, \lambda_3$ a $\mathbf{C}$ mátrix sajátértékeit növekvő sorrendben. Hatványmódszert hajtunk végre az $\mathbf{A} = \mathbf{C} - 10\mathbf{E}$ mátrixszal. A $\mathbf{C}$ mátrix melyik sajátvektora határozható meg az előállított $\overline{\mathbf{y}}^{(k)}$ vektorsorozattal? Hajtsunk végre egy iterációs lépést a $[2/3, 1/3, 2/3]^T$ vektorral, majd adjunk becslést az eredmény alapján a $\mathbf{C}$ mátrix megfelelő sajátértékére!

$$\mathbf{C} = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 5 & 1 \\ 1 & 1 & 10 \end{bmatrix}$$

$\Longrightarrow$

4.10. Határozzuk meg az $\mathbf{A}$ mátrix domináns sajátértékének egy közelítését úgy, hogy a hatványmódszer segítségével elvégzünk 4 iterációs lépést az $[1,1,1]^T$ vektorról indulva, majd utána a sajátértéket a Rayleigh-hányadossal becsüljük!

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

$\Longrightarrow$

4.11. Tekintsük az

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

mátrixot! Jelölje a mátrix legkisebb sajátértékét $\lambda_{\min}$. Igazoljuk, hogy $\lambda_{\min} = 4 - \varrho(\mathbf{A} - 4\mathbf{E})$! Adjunk becslést a $\lambda_{\min}$ értékre úgy, hogy az $\mathbf{A}-4\mathbf{E}$ mátrixra alkalmazzuk a hatványmódszert az $\overline{\mathbf{x}}^{(0)} = [1,1,1,1]^T$ kezdővektorról indulva, három iterációs lépést végrehajtva! $\Longrightarrow$

4.12. Egy $\mathbf{A}$ 4×4-es mátrixról tudjuk, hogy sajátértékei a 20, 10, 5 és 1 számok közelében vannak. Milyen α számmal alkalmazzuk a hatványmódszert az $\mathbf{A} - \alpha\mathbf{E}$ mátrixra, hogy az 1 közeli sajátértéket és a hozzá tartozó sajátvektort adja meg? $\longrightarrow \Longrightarrow$

4.13. (⊞) Határozzuk meg a 4.10. feladat mátrixának legnagyobb és legkisebb sajátértékét a hatványmódszer segítségével! $\Longrightarrow$

4.14. (⊟) Módosítsuk a powmeth.m programot úgy, hogy az inverz iterációt hajtsa végre, és az iterációt gyorsítsuk a mátrix LU-felbontásának kiszámolásával! $\Longrightarrow$

4.15. (⊞) A 6×6-os Hilbert-mátrixnak van egy sajátértéke $1/4$ közelében. Határozzuk meg ezt a sajátértéket és a hozzá tartozó sajátvektort az inverz iteráció alkalmazásával! $\Longrightarrow$

4.16. (⊞) Határozzuk meg az

$$\mathbf{A} = \begin{bmatrix} 5 & 9 & 8 & \dots & 1 \\ 9 & 5 & 9 & \ddots & 2 \\ 8 & 9 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 9 \\ 1 & 2 & \dots & 9 & 5 \end{bmatrix}$$

mátrix legnagyobb és legkisebb abszolút értékű sajátértékeit, ill. azt a sajátértéket, ami 15 körül van! $\Longrightarrow$

4.17. (⊞) Határozzuk meg az 5×5-ös Hilbert-mátrix legnagyobb és második legnagyobb sajátértékét rangcsökkentéssel! $\Longrightarrow$

4.18. (⊞) Oldjuk meg a 4.17. feladatot a Householder-féle deflációs eljárás segítségével! $\Longrightarrow$

4.2.3. Jacobi- és QR-iterációk

4.19. A Jacobi-módszernél az

$$\mathbf{A} = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

mátrixhoz keresnünk kell egy olyan

$$\mathbf{S} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

forgatási mátrixot, mellyel az $\mathbf{S}^T\mathbf{AS}$ mátrix diagonális lesz. A θ szögre teljesülnie kell, hogy $\cos(2\theta) = 0$ ($a = d$ esetén) vagy hogy $\operatorname{ctg}(2\theta) = (d-a)/(2b)$ ($a \neq d$ esetén), ill. a Pitagorasz-tételnek (4.5. tétel). Igazoljuk, hogy az $s := \sin\theta$ és $c := \cos\theta$ értékek megkaphatók a θ szög explicit kiszámítása nélkül is! Igazoljuk először, hogy olyan szögekre, melyekre a szereplő függvények értelmezve vannak, igaz a

$$\operatorname{tg}^2\theta + 2\operatorname{ctg}(2\theta)\cdot\operatorname{tg}\theta - 1 = 0$$

egyenlőség, majd adjuk meg ebből az s és c értékeket! $\Longrightarrow$

4.20. Adjuk meg az

$$\mathbf{A} = \begin{bmatrix} 2 & 4 \\ 4 & 2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix}$$

mátrixokhoz tartozó $\mathbf{S}$ Jacobi-transzformációs mátrixokat! $\Longrightarrow$

4.21. Végezzünk el két lépést az

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$$

mátrixszal a Jacobi-módszerrel (generáljuk a Jacobi-transzformációkat az első sor harmadik és a második sor harmadik elemeivel), majd adjunk becslést a mátrix sajátértékeire a kapott iterációs mátrix alapján! $\Longrightarrow$

4.22. Végezzünk el két lépést az

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & 2 & 2 \\ 2 & 3 & 2 & 2 \\ 2 & 2 & 3 & 2 \\ 2 & 2 & 2 & 3 \end{bmatrix}$$

mátrixszal a Jacobi-módszerrel (generáljuk a Jacobi-transzformációkat az első sor negyedik és a második sor negyedik elemeivel), majd adjunk becslést a mátrix sajátértékeire a kapott iterációs mátrix alapján! $\Longrightarrow$

4.23. (⊞) Végezzünk el a Jacobi-módszerrel (balról jobbra és fentről lefelé haladva a mátrix főátló feletti részén) annyi iterációs lépést az $\mathbf{A} = \text{tridiag}([-1, 2, -1]) \in \mathbb{R}^{5\times 5}$ mátrixszal, míg a mátrix főátlón kívüli részének Frobenius-normája az eredeti mátrix Frobenius-normájának 1/1000-e nem lesz! Adjunk becslést a mátrix sajátértékeire a kapott iterációs mátrix segítségével! $\Longrightarrow$

4.24. (⊞) Határozzuk meg annak a 10×10-es mátrixnak a sajátértékeit, melynek a főátlójában 4-esek állnak a többi elem pedig 1-es! $\Longrightarrow$

4.25. Alkalmazzuk a QR-iterációt az

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 2 & 2 \end{bmatrix}$$

mátrix sajátértékeinek meghatározására! Végezzünk el két iterációs lépést, és ez alapján adjunk becslést a sajátértékekre! $\Longrightarrow$

4.26. Határozzuk meg az

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix}$$

mátrix QR-felbontását valamelyik tanult módszer segítségével, és végezzünk el egy iterációs lépést a QR-iterációval! $\Longrightarrow$

4.27. (⊟) Írjunk MATLAB programot, amely a QR-iterációt hajtja végre az

```
[s,h]=qriter(A,nmax,toll)
```

paranccsal, ahol A az a mátrix, aminek a sajátértékeit meg szeretnénk határozni, nmax a maximális iterációszám, és a toll toleranciaszint egy olyan leállási feltételt ad, hogy ha az iterációs mátrix főátlón kívüli részének Frobenius-normája kisebb, mint az A mátrix Frobenius-normájának toll szorosa, akkor már leállíthatjuk az iterációt! A kimenő paraméterek az iterációs mátrix főátlójának elemei (ezek) a sajátértékbecslések és ezek hibája (a Gersgorin-tétel alapján). $\Longrightarrow$

4.28. Hozzunk létre egy olyan felső Hessenberg-mátrixot, melynek ugyanazok a sajátértékei, mint az alábbi $\mathbf{A}$ mátrixnak! Alkalmazzunk Householder-tükrözést a transzformációhoz!

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & 3 \\ 4 & 4 & 4 \\ 3 & 1 & 4 \end{bmatrix}$$

$\Longrightarrow$

4.29. Hozzunk létre egy olyan felső Hessenberg-mátrixot, melynek ugyanazok a sajátértékei, mint az alábbi $\mathbf{A}$ mátrixnak! Alkalmazzunk Householder-tükrözést a transzformációhoz! Milyen alakú lesz a transzformált mátrix azon túl, hogy felső Hessenberg?

$$\mathbf{A} = \begin{bmatrix} 4 & 4 & 3 \\ 4 & 4 & 4 \\ 3 & 4 & 4 \end{bmatrix}$$

$\Longrightarrow$

4.30. (⊞) MATLAB-ban számolva a részszámításokat, hozzunk létre egy olyan felső Hessenberg-mátrixot, melynek ugyanazok a sajátértékei, mint az alábbi $\mathbf{A}$ mátrixnak! Alkalmazzunk Householder-tükrözést a transzformációhoz!

$$\mathbf{A} = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

$\Longrightarrow$

4.31. (⊟) Módosítsuk a 4.27. feladatbeli programot úgy, hogy a program hozza az A mátrixot Hessenger-alakra a QR-iteráció megkezdése előtt! $\Longrightarrow$

4.32. (⊞) Adjuk meg a 4.30. feladatban szereplő mátrix sajátértékeit a 4.31. feladatbeli QR-iterációs program segítségével! Legyen a toleranciaszint 10^{-6}! $\Longrightarrow$

4.33. (⊞) Adjuk meg az $\mathbf{A} = \text{tridiag}(1, -2, 1) \in \mathbb{R}^{20\times 20}$ mátrix sajátértékeit a 4.31. feladatbeli QR-iterációs program segítségével! Legyen a toleranciaszint 10^{-8}! $\Longrightarrow$

5. fejezet

Nemlineáris egyenletek és egyenletrendszerek megoldása

5.1. Képletek, összefüggések

Az $f(x) = 0$ egyenlet megoldását keressük, ahol általában $f : \mathbb{R} \to \mathbb{R}$ folytonos függvény.

A megoldást először a zérushelyek elkülönítésével kezdjük, azaz megadunk olyan intervallumokat, amelyek tartalmazzák a zérushelyeket. Ebben segít az alábbi tétel.

5.1. Tétel ***(Elégséges feltétel zérushely létezésére egy intervallumban.)*** *Ha egy folytonos függvény esetén $f(a) \cdot f(b) < 0$ $(a < b)$, akkor van olyan $c \in (a, b)$, melyre $f(c) = 0$, sőt ha f szigorúan monoton, akkor pontosan egy ilyen zérushely van csak.*

Polinomok zérushelyeinek lokalizációját segíti az alábbi tétel.

5.2. Tétel ***(Polinomok zérushelyeinek lokalizációja.)*** *A $p(x) = a_n x^n + \ldots + a_1 x + a_0$ $(a_n, a_0 \neq 0)$ polinom zérushelyei az origó közepű $R = 1 + A/|a_n|$ és $r = 1/(1 + B/|a_0|)$ sugarak által meghatározott körgyűrűben vannak, ahol*

$$A = \max\{|a_{n-1}|, \ldots, |a_0|\}, \;\; B = \max\{|a_n|, \ldots, |a_1|\}.$$

Az alábbiakban felsoroljuk a legfontosabb nemlineáris egyenlet megoldó eljárásokat. A tételekben $k_{\max}$ mindig a maximális iterációszámot és *toll* a leállási feltételekben használt toleranciaszintet jelentik. Az $x^\star$ érték az f függvény egy zérushelye. Az algoritmus végrehajtása után k értékéből tudhatjuk, hogy elértük-e a kívánt pontosságot: ha $k < k_{\max}$, akkor a pontosságot elértük, ha $k = k_{\max}$, akkor amiatt állt le az algoritmus, mert elértük a maximális lépésszámot.

5.3. Tétel ***(Az intervallumfelezési módszer konvergenciája.)*** *Tegyük fel, hogy az f függvény folytonos az $[a,b]$ intervallumon és $f(a)\cdot f(b)<0$. Ekkor az*

$k = 0$
while $k < k_{\max}$ ***and*** $(b-a)/2 > toll$
$x := a + (b-a)/2$
if $f(x) = 0$ ***then***
end
else
if $f(x)\cdot f(a) > 0$ ***then***
$a = x$
else
$b = x$
end if
end if
$k := k + 1$
end while

algoritmus által szolgáltatott x_k sorozatra $x_k \to x^\star$, ahol $x^\star$ az f függvény egyik $[a,b]$-be eső zérushelye, továbbá igaz az

$$|x_k - x^\star| \le \frac{b-a}{2^{k+1}}$$

hibabecslés.

Azt mondjuk, hogy az f függvény *kielégíti az alapfeltevéseket* az $[a,b]$ intervallumon, ha van $[a,b]$ belsejében zérushelye, legalább kétszer folytonosan deriválható, és megfelelő pozitív konstansokkal $0 < m_1 \le |f'(x)| \le M_1 < \infty$ és $0 < m_2 \le |f''| \le M_2 < \infty$ is igaz minden $x \in [a,b]$ pontban.

5.4. Tétel ***(A húrmódszer konvergenciája.)*** *Elégítse ki f az alapfeltevéseket az*

$[a,b]$ *intervallumon! Ekkor az*

$fa := f(a), fb := f(b)$
$k := 0, fx := 1$
while $k < k_{\max}$ ***and*** $|fx| > toll$
$x := b - fb \cdot (b-a)/(fb-fa), fx = f(x)$
if $fx \cdot fa < 0$ ***then***
$b := x, fb := fx$
else
$a := x, fa := fx$
end if
$k := k + 1$
end while

algoritmussal előállított x_k *sorozat tart az* $f(x) = 0$ *egyenlet egyetlen* $x^\star$ *megoldásához, a konvergencia elsőrendű, és érvényes az*

$$|x_{k+1} - x^\star| \leq C|x_k - x^\star|$$

becslés, ahol $C = |x_0 - x^\star|M_2/(2m_1)$.

5.5. Tétel ***(A szelőmódszer konvergenciája.)*** *Teljesítse az* f *függvény az alapfeltevéseket az* $[a,b]$ *intervallumon! Ekkor, ha* $\max\{|a - x^\star|, |b - x^\star|\} < 2m_1/M_2$, *akkor a*

$fa := f(a), fb := f(b)$
$k := 0, fx := 1$
while $k < k_{\max}$ ***and*** $|fx| > toll$
$x := b - fb \cdot (b-a)/(fb-fa), fx = f(x)$
if $|fx| < toll$ ***then***
end
else
$a := b, b := x$
end if
end while

szelőmódszerrel előállított x_k *sorozat monoton módon* $x^\star$*-hoz tart, és a konvergencia rendje* $(1+\sqrt{5})/2 \approx 1.618$. *Továbbá érvényes az*

$$|x_{k+1} - x^\star| \leq C|x_k - x^\star||x_{k-1} - x^\star|$$

becslés a $C = M_2/(2m_1)$ *választással.*

5.6. Tétel ***(A Newton-módszer konvergenciája.)*** *Teljesítse az f függvény az alapfeltevéseket az $[a,b]$ intervallumon! Ha az*

$$
\begin{aligned}
&x := x_0, dx := 1, k := 0\\
&\textbf{\textit{while}}\ k < k_{\max}\ \textbf{\textit{and}}\ |dx| > toll\\
&\ dx = f(x)/f'(x)\\
&\ x := x - dx\\
&\ k := k + 1\\
&\textbf{\textit{end while}}
\end{aligned}
$$

algoritmust olyan x_0 pontból indítjuk, melyre $|x_0 - x^\star| < \min\{|a-x^\star|, |b-x^\star|, 2m_1/M_2\}$, akkor a módszer által előállított x_k sorozat másodrendben és monoton módon konvergál az $x^\star$ határértékhez, továbbá érvényes az

$$|x_{k+1} - x^\star| \le C|x_k - x^\star|^2$$

becslés a $C = M_2/(2m_1)$ választással.

5.7. Tétel ***(Newton-módszer monoton konvergenciája.)*** *Tegyük fel, hogy az f függvény első és második deriváltja sem vesz fel nulla értéket az $x^\star$ zérushely és az x_0 kezdőpontok által meghatározott intervallumon, és $f(x_0)f''(x_0) > 0$! Ekkor a Newton-módszer által generált $\{x_k\}$ sorozat szigorúan monoton sorozat lesz és $x^\star$-hoz tart.*

Megjegyezzük, hogy a műveletszámokat is figyelembe véve a fenti módszerek közül a szelőmódszer a leggyorsabb. Az intervallumfelezési módszer és a húrmódszer mindenképpen megtalálja valamelyik zérushelyet az intervallum belsejében, a szelő és a Newton-módszer pedig csak akkor találja meg a zérushelyet, ha megfelelő helyről indítjuk őket.

Nemlineáris egyenletek egy másfajta megoldási módszere az ún. fixpont iteráció, amely a Banach-féle fixponttételt használva állít elő egy $x^\star$-hoz tartó sorozatot. Ehhez az $f(x) = 0$ egyenletet átírjuk a vele ekvivalens $x = F(x)$ alakra egy megfelelő F függvénnyel. Az F függvény kontraktivitásának igazolásához használhatjuk az 1.23. feladat eredményét.

5.8. Tétel *Legyen $F : [a,b] \to [a,b]$ kontrakció, továbbá legyen F legalább r-szer folytonosan differenciálható úgy, hogy*

$$F'(x^\star) = \ldots = F^{(r-1)}(x^\star) = 0,$$

és $F^{(r)}(x^\star) \neq 0$. Ekkor az F által meghatározott fixpont iteráció $[a,b]$ bármelyik pontjából indítva r-edrendben tart az F függvény egyetlen $[a,b]$-beli fixpontjához.

Nemlineáris egyenletrendszerek esetén az egyenletrendszer egy $\bar{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^n$,

$$\bar{\mathbf{f}}(x_1, \ldots, x_n) \mapsto (f_1(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n))$$

vektor-vektor függvény segítségével felírható $\bar{\mathbf{f}}(\overline{\mathbf{x}}) = \mathbf{0}$ alakban, ahol $\overline{\mathbf{x}} = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

Amennyiben ekvivalens módon az $\bar{\mathbf{f}}(\overline{\mathbf{x}}) = \mathbf{0}$ egyenletet olyan $\overline{\mathbf{x}} = \overline{\mathbf{F}}(\overline{\mathbf{x}})$ alakra tudjuk átírni megfelelő $\overline{\mathbf{F}} : \mathbb{R}^n \to \mathbb{R}^n$,

$$\overline{\mathbf{F}}(x_1, \ldots, x_n) \mapsto (F_1(x_1, \ldots, x_n), \ldots, F_n(x_1, \ldots, x_n))$$

függvénnyel, melyre egy adott $\overline{\mathbf{x}}_0$ helyről indítva az $\overline{\mathbf{x}}_{k+1} = \overline{\mathbf{F}}(\overline{\mathbf{x}}_k)$ iterációt az konvergál egy $\overline{\mathbf{x}}^\star$ vektorhoz, akkor $\overline{\mathbf{x}}^\star$ nyilvánvalóan az egyenletrendszer egy megoldása lesz.

Az $\overline{\mathbf{x}}_0$ kezdővektort kitalálhatjuk pl. az egyenletrendszer megoldására vonatkozó várakozásainkból, vagy $n = 2$ esetén ábrázolhatjuk az f_1 és f_2 koordinátafüggvények 0-hoz tartozó szintvonalait és megsejthetjük ezek körülbelüli metszéspontját, vagy egyszerűen csak találomra elindítjuk néhány helyről az iterációt, bízva abban, hogy az konvergálni fog.

Ha tudjuk igazolni, hogy teljesülnek a Banach-féle fixponttétel feltételei egy bizonyos halmazon, akkor az biztosítja, hogy a halmazból tetszőleges pontról indítva a fixpont iterációt, az az egyértelműen létező fixponthoz fog tartani. A Banach-féle fixponttételt alkalmazhatjuk pl. az alábbi alakban.

5.9. Tétel ***(A Banach-féle fixponttétel nemlineáris egyenletrendszerekre. [3, 547. oldal, Theorem 10.6])*** *Tegyük fel, hogy az $\overline{\boldsymbol{x}} = \overline{\boldsymbol{F}}(\overline{\boldsymbol{x}})$ egyenlet $\overline{\boldsymbol{F}}$ iterációs függvénye a D n-dimenziós téglatartományt önmagába képezi, és folytonosan deriválható koordinátafüggvényei mindegyikére igaz egy $0 \leq q < 1$ számmal, hogy*

$$\left| \frac{\partial F_i}{\partial x_j} \right| \leq \frac{q}{n}.$$

Ekkor az $\overline{\boldsymbol{x}}_{k+1} = \overline{\boldsymbol{F}}(\overline{\boldsymbol{x}}_k)$ iteráció tetszőleges $\overline{\boldsymbol{x}}_0 \in D$ kezdővektor esetén az $\overline{\boldsymbol{F}}$ függvény egyetlen D-beli $x^\star$ fixpontjához tart, továbbá érvényes az

$$\|\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{x}}^\star\|_\infty \leq \frac{q^k}{1-q} \|\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_0\| \infty$$

hibabecslés.

Egy speciális fixpont iterációt állít elő a Newton-iteráció is.

5.10. Tétel ***(Newton-iteráció konvergenciája nemlineáris egyenletrendszerekre. [10, 283. oldal, Theorem 7.1])*** *Ha az $\overline{\boldsymbol{f}} : \mathbb{R}^n \to \mathbb{R}^n$ folytonosan deriválható függvénynek van $\overline{\boldsymbol{x}}^\star$ zérushelye egy $D \subset \mathbb{R}^n$ nyílt, konvex halmazon, valamint $\overline{\boldsymbol{x}}^\star$*

egy környezetében az $\overline{\boldsymbol{f}}$ függvény $\boldsymbol{J}$ Jacobi-mátrixa Lipschitz-folytonos, továbbá $\overline{\boldsymbol{x}}^\star$-ban invertálható, akkor létezik olyan környezete $\overline{\boldsymbol{x}}^\star$-nak, melynek bármelyik pontjából elindítva az

$$\overline{\boldsymbol{x}}_{k+1} = \overline{\boldsymbol{x}}_k - (\boldsymbol{J}(\overline{\boldsymbol{x}}_k))^{-1}\overline{\boldsymbol{f}}(\overline{\boldsymbol{x}}_k)$$

iterációt az másodrendben az egyenlet $\overline{\boldsymbol{x}}^\star$ megoldásához tart.

5.2. Feladatok

5.2.1. Sorozatok konvergenciarendje, hibabecslése

5.1. Határozzuk meg az $a_k = 1/k$ és $b_k = 2^{-k}$ sorozatok konvergenciarendjét! ⟶ ⟹

5.2. Határozzuk meg az $e_k = 10^{-2^k}$ és $f_k = 10^{-k^2}$ sorozatok konvergenciarendjét! ⟶ ⟹

5.3. Mekkora az alábbi 2-höz tartó számsorozat konvergenciarendje? ⟶ ⟹

```
2.100000000000000
2.040000000000000
2.001024000000000
2.000000000439805
```

5.4. Igazoljuk, hogy az alábbi 5-höz tartó sorozat konvergenciarendje (legalább) kettő! ⟹

```
5.200000000000000
5.080000000000000
5.012800000000000
5.000327680000000
5.000000214748365
5.000000000000092
```

5.5. Tegyük fel, hogy $x^\star$ zérushelye egy f valós-valós függvénynek, és hogy x egy tetszőleges olyan érték, hogy az $x^\star$ és x közti zárt szakasz minden pontjában f folytonosan deriválható, és van olyan $m_1 > 0$ konstans, mellyel $|f'(x)| \geq m_1$. Igazoljuk, hogy érvényes az

$$|x - x^\star| \leq \frac{|f(x)|}{m_1}$$

becslés! ⟶ ⟹

5.2.2. Zérushelyek lokalizációja

5.6. Igazoljuk, hogy az $f(x) = x \ln x - 1$ függvénynek van zérushelye az $[1, e]$ intervallumban! Hány zérushely van itt? ⟶ ⟹

5.7. Hány zérushelye van a $p(x) = x^3 - 2x^2 + 4x - 4$ polinomnak? Adjunk meg olyan 2-nél nem hosszabb intervallumokat, melyekben benne vannak a zérushelyek! ⟶ ⟹

5.8. Hány zérushelye van a $p(x) = x^3 - 2x^2 + x - 1/10$ polinomnak? Adjunk meg olyan 1-nél nem hosszabb intervallumokat, melyekben benne vannak a zérushelyek! ⟶ ⟹

5.9. Hány megoldása van az $x^2 e^x = \sin x$ egyenletnek? Hány pozitív megoldás van? Lokalizáljuk őket! Lokalizáljuk a legnagyobb negatív megoldást! ⟹

5.10. Adjunk alsó és felső becslést a $p(x) = x^5 - 4x^4 + 3x^2 + 5x - 4$ polinom zérushelyeinek abszolút értékeire! ⟶ ⟹

5.2.3. Intervallumfelezési módszer

5.11. Adjuk meg az $x^3 + x - 4$ polinom zérushelyét 10^{-2}-nál kisebb hibával úgy, hogy az intervallumfelezési módszert használjuk az [0,4] intervallummal indítva az iterációt! Becsüljük meg előre, hogy hány lépésre lesz szükségünk az adott pontosságú megoldáshoz! ⟶ ⟹

5.12. Adjunk 1/10-nél kisebb hibával becslést $\sqrt[3]{25}$-re az intervallumfelezési módszert használva! Becsüljük meg előre a szükséges lépésszámot! ⟶ ⟹

5.13. (⊟) Írjunk MATLAB programot, amely az intervallumfelezési módszert hajtja végre! ⟹

5.14. (⊞) Adjuk meg az $f(x) = e^x - x^2 - 3x + 2$ függvény [0,1] és a $g(x) = 2x \cos(2x) - (x + 1)^2$ függvény [-3,-2] intervallumbeli zérushelyeit! ⟹

5.2.4. Newton-módszer

5.15. A Newton-módszert használjuk a $2 \sin x = x$ egyenlet megoldására az $x_0 = 2$ pontból indítva. Az x_1 első iterációs lépés értékére 1.9010 adódott. Adjunk hozzávetőleges becslést arra, hogy hány iterációs lépést kell elvégeznünk ahhoz, hogy a megoldást megkapjuk legalább 5 helyes tizedesjegyre! [7, 794. oldal] ⟶ ⟹

5.16. Hány megoldása van az $e^{-x} + x^2 - 10 = 0$ egyenletnek? Határozzuk meg a pozitív megoldás(oka)t a Newton-módszer segítségével legalább hat helyes tizedesjegyre! ⟹

5.17. Határozzuk meg az $e^{-x} = \sin x$ egyenlet legkisebb pozitív megoldását a Newton-módszer segítségével négy helyes tizedesjegy pontossággal! $\Longrightarrow$

5.18. Hány valós zérushelye van a $p(x) = x^3 - x - 4$ polinomnak? Az egyik meghatározására használjuk a Newton-módszert! Végezzünk el annyi iterációs lépést, hogy két egymás utáni közelítés eltérése kisebb legyen már, mint 0.01! $\Longrightarrow$

5.19. Határozzuk meg az $x^4 - x - 10 = 0$ egyenlet legkisebb pozitív megoldását három helyes tizedesjegy pontossággal! $\Longrightarrow$

5.20. Alkalmazzuk az 5.5. feladat eredményét az $f(x) = x^2 - 2$ függvény pozitív zérushelyének Newton-módszeres megkeresésének leállási feltételeként! $\Longrightarrow$

5.21. Alkalmazzuk az 5.5. feladat eredményét az $f(x) = \cos x - x$ függvény zérushelyének Newton-módszeres megkeresésének leállási feltételeként! $\Longrightarrow$

5.22. Az $f(x) = x^3 - 3x + 2$ függvény $x^\star = 1$ zérushelyének meghatározására használtuk a Newton-módszert. Az iteráció az alábbi sorozatot generálja, ami nyilvánvalóan nem másodrendben tart 1-hez. Mi ennek az oka? Módosítsuk úgy a Newton-módszert az adott feladatra úgy, hogy az másodrendű legyen! $\longrightarrow$ $\Longrightarrow$

```
2
1.5556
1.2979
1.1554
1.0796
1.0403
1.0203
```

5.23. Igazoljuk, hogy ha f m-szer folytonosan deriválható és $x^\star$ f m-szeres zérushelye, akkor az

$$x_{k+1} = x_k - m\frac{f(x_k)}{f'(x_k)}$$

módosított Newton-iteráció másodrendben konvergens lesz! $\longrightarrow$ $\Longrightarrow$

5.24. Igazoljuk, hogy a

$$g(x) := \frac{f(x)}{f'(x)}$$

függvénynek $f(x)$ minden zérushelyénél egyszeres zérushelye van! Javasoljunk egy olyan módosított Newton-módszert ez alapján, amelynél nem fordul elő konvergenciarend-csökkenés! $\longrightarrow$ $\Longrightarrow$

5.25. Az $f(x) = -3x^3 - 5x^2 + x + 1$ függvénynek van egy zérushelye a $[-1, 0]$ intervallumban. Használhatjuk-e a megoldás megkeresésére a Newton-módszert az $x_0 = 0$ pontból indulva? $\longrightarrow$ $\Longrightarrow$

5.2.5. Húr- és szelőmódszer

5.26. Határozzuk meg az $f(x) = \cos(2x-1)$ függvény [-1,1] intervallumbeli zérushelyét a húrmódszer segítségével! Végezzünk el négy lépést a módszerrel! $\Longrightarrow$

5.27. Oldjuk meg az 5.26. feladatot a szelőmódszer segítségével! Végezzünk el öt iterációs lépést! $\Longrightarrow$

5.28. (⊞) Hasonlítsuk össze a szelő- és a Newton-módszert az

$$f(x) = \ln(\sin(x^8 - 12x^3)e^{x^2-1} + 1$$

függvény legkisebb pozitív zérushelyének megkeresésekor! Indítsuk mindkét módszert az $x_0 = 0.7$-es értékről! A szelőmódszer esetén legyen $x_1 = 0.69$. $\Longrightarrow$

5.2.6. Fixpont iterációk

5.29. Szemléltessük grafikonon az $x_{x+1} = F(x_k)$ alakú fixpont iterációkat! Mutassunk példát egy konvergens és egy nem konvergens esetre! $\Longrightarrow$

5.30. Az $x_{k+1} = \ln(1 + x_k) - (x_k - x_k^2/2)$ iteráció fixpontja $x^\star = 0$. Adjuk meg a fixpont egy olyan környezetét, ahonnét az iterációt indítva az a fixponthoz tart! Mekkora a konvergencia rendje? $\longrightarrow \Longrightarrow$

5.31. Az $f(x) = x^2 - 2 = 0$ egyenlet megoldásának meghatározására szeretnénk használni az

$$x_{k+1} = x_k + A\left(\frac{x_k^2 - 2}{x_k}\right) + B\left(\frac{x_k^2 - 2}{x_k^3}\right)$$

iterációt. Határozzuk meg úgy A és B értékét, hogy a lehető legmagasabb rendű legyen a konvergencia! $\longrightarrow \Longrightarrow$

5.32. Adjunk meg olyan $x_{k+1} = F(x_k)$ alakú fixpont iterációt, amely az $x_0 = 0$ pontból indítva a $2 - x^2 = 0$ egyenlet pozitív megoldásához tart! Azt is adjuk meg, hogy a javasolt módszerrel, mennyit kellene lépni ahhoz, hogy a megoldást 10^{-6}-nál pontosabban megközelítsük! $\Longrightarrow$

5.33. Az $x = 0.5 + \sin x$ egyenlet megoldására alkalmaztuk az

$$x^{(k+1)} = 0.5 + \sin x^{(k)},\ x^{(0)} = 1$$

iterációt, és eredményül az $x^\star = 1.497300\ldots$ értéket kaptuk. Mutassuk meg, hogy 10 iteráció után már megkaphattuk ezt a megoldást 6 helyes tizedesjegyre! $\Longrightarrow$

5.34. Igazoljuk, hogy az

$$x_{k+1} = \frac{x_k}{3} + \frac{1}{x_k}$$

iterációval előállított sorozat tetszőleges $x_0 \in [1, 2]$ kezdőérték esetén $\sqrt{3/2}$-hez tart! Ha $x_0 = 2$-ről indítjuk az iterációt, akkor hányadik tagtól esnek már a sorozat elemei a határérték 10^{-3}-os környezetébe? $\Longrightarrow$

5.35. Az alábbi fixpont iterációkat használjuk $\sqrt[3]{21}$ meghatározására az $x_0 = 3$ pontból indulva. Vizsgáljuk meg a módszereket konvergencia és konvergenciasebesség szempontjából! [3, 54. oldal]

$$a)\ x_k = \frac{20x_{k-1} + 21/x_{k-1}^2}{21}, \quad b)\ x_k = x_{k-1} - \frac{x_{k-1}^3 - 21}{3x_{k-1}^2}, \quad c)\ x_k = x_{k-1} - \frac{x_{k-1}^4 - 21x_{k-1}}{x_{k-1}^2 - 21}$$

$\longrightarrow \Longrightarrow$

5.36. Igazoljuk, hogy az

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k) - f(x_k)f''(x_k)/(2f'(x_k))}$$

iteráció harmadrendben konvergál az $f(x)$ függvény $x^\star$ zérushelyéhez, ha az egyszeres zérushely! (Ez az ún. Halley-féle iteráció.) $\Longrightarrow$

5.2.7. Nemlineáris egyenletrendszerek megoldása

5.37. Tekintsük az alábbi nemlineáris egyenletrendszert [3, 548. oldal]

$$\begin{aligned} 3x_1 - \cos(x_2x_3) - 1/2 &= 0, \\ x_1^2 - 81(x_2 + 0.1)^2 + \sin x_3 + 1.06 &= 0, \\ e^{-x_1x_2} + 20x_3 + (10\pi - 3)/3 &= 0! \end{aligned}$$

Igazoljuk, hogy az egyenletrendszernek pontosan két megoldása van a $[-1, 1] \times [-1, 1] \times [-1, 1]$ kockában! $\longrightarrow \Longrightarrow$

5.38. (⊞) Határozzuk meg az 5.37. feladat megoldásait 10^{-6}-os, maximumnormában mért hibával! $\Longrightarrow$

5.39. (⊞) Hány megoldása van az

$$\begin{aligned} x_1^2 - 10x_1 + x_2^2 + 8 &= 0 \\ x_1x_2^2 + x_1 - 10x_2 + 8 &= 0 \end{aligned}$$

nemlineáris egyenletrendszernek a $[-10, 10] \times [-10, 10]$ négyzet belsejében? [3, 551. oldal, 5. feladat] $\Longrightarrow$

5.40. (⊞) Igazoljuk, hogy az 5.39. feladatban szereplő nemlineáris egyenletrendszernek pontosan egy megoldása van a $D = [0, 1.5] \times [0, 1.5]$ négyzet belsejében! Határozzuk meg ezt a megoldást 10^{-6}-os pontossággal maximumnormában! $\Longrightarrow$

5.41. (⊞) Határozzuk meg az 5.39. nemlineáris egyenletrendszer megoldásait a Newton-módszer segítségével! $\longrightarrow \Longrightarrow$

5.42. (⊞) Határozzuk meg az

$$\begin{aligned} 5x^2 - y^2 &= 0 \\ y - 0.25(\sin x + \cos y) &= 0 \end{aligned}$$

nemlineáris egyenletrendszer $[1/4, 1/4]^T$ közelébe eső megoldását a Newton-módszer segítségével! [3, 552. oldal, 6. feladat] $\Longrightarrow$

6. fejezet

Interpoláció és approximáció

6.1. Képletek, összefüggések

Az interpolációs alapfeladat az, hogy a koordináta-rendszerben adott különböző abszcisszájú pontokhoz megkeressük egy bizonyos függvényosztályból azokat az ún. interpolációs függvényeket, melyek grafikonja átmegy az összes ponton. Csak azokkal az esetekkel foglalkozunk, amikor az interpolációs függvényt a polinomok ill. a trigonometrikus polinomok körében keressük.

6.1.1. Polinominterpoláció

6.1. Tétel ***(Interpolációs polinom egyértelműsége.)*** *Adott (x_k, f_k) $(k = 0, \ldots, n)$ pontok esetén egyértelműen létezik egy olyan L_n legfeljebb n-edfokú polinom, melynek grafikonja átmegy az összes adott ponton.*

6.2. Tétel ***(Lagrange-féle előállítás.)*** *Az L_n interpolációs polinom az*

$$L_n(x) = \sum_{k=0}^{n} f_k l_k(x)$$

képlettel adható meg, ahol

$$l_k(x) = \frac{(x - x_0) \ldots (x - x_{k-1})(x - x_{k+1}) \ldots (x - x_n)}{(x_k - x_0) \ldots (x_k - x_{k-1})(x_k - x_{k+1}) \ldots (x_k - x_n)}$$

az x_k ponthoz tartozó ún. Lagrange-féle alappolinom.

6.3. Tétel ***(Newton-féle előállítás.)*** *Az L_n interpolációs polinom előállítható*

$$\begin{aligned} L_n(x) = &[x_0]f \\ &+ [x_0, x_1]f \cdot (x - x_0) \\ &+ [x_0, x_1, x_2]f \cdot (x - x_0)(x - x_1) \\ &+ \dots \\ &+ [x_0, x_1, \dots, x_n]f \cdot (x - x_0) \dots (x - x_{n-1}) \end{aligned} \tag{6.1}$$

alakban, ahol az együtthatóként szereplő ún. osztott differenciák rekurzív módon határozhatók meg az $[x_k]f = f_k$ és

$$[x_0, \dots, x_s]f = \frac{[x_1, \dots, x_s]f - [x_0, \dots, x_{s-1}]f}{x_s - x_0}$$

képletek segítségével.

Ha egy ismert f függvény grafikonjáról választjuk az interpolálandó pontokat, akkor mérhetjük az f függvény és a pontokra illesztett L_nf interpolációs polinom pontonkénti eltérését. Az $E_n(x) = (L_nf)(x) - f(x)$ értéket az x pontbeli interpolációs hibának nevezzük.

6.4. Tétel ***(Interpolációs hiba.)*** *Amennyiben $f \in C^{n+1}$ az x pont és az $x_0, \dots, x_n$ alappontok által meghatározott I intervallumban, akkor ez az interpolációs hiba az*

$$E_n(x) = -\frac{f^{(n+1)}(\xi_x)}{(n+1)!} w_{n+1}(x)$$

alakban írható, ahol ξ_x az I intervallum belsejébe eső megfelelő konstans, és $w_{n+1}(x) = (x - x_0) \cdot \ldots \cdot (x - x_n)$ az ún. alappontpolinom. Az alappontpolinom abszolút értéke becsülhető a

$$|w_{n+1}(x)| \le \frac{h^{n+1} n!}{4}$$

formulával, ahol h a leghosszabb osztóintervallum hossza.

Általában nem garantálható, hogy az interpolációs hiba nullához tartson, ha egyre több alappontot veszünk fel.

6.5. Tétel ***(Az egyenletes konvergencia egy elégséges feltétele.)*** *Tegyük fel, hogy az f függvény tetszőlegesen sokszor deriválható az $I = [a, b]$ intervallumon és van olyan M pozitív szám, hogy az intervallum minden x pontjában $|f^{(n)}(x)| \le M^n$. Ekkor, ha az interpolációs alappontok mindig az I intervallumból kerülnek ki, akkor az L_nf interpolációs polinomsorozat egyenletesen tart az f függvényhez az I intervallumon, továbbá*

$$\|L_nf - f\|_{C[a,b]} \le \frac{M^{n+1}}{(n+1)!} (b - a)^{n+1}.$$

Az interpolációs hiba csökkenthető, ha az interpolációt ún. Csebisev-alappontokon végezzük el. A Csebisev-polinomok a $[-1,1]$ intervallumon a

$$T_0(x) = 1,\ T_1(x) = x,\ T_{n+1}(x) = 2xT_{n-1}(x) - T_{n-2}(x)$$

rekurzióval értelmezettek. T_{n+1} zérushelyei

$$x_{n+1,k} = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right)\ \ k = 0, \ldots, n$$

alakban írhatóak.

6.6. Tétel ***(Csebisev-alappontos interpoláció hibabecslése.)*** *Amennyiben $n+1$ Csebisev-alapponton interpolálunk, akkor az interpolációs hiba felső becslésére teljesül, hogy*

$$|E_n(x)| \leq \frac{M_{n+1}}{(n+1)!2^n},$$

ahol $M_{n+1} = \max_{x\in[-1,1]}\{|f^{(n+1)}(x)|\}$. A Lipschitz-folytonos függvényeket Csebisev-alappontokon interpolálva $L_nf \to f$ egyenletesen a $[-1,1]$ intervallumon.

Amennyiben olyan polinomot keresünk az interpoláció során, amelynek értékei és deriváltjai is adottak az alappontokban, akkor Hermite–Fejér-interpolációról beszélünk.

6.7. Tétel ***(Hermite–Fejér-féle intepolációs polinom egyértelműsége.)*** *Ha $n+1$ alappont adott, akkor egyértelműen létezik olyan legfeljebb $2n+1$-ed fokú H_{2n+1} polinom, amely az alappontokban az előre megadott értékeket és deriváltértékeket vesz fel.*

6.8. Tétel ***(Hermite–Fejér-féle interpolációs polinom előállítása.)*** *Az Hermite–Fejér-interpolációs előállítható a*

$$H_{2n+1}(x) = \sum_{k=0}^{n} f_k^{(0)}(1 - 2(x-x_k)l_k'(x_k))l_k^2(x) + \sum_{k=0}^{n} f_k^{(1)}(x-x_k)l_k^2(x)$$

képlettel, ahol l_k a k-adik alapponthoz tartozó Lagrange-féle alappolinom. Másfajta előállítás nyerhető az interpolációs polinom Newton-féle előállításához hasonlóan, ha minden alappontot kétszer szerepeltetünk az osztott differencia táblázatban, és két egyforma pont elsőrendű osztott differenciája helyett a pontbeli deriváltat szerepeltetjük.

6.9. Tétel ***(Hermite–Fejér-féle interpoláció hibája.)***

$$E_n(x) = H_{2n+1}(x) - f(x) = -\frac{f^{(2n+2)}(\xi_x)}{(2n+2)!}w_{n+1}^2(x),$$

ahol ξ_x egy, az I intervallum belsejébe eső megfelelő konstans.

Az eddigiekben a tejes intervallumon ugyanazzal a polinommal interpoláltunk. Most azt az esetet vizsgáljuk, amikor minden részintervallumon más-más polinom biztosítja az interpolációt.

6.10. Tétel ***(Szakaszonként lineáris interpoláció hibája.)*** *Legyen $f \in C^2$. Ha olyan folytonos s függvénnyel interpolálunk, amely minden részintervallumon legfeljebb elsőfokú, akkor az interpolációs hibára az*

$$|s(x) - f(x)| \leq \frac{M_2}{8} h^2$$

becslés érvényes, ahol M_2 egy felső korlát f második deriváltjára, és h a szomszédos alappontok közötti maximális távolság.

6.11. Tétel ***(Harmadfokú természetes spline-függvény előállítása.)*** *Tegyük fel, hogy az $x_0, \ldots, x_n$ alappontok egyforma h távolsága vannak egymástól. Ekkor az a legalább kétszer folytonosan differenciálható függvény melynek második deriváltja négyzetének integrálja a teljes intervallumon minimális egy szakaszonként legfeljebb harmadfokú s függvény lesz, amely az alábbi módon határozható meg: Megoldjuk a*

$$\frac{h}{3}\begin{bmatrix} 2 & 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & & & & & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \ldots & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_n \end{bmatrix} = \begin{bmatrix} f_1 - f_0 - D_0 h^2/6 \\ f_2 - f_0 \\ f_3 - f_1 \\ \vdots \\ f_n - f_{n-1} + D_n h^2/6 \end{bmatrix}$$

lineáris egyenletrendszert a $d_0, \ldots, d_n$ értékekre, melyek s deriváltjait adják meg az alappontokban. Ezek után az egyes szakaszok legfeljebb harmadfokú polinomjai Hermite–Fejér-interpolációval határozhatók meg.

6.12. Tétel ***(Harmadfokú természetes spline hibája.)*** *Legyen $f \in C^4[x_0, x_n]$ és s az f függvény harmadfokú spline-approximációja a h lépésközű ekvidisztáns $x_0 < x_1 < \ldots < x_n$ alappontokon. Ekkor*

$$\|s^{(r)} - f^{(r)}\|_{C[x_0,x_n]} \leq C_r h^{4-r} \|f^{(4)}\|_{C[x_0,x_n]}, \ r = 0, 1, 2, 3,$$

ahol $C_0 = 5/384$, $C_1 = 1/24$, $C_2 = 3/8$ és $C_3 = 1$.

6.1.2. Trigonometrikus interpoláció

6.13. Tétel ***(Diszkrét Fourier-együtthatók számítása páros sok alappont esetén.)*** *Tegyük fel, hogy az $x_k = 2\pi k/(n+1)$ alappontokban adottak az $f_k \in \mathbb{R}$ értékek ($k = 0, \ldots, n$). Tegyük fel, hogy n páratlan. Ekkor egyértelműen létezik egy olyan*

$m = (n+1)/2$-ed fokú kiegyensúlyozott t_m trigonometrikus polinom, melyre $t_m(x_k) = f_k$ $(k = 0, \ldots, n)$. A valós diszkrét Fourier-együtthatók az alábbi módon számolhatók:

$$a_0 = \frac{1}{n+1}\sum_{k=0}^{n} f_k, \quad a_m = \frac{1}{n+1}\sum_{k=0}^{n} f_k \cos(mx_k),$$

$$a_j = \frac{2}{n+1}\sum_{k=0}^{n} f_k \cos(jx_k) \quad (j = 1, \ldots, m-1),$$

$$b_j = \frac{2}{n+1}\sum_{k=0}^{n} f_k \sin(jx_k) \quad (j = 1, \ldots, m-1).$$

Hasonló képletek érvényesek akkor is, ha az alappontok száma páratlan. Lásd a [4] jegyzet 6.6. fejezet.

A trigonometrikus interpoláció műveletszáma jelentősen csökkenthető a gyors Fourier-transzformáció módszerének alkalmazásával. Ennek részletes leírása megtalálható a a [4] jegyzet 6.7. fejezetében.

6.1.3. Approximáció polinomokkal

Itt a célunk adott alappontokhoz meghatározni az alappontokat legkisebb négyzetek értelemben legjobban közelítő adott fokszámú polinomot. Ezt megtehetjük normálegyenlet és ortogonális függvények segítségével is.

6.14. Tétel ***(Alappontokat legkisebb négyzetek értelemben legjobban közelítő polinom meghatározása normálegyenlettel.)*** *Az (x_i, f_i) $(i = 1, \ldots, n)$ pontokat legkisebb négyzetek értelemben legjobban közelítő, legfeljebb k-adfokú ($k \leq n_{kül.} - 1$) $a_k x^k + \ldots + a_1 x + a_0$ polinom együtthatóit az*

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^k \\ 1 & x_2 & x_2^2 & \ldots & x_2^k \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^k \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_n \end{bmatrix}$$

túlhatározott lineáris egyenletrendszer legkisebb négyzetek értelemben legjobb $\overline{\mathbf{a}}_{LS}$ megoldása adja.

6.15. Tétel ***(Legkisebb négyzetek értelemben legjobb közelítés ortogonális függvények segítségével.)*** *Legyenek $\phi_1, \ldots, \phi_k$ páronként ortogonálisak és normáltak az*

$x_1, \ldots, x_n$ alappontokon, és legyen $\mathcal{F} = lin\{\phi_1, \ldots, \phi_k\}$, azaz a ϕ_i függvények összes lineáris kombinációja. Az (x_i, f_i) $(i = 1, \ldots, n)$ (különböző abszcisszájú) pontokat legkisebb négyzetek értelemben legjobban közelítő $\phi^\star$ függvény az $\mathcal{F}$ halmazból a

$$\phi^\star(x) = \sum_{i=1}^{k} (\phi_i^T(\overline{\boldsymbol{x}})\overline{\boldsymbol{f}})\phi_i(x)$$

alakban írható.

Természetesen az utóbbi tétel alkalmazásához először elő kell állítani az alappontokon ortogonális polinomokat. Mivel az $1, \sin(jx), \cos(jx)$ polinomok ortogonális rendszert alkotnak a szokásos alappontrendszereken, így a legjobban közelítő trigonometrikus polinom az interpolációs polinom megfelelő fokszámú polinomra való csonkolása lesz.

6.2. Feladatok

6.2.1. Polinominterpoláció

Interpoláció Lagrange és Newton módszerével általános alappontokon

6.1. Határozzuk meg a $(-1, 2)$, $(2, 4)$, $(3, 0)$, $(4, 2)$ pontok esetén az alappontokhoz tartozó Lagrange-féle alappolinomokat és a pontokra illeszkedő interpolációs polinomot Lagrange módszerével! ⟶ ⟹

6.2. Határozzuk meg a $(-1, 2)$, $(2, 4)$, $(3, 0)$, $(4, 2)$ pontokra illeszkedő interpolációs polinomot Newton módszerével (vö. 6.1. feladat)! Állítsuk elő az interpolációs polinomot Horner-alakban is! ⟹

6.3. Határozzuk meg az $(1, 5)$, $(3, 2)$, $(4, 3)$ pontokra illeszkedő interpolációs polinomot Lagrange és Newton módszerével is! ⟹

6.4. Hány műveletre van szükség $n + 1$ alappont esetén a Lagrange- és a Newton-féle interpolációs polinomok helyettesítési értékeinek kiszámításához? ⟹

6.5. Hogyan csökkenthető a helyettesítési értékek meghatározásának műveletszáma az interpolációs polinom Lagrange-alakjának megfelelő átalakításával? ⟶ ⟹

6.6. Határozzuk meg a $(-1, 2)$, $(2, 4)$, $(3, 0)$, $(4, 2)$ pontokra illeszkedő interpolációs polinomot a baricentrikus interpolációs formulával (vö. 6.1. feladat)! A baricentrikus interpolációs formulát lásd a 6.5. feladatban. ⟹

6.7. Határozzuk meg az alábbi pontokat interpoláló interpolációs polinomokat valamelyik tanult módszerrel!

a)

x_k	1	3
f_k	4	6

b)

x_k	1	3	4
f_k	4	6	8

c)

x_k	1	3	4	5
f_k	4	6	8	0

$\Longrightarrow$

6.8. Igazoljuk, hogy az $l_k(x)$ $(k = 0, \ldots, n)$ Lagrange-féle alappolinomokra teljesül az

$$\sum_{k=0}^{n} x_k^s l_k(x) = x^s$$

egyenlőség tetszőleges $s = 0, \ldots, n$ természetes szám esetén! $\longrightarrow \Longrightarrow$

6.9. A $\log_2 3$ értéket szeretnénk közelíteni az $f(x) = \log_2 x$ függvény $x_0 = 2$, $x_1 = 4$ és $x_2 = 8$ alappontokra illeszkedő interpolációs polinomja segítségével. Mekkora értéket ad ez a közelítés, és mekkora a várható hiba? $\Longrightarrow$

6.10. Közelítsük $\sqrt{5}$ értékét az $f(x) = \sqrt{x}$ függvényt az $x = 0, 1, 4, 9$ alappontokon interpoláló polinom $x = 5$ pontbeli értékével! $\Longrightarrow$

6.11. Az $f(x) = 1/x$ függvényt szeretnénk közelíteni a $[0.5, 1]$ intervallumon az ekvidisztáns felosztáshoz tartozó alappontokbeli függvényértékekre illesztett $p(x)$ interpolációs polinommal. Mekkora interpolációs hibára számíthatunk, ha az osztóintervallumok száma 10? $\Longrightarrow$

6.12. Hogyan egyszerűsíthető az interpolációs polinom meghatározása a Newton-módszerrel, ha az alappontok egyforma távol vannak egymástól? $\longrightarrow \Longrightarrow$

6.13. Határozzuk meg a 6.12. feladat módszerével a (4,1), (6,3), (8,8) és (10,20) pontokhoz tartozó interpolációs polinomot! $\Longrightarrow$

6.14. Az $f(x) = \ln x$ függvényt szeretnénk közelíteni az $[1, 2]$ intervallumon az ekvidisztáns felosztáshoz tartozó alappontokbeli függvényértékekre illesztett $p(x)$ interpolációs polinommal. Ha 20 osztóintervallumot használunk, akkor mekkora interpolációs hibára számíthatunk? $\Longrightarrow$

6.15. Az $f(x) = \ln x$ függvényt interpoláljuk az $[1, 2]$ intervallumon ekvidisztáns alappontokon. Igaz-e, hogy az interpolációs polinomok egyenletesen tartanak az $\ln x$ függvényhez az adott intervallumon, ha a felosztások száma végtelenhez tart? $\Longrightarrow$

6.16. Határozzuk meg az $f(x) = 1/x$ függvény esetén az $[x_0, \ldots, x_n]f$ n-edrenű osztott differenciát! $\Longrightarrow$

6.17. Határozzuk meg az $f(x) = x^{n+1}$ függvény esetén az $[x_0, \ldots, x_n]f$ n-edrenű osztott differenciát! $\Longrightarrow$

6.18. (⊟) Írjunk MATLAB programot, amely meghatározza a Newton-féle osztott differenciákat, és adott pontokban kiszámítja az interpolációs polinom értékét! $\Longrightarrow$

6.19. (⊞) Adjunk becslést az

$$\int_0^1 \sin(x^2)\,\mathrm{d}x$$

integrálra úgy, hogy az integrált a 11 ekvidisztáns eloszlású alappontban interpoláló polinom integráljával közelítjük! $\Longrightarrow$

6.20. (⊞) A vízgőz nyomása (Hgmm) az alábbi módon függ a hőmérséklettől (°C):

T	40	48	56	64	72
p	55.3	83.7	123.8	179.2	254.5

Határozzuk meg az interpolációs polinomot, és becsüljük a gőznyomást $T = 50°C$ esetén [2, 151. oldal]! $\Longrightarrow$

6.21. (⊞) A

$$K(m) = \int_0^{\pi/2} \frac{\mathrm{d}t}{\sqrt{1 - m\sin^2 t}}$$

elliptikus integrál értékei különböző m értékekre az alábbiak (két tizedesjegyre kerekítve).

m	0.00	0.20	0.40	0.60	0.80
K	1.57	1.66	1.78	1.95	2.26

Határozzuk meg az interpolációs polinomot és becsüljük az integrál értékét az alábbi m értékek esetén: $m = 0.1, 0.3, 0.5, 0.7, 0.9$! $\Longrightarrow$

Interpoláció Csebisev-alappontokon

6.22. Interpoláljuk a $\sin(\pi x/2)$ függvényt a $[-1, 1]$ intervallumon két Csebisev-alappontot használva! Írjuk fel az interpolációs polinomot és becsüljük meg az interpolációs hibát! $\Longrightarrow$

6.23. Hány Csebisev-alapponton kellene interpolálni a $\sin x$ függvényt a $[0, \pi]$ intervallumon, hogy az interpolációs hiba 10^{-6}-nál kisebb legyen? $\Longrightarrow$

6.24. (⊞) Adjunk becslést az

$$\int_0^1 \sin(x^2)\,\mathrm{d}x$$

integrálra úgy, hogy az integrált a három Csebisev-alappontban interpoláló polinom integráljával közelítjük. ⟹

6.25. Igazoljuk, hogy a Runge-példában szereplő $f(x) = 1/(1+x^2)$ függvényt a [-5,5] intervallumon Csebisev-alappontokon interpolálva az interpolációs polinomok egyenletesen tartanak f-hez, ha az alappontok száma végtelenhez tart! ⟶ ⟹

Hermite-interpoláció

6.26. Tekintsük azt a legalacsonyabb fokú q polinomot, amely átmegy az (1,0), (2,3), (3,1) pontokon és $q'(1) = q'(2) = q'(3) = 1$. Mekkora ezen polinom helyettesítési értéke az $x = 4$ pontban? ⟹

6.27. Közelítsük az $f(x) = \sin x$ függvényt Hermite–Fejér-féle interpolációs polinommal az $x = 0$, $x = \pi/2$ alappontokon! Becsüljük meg az eredmény alapján $\sin(\pi/4)$ értékét! ⟹

Szakaszonkénti polinomiális interpoláció

6.28. Tekintsük az $f(x) = \sin^2 x$ függvény grafikonjáról a $(k\pi/(n+1), f(k\pi/(n+1))$ pontokat $(k = 0, 1, \ldots, n+1)$! Tegyük fel, hogy az adott pontok közül a szomszédosakhoz tartozó szakaszokon legfeljebb elsőfokú polinommal interpolálunk, és így az egész $[0, \pi]$ intervallumon a $p(x)$ interpolációs függvényhez jutunk. Mekkora legyen n értéke, hogy $\|f - p\|_{C[0,\pi]} < 10^{-6}$ teljesüljön? ⟹

6.29. Tekintsük az $f(x) = \sqrt{x}$ függvény értékeit az $x_k = 1 + k/n$ $(k = 0, 1, \ldots, n,$ n pozitív egész) alappontokban! Minden részintervallumon illesszünk az intervallum két szélén felvett függvényértékekre és az intervallum felezőpontjában vett függvényértékre egy-egy legfeljebb másodfokú polinomot! Jelöljük azt a függvényt s-sel, amelynek az egyes intervallumokra való leszűkítései éppen a fenti interpolációs polinomok! Mekkora legyen n értéke legalább, hogy tetszőleges $\bar{x} \in [1, 2]$ pontban igaz legyen, hogy $s(\bar{x}) - f(\bar{x})| \leq 10^{-8}$? ⟹

6.30. Igazoljuk, hogy ha egy $f \in C^2$ függvényt interpolálunk három ekvidisztáns h távolságú rácspontban, akkor az interpolációs hibát felülről becsli a

$$\frac{h^3}{9\sqrt{3}} \max_x \{|f'''(x)|\}$$

kifejezés [5, 3.12. feladat]! ⟹

6.31. Jelölje $s(x)$ az $(x_0 - h, f_{-1})$, (x_0, f_0) és $(x_0 + h, f_1)$ pontokat interpoláló, szakaszonként harmadfokú természetes spline-függvényt! Igazoljuk, hogy $s'(x_0)$ megegyezik az első derivált adott alappontokon vett másodrendű központi közelítésével! $\Longrightarrow$

6.32. Határozzuk meg a (-1,2), (0,0) és (1,1) pontokat összekötő szakaszonként harmadfokú természetes spline-függvényt! $\Longrightarrow$

6.2.2. Trigonometrikus interpoláció

6.33. Határozzuk meg a $(0, 1)$, $(2\pi/3, 2)$ és $(4\pi/3, 0)$ pontokra illeszkedő legalacsonyabb fokszámú trigonometrikus polinomot! $\longrightarrow$ $\Longrightarrow$

6.34. Adjuk meg a $(0, -1), (\pi/2, 3), (\pi, 0), (3\pi/2, 1)$ pontokhoz tartozó legalacsonyabb fokú trigonometrikus interpolációs polinomot! $\Longrightarrow$

6.35. Mutassuk be a gyors Fourier-transzformáció előnyét páros alappont esetén! $\Longrightarrow$

6.36. Mutassuk be a gyors Fourier-transzformáció előnyét akkor, ha az alappontok $n+1$ száma $n + 1 = t_1 t_2$ alakban írható, ahol t_1 és t_2 két pozitív egész szám! $\Longrightarrow$

6.2.3. Approximáció polinomokkal és trigonometrikus polinomokkal

6.37. Adjuk meg a (0,1), (0,2), (1,2) és (3,0) pontokat legjobban közelítő legfeljebb elsőfokú polinomot a normálegyenlet segítségével! $\longrightarrow$ $\Longrightarrow$

6.38. Határozzuk meg a 6.38. feladatban szereplő pontokat legkisebb négyzetek értelemben legjobban közelítő legfeljebb másodfokú polinomot a normálegyenlet segítségével! $\longrightarrow$ $\Longrightarrow$

6.39. Határozzuk meg a (-1,2), (0,1), (1,3) és (3,0) pontokat legkisebb négyzetek értelemben legjobban közelítő legfeljebb elsőfokú polinomot ortogonális polinomok segítségével! $\Longrightarrow$

6.40. Határozzuk meg a 6.39. feladatban szereplő pontok legkisebb négyzetek értelemben legjobb közelítését ortogonális polinomok segítségével! $\Longrightarrow$

6.41. Melyik az az elsőfokú trigonometrikus polinom, amelyik legkisebb négyzetek értelemben a legjobban közelíti a $(0, 0)$, $(\pi/3, 1)$, $(2\pi/3, 2)$, $(3\pi/3, 3)$, $(4\pi/3, 4)$, $(5\pi/3, 5)$ pontokat? $\longrightarrow$ $\Longrightarrow$

7. fejezet

Numerikus deriválás és numerikus integrálás

7.1. Képletek, összefüggések

Numerikus deriválás

Ebben a fejezetben azzal foglalkozunk, hogy hogyan lehet egy függvény deriváltjait közelíteni adott pontokban ismert függvényértékek segítségével.

Ha egy megfelelően sokszor deriválható f függvény egy tetszőleges deriváltját az x_0 pontban Df, és ennek közelítését $\Delta f(h)$ jelöli (h argumentum azt fejezi ki, hogy a közelítés függ az alappontok h távolságától), akkor a közelítés rendje r, ha $|Df - \Delta f(h)| = \mathcal{O}(h^r)$. Bevezettük a haladó, retrográd és központi differenciákat.

- A haladó differencia az x_0 pontban: $\Delta f_+ = \dfrac{f(x_0+h) - f(x_0)}{h}$.
- A retrográd differencia az x_0 pontban: $\Delta f_- = \dfrac{f(x_0) - f(x_0-h)}{h}$.
- A központi differencia $\Delta f_c := \dfrac{\Delta f_+ + \Delta f_-}{2} = \dfrac{f(x_0+h) - f(x_0-h)}{2h}$.

A második derivált közelítésére a

$$\Delta^2 f_c := \frac{\Delta f_+ - \Delta f_-}{h} = \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2}$$

formulát alkalmazzuk.

A haladó és a retrográd formulák $f \in C^2$ esetén elsőrendű, a központi formula $f \in C^3$, a $\Delta^2 f_c$ formula $f \in C^4$ esetén másodrendű közelítést adnak.

A lépéstávolság-dilemma a hibával terhelt adatokat tartalmazó numerikus deriválási lépésköz megválasztására vonatkozik.

A fenti fogalmak részletesen megismerhetők a [4] könyv 7. fejezetéből.

Numerikus integrálás

A numerikus integrálás azt vizsgálja, hogy egy függvény néhány helyen vett függvényértékének segítségével hogyan lehet közelíteni a függvény határozott integrálját. Erre szolgálnak az interpolációs módszerek, amikor a függvényértékekre illesztett interpolációs polinomok integráljával közelítjük a tényleges integrálértéket.

Speciális klasszikus kvadratúraformulák a trapéz, érintő- és Simpson-formulák, amikor a függvény két illetve három pontjára illesztünk alacsony fokszámú interpolációs polinomot. Ha finomodó, h lépésközű rácshalókra illesztünk alacsony fokszámú interpolációs polinomot, akkor klasszikus összetett kvadratúraformulákról beszélünk. Utóbbiak konvergenciája és annak sebessége lényeges kérdés. Egy összetett kvadratúraformula közelítését r-ed rendűnek nevezzük a h lépésközű ekvidisztáns rácshálón, ha a pontos és a numerikus integrál eltérése $\mathcal{O}(h^r)$.

- Az összetett trapézformula $f \in C^2[a,b]$ függvények esetén másodrendű.
- Az összetett érintőformula $f \in C^2[a,b]$ függvények esetén másodrendű.
- Az összetett Simpson-formula $f \in C^4[a,b]$ függvények esetén negyedrendű.

A Richardson-extrapoláció a különböző rácshálókon vett közelítések kombinálásával növeli a pontosságot (rendet). A numerikus integráló formulák közül a Romberg-módszer ezen alapul. A Gauss-féle alappontmegválasztással a numerikus integrálás rendjét tudjuk növelni.

A fenti fogalmak közül a klasszikus kvadratúraformulák részletesen megismerhetők a [4] könyv 8.2 fejezetéből. Az összetett kvadratúraformulákat a 8.3 fejezet tartalmazza. A Romberg-módszer a 8.4 fejezetből ismerhető meg, míg a Gauss-kvadratúrákkal a 8.5 fejezet foglalkozik.

7.2. Feladatok

7.2.1. Numerikus deriválás

7.1. Mit approximál a

$$\frac{-3f(x_0) + 4f(x_0+h) - f(x_0+2h)}{2h}$$

kifejezés? Határozzuk meg a közelítés hibáját! $\Longrightarrow$

7.2. Mit approximál az

$$\frac{f(x_0-2h)-8f(x_0-h)+8f(x_0+h)-f(x_0-2h)}{12h}$$

kifejezés? Határozzuk meg a közelítés hibáját! $\longrightarrow$

7.3. Mit approximál a

$$\frac{-f(x_0-2h)+16f(x_0-h)-30f(x_0)+16f(x_0+h)-f(x_0+2h)}{12h}$$

kifejezés? Határozzuk meg a közelítés hibáját! $\longrightarrow$

7.4. Mit approximál az

$$\frac{f(x_0-2h)-4f(x_0-h)+6f(x_0)-4f(x_0+h)+f(x_0+2h)}{h^4}$$

kifejezés? Határozzuk meg a közelítés hibáját! $\Longrightarrow$

7.5. Adjuk meg az

$$f'(x_0)\approx\frac{f(x_0+h)-f(x_0+2h)}{2h}$$

középponti szabály ϵ-hibával megadott függvényértékek melletti optimális h lépéshossz megválasztását! $\Longrightarrow$

7.6. Adjuk meg a 7.2. feladatban szereplő kifejezés felső határoló függvényét ϵ pontosságú adatok esetén! Határozzuk meg az optimális h lépéshossz értékét! $\longrightarrow$

7.7. Határozzuk meg a második deriváltat másodrendben közelítő centrális differencia felső határoló függvényét ϵ pontosságú adatok esetén! Határozzuk meg az optimális h lépéshossz értékét! $\longrightarrow$

7.8. Approximáljuk az $f''(x_0)$-t az $f(x_0-h)$, $f(x_0)$ és $f(x_0+h)$ értékekből az

$$Af(x_0-h)+Bf(x_0)+Cf(x_0+h)$$

kifejezéssel, ahol A, B és C adott állandók! Adjuk meg a pontos feltételt az A, B, C számokra! $\Longrightarrow$

7.9. (⊟) Írjunk olyan MATLAB programot, amely a $\sin''(0.5)=-0.479425538604203$ értéket a másodrendű centrális differenciával közelíti! Magyarázzuk meg, hogy miért ingadozik a 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} és 10^{-6} lépésközök mellett az abszolút értékben vett hiba! $\Longrightarrow$

7.10. (⊞) Az alábbi közelítő kifejezések közül válasszuk ki azokat, amelyekkel lehet közelíteni az $f'(1)$, $f''(1)$ és $f'''(1)$ deriváltakat és határozzuk meg, hogy a módszerek milyen közelítő értéket adnak!

(a) $\dfrac{f(x_0+h)-f(x_0)}{h}$,

(b) $\dfrac{f(x_0)-f(x_0-h)}{h}$,

(c) $\dfrac{f(x_0+h)-2f(x_0)+f(x_0-h)}{h^2}$,

(d) $\dfrac{f(x_0+2h)-2f(x_0+h)+f(x_0)}{h^2}$,

(e) $\dfrac{-f(x_0+h)+3f(x_0+2h)-3f(x_0+3h)+f(x_0+4h)}{h^3}$,

(f) $\dfrac{-2f(x_0+h)-3f(x_0+2h)+2f(x_0+3h)+2f(x_0+4h)}{h^3}$,

ha az f függvény függvényértékeit az alábbi táblázat tartalmazza:

x	1.00	1.05	1.10	1.15	1.20	1.25
$f(x)$	1.00000	1.02470	1.04881	1.07238	1.09544	1.11803

7.1. táblázat. Adott alappontokhoz tartozó függvényértékek.

⟶

7.2.2. Numerikus integrálás

7.11. Számítsuk ki az

$$\int_0^1 x^2\,\mathrm{d}x$$

integrál értékét érintő-, trapéz- és Simpson-formulával! Mekkora a hiba? ⟹

7.12. Számoljuk ki az

$$\int_0^1 \frac{1}{1+x^2}\,\mathrm{d}x$$

integrál értékét a $[0,1]$ intervallum három részre való felosztásával összetett trapézformulával! Mekkora a hiba? ⟹

7.13. Határozzuk meg a 7.12. feladat esetén hány intervallum kell ahhoz, hogy 10^{-5} pontossággal megkaphassuk a pontos értéket! $\Longrightarrow$

7.14. (⊞) Határozzuk meg a 7.12. feladatbeli integrál pontos értékét! Az intervallumszámok növelésével vizsgáljuk meg az egyes összetett kvadratúraformulák konvergenciarendjét számítógép segítségével! $\longrightarrow$

7.15. (⊞) Számoljuk ki az

$$\int_{-2}^{2} (x^5 - 3x^3 + 2x + 1)\,\mathrm{d}x$$

integrál értékét a $[-2, 2]$ intervallum 23 részre való felosztásával összetett trapézformulával! $\longrightarrow$

7.16. (⊟) Írjunk olyan MATLAB programot, amely n részre történő osztással, összetett trapézformulával közelíti az integrál értékét! $\Longrightarrow$

7.17. (⊟) Módosítsuk a 7.16. feladatban megírt programunkat úgy, hogy az előző feladatot összetett érintőformulával oldja meg! $\longrightarrow$

7.18. (⊟) Írjunk olyan MATLAB programot, melyben kiválaszthatjuk, hogy az adott integrál értékét mely módszerrel (összetett érintő-, trapéz- és Simpson-formula) és hány intervallumra történő osztással közelítjük! $\Longrightarrow$

7.19. Határozzuk meg a zárt $N^{4,k}$ Newton–Cotes-együtthatókat! $\longrightarrow$

7.20. Határozzuk meg az $2 + \cos(2\sqrt{x})$ függvény közelítő integrálját a $[0, 2]$ intervallumon a 7.19. feladatban kiszámolt együtthatók segítségével! $\Longrightarrow$

7.21. Készítsük el a három pontra illeszkedő Gauss–Legendre-formulát! $\Longrightarrow$

7.22. Készítsük el a három pontra illeszkedő Gauss–Csebisev-formulát! $\longrightarrow$

7.23. Határozzuk meg a Gauss–Csebisev-formulával az

$$\int_{-1}^{1} \frac{x^4}{\sqrt{1 - x^2}}\,\mathrm{d}x$$

integrál közelítő értékét! $\Longrightarrow$

7.24. Keressünk olyan c_i konstansokat, hogy az

$$\int_{0}^{4} f(x)\,\mathrm{d}x \approx c_1 f(1) + c_2 f(2) + c_3 f(4)$$

közelítő integrálás minden legfeljebb másodfokú polinomra pontos legyen! $\Longrightarrow$

7.25. (⊞) Tekintsük az alábbi integrált:

$$\int_0^{0.8} (0.2 + 25x - 200x^2 + 675x^3 - 900x^4 + 400x^5)\,\mathrm{d}x.$$

Határozzuk meg a közelítő integrál értékét a Richardson-extrapolációval, ha a MATLAB-ban a Crank–Nicolson-sémát használtuk a módszer indításához szükséges numerikus értékek számításához 1, 2 és 4 intervallumszám esetén! Számítsuk ki a hibaszámításhoz az integrál pontos értékét és vessük össze a módszerek jóságát is! ⟹

7.26. A 7.25. eredményeit figyelembe véve alkalmazzuk a Romberg-módszert úgy, hogy a módszer a pontos integrál értékét negyed-, hatod-, illetve nyolcadrendben közelítse! ⟹

7.27. Határozzuk meg Romberg-módszerrel 10^{-8} pontossággal a Gauss-függvény integrálját a $[0, 1]$ intervallumon! ⟶

7.28. (⊟) Írjunk olyan MATLAB programot, amely Romberg-módszerrel közelíti az

$$\int_0^{\pi} \sin(x)\,\mathrm{d}x = 2$$

integrál értékét, ha a függvény bemenő paramétere az extrapolációs lépésszám! ⟶

8. fejezet

A közönséges differenciálegyenletek kezdetiérték-feladatainak numerikus módszerei

8.1. Képletek, összefüggések

Ebben a fejezetben a közönséges differenciálegyenletek kezdetiérték-feladatainak numerikus megoldásait vizsgáljuk. Szokásos módon az elsőrendű differenciálegyenleteket vizsgáljuk, azaz az $y'(t) = f(t, y(t))$ egyenletet az $y(0) = y_0$ kezdeti feltétellel. A numerikus megoldás az ismeretlen $y(t)$ függvény egy $t_i = ih$ $(i = 0, 1, \ldots, N)$ rácshálón való közelítését jelenti, ahol az $y(t_i)$ közelítését jelentő y_i értéket valamilyen képlet segítségével határozzuk meg.

Megkülönböztetjük az *egylépéses módszereket* (amikor csak a t_{i-1} pontbeli közelítést használjuk y_i kiszámolására) és a *többlépéses módszereket* (amikor több megelőző pontbeli közelítést használunk y_i kiszámolására). A módszerek pontosságát a lokális approximációs hiba jellemzi, amely azt fejezi ki, hogy a pontos megoldás rácshálón vett vetülete h milyen rendjében elégíti ki a numerikus megoldást meghatározó sémát.

Az egylépéses módszerek közül kiemeljük az alábbiakat.

- Explicit Euler-módszer: $y_i = y_{i-1} + hf(t_{i-1}, y_{i-1})$.
- Implicit Euler-módszer: $y_i = y_{i-1} + hf(t_i, y_i)$.
- A Crank–Nicolson-módszer: $y_i = y_{i-1} + 0.5h(f(t_{i-1}, y_{i-1}) + f(t_i, y_i))$.

Az első két módszer elsőrendű, míg a harmadik módszer másodrendű. Ezek általánosítása a θ-módszer, amelynek speciális esetei a fenti módszerek. A fenti módszerek közül a második és harmadik is implicit, azaz y_i meghatározása csak egy egyenlet megoldásával

lehetséges. Ennek kiküszöbölésére ezeket a módszereket explicitté tehetjük az algoritmus módosításával. Így származtathatók a javított Euler-, illetve az Euler–Heun-módszerek, ill. ezek általánosításaként a Runge–Kutta-típusú módszerek, amikor is az ún. Butcher-táblázat segítségével több köztes érték segítségével számoljuk ki az y_{i-1} értékből az y_i értékét. Ezek a módszerek a köztes értékek számától (az ún. lépcsőszámtól) függően általában magasabb rendben pontosak.

Az egylépéses módszerek általánosítása a lineáris többlépéses módszerek, amelyek alakja

$$a_0 y_i + a_1 y_{i-1} + \cdots + a_m y_{i-m} = h(b_0 f_i + b_1 f_{i-1} + \cdots + b_m f_{i-m}), \quad i = m, m+1, \ldots,$$

ahol $f_i = f(t_i, y_i)$, és a_k és b_k a módszert definiáló adott paraméterek. Ezek a módszerek b_0 értékétől függően szintén lehetnek explicitek ($b_0 = 0$) és implicitek ($b_0 \neq 0$). Pontosságukat az m lépésszám határozza meg.

Fontos kérdés a numerikus megoldás rögzített rácshálón való viselkedésének vizsgálata. Ilyenek az A-stabilitás, illetve az erős stabilitás.

A fenti fogalmak részletesen megismerhetők a [4] könyv 9. fejezetéből. Ezen belül a 9.3.2 szakasz a nevezetes egylépéses módszerekkel, a 9.4 szakasz pedig a Runge–Kutta-módszerekkel foglalkozik. A lineáris többlépéses módszerekkel a 9.5 fejezetben ismerkedhetünk meg. Az A-stabilitást a 9.6 fejezet ismerteti.

8.2. Feladatok

8.2.1. Egylépéses módszerek

8.1. Határozzuk meg az alábbi módszerek konzisztenciarendjét:

(a) explicit Euler,

(b) implicit Euler,

(c) Crank–Nicolson,

(d) θ-módszer!

$\longrightarrow \Longrightarrow$

8.2. Tekintsük az

$$\begin{cases} \dot{y}(t) = 1 - 10y(t) \\ y(0) = 0 \end{cases}$$

kezdetiérték-feladatot. Számítsuk ki a megoldás közelítő értékét a $t = 2$ pontban $h = 1/2, 1/4, 1/8, 1/16$ lépésközök esetén, ha a módszer

(a) explicit Euler,

(b) implicit Euler,

(c) Crank–Nicolson,

(d) javított Euler,

(e) Euler–Heun!

$\longrightarrow \Longrightarrow$

8.3. Tekintsük az

$$\begin{cases} \dot{y}(t) = \dfrac{2y(t)}{t} \\ y(1) = 1 \end{cases}$$

kezdetiérték-feladatot. Számítsuk ki a megoldás közelítő értékét a $t = 2$ pontban $h = 1/2, 1/4, 1/8, 1/16$ lépésközök esetén, ha a módszer

(a) explicit Euler,

(b) implicit Euler,

(c) Crank–Nicolson,

(d) javított Euler,

(e) Euler–Heun!

$\longrightarrow$

8.4. Tekintsük az

$$\begin{cases} 4\dot{y}(t) = ty(t) + 2 \\ y(0) = 3 \end{cases}$$

kezdetiérték-feladatot. Számítsuk ki a megoldás közelítő értékét a $t = 2$ pontban $h = 1/2, 1/4, 1/8, 1/16$ lépésközök esetén, ha a módszer

(a) explicit Euler,

(b) implicit Euler,

(c) Crank–Nicolson,

(d) javított Euler,

(e) Euler–Heun!

⟶

8.5. Tekintsük az

$$\begin{cases} \dot{y}(t) + 0.4y(t) = 3e^{-t} \\ y(0) = 5 \end{cases}$$

kezdetiérték-feladatot. Számítsuk ki a megoldás közelítő értékét a $t = 3$ pontban $h = 1/2, 1/4, 1/8, 1/16$ lépésközök esetén, ha a módszer

(a) explicit Euler,

(b) javított Euler,

(c) Euler–Heun!

⟶

8.6. (⊟) Határozzuk meg a 8.2.-8.5. feladatok módszerei közül melyek explicit módszerek! Írjunk olyan MATLAB programokat, amelyek megoldják a 8.2.-8.5. feladatokat!
⟶

8.7. (⊞) Alkalmazzuk a 8.2.-8.5. feladatokra a MATLAB ODE45 beépített módszerét!
⟶

8.8. (⊞) Számítsuk ki a 8.3. feladat pontos megoldását és vessük össze a kapott numerikus megoldásokkal! A lépésköz felezésével a hiba különböző mértékben csökken. Mivel magyarázható ez? ⟹

8.9. Válasszuk meg az $y_{n+1} = y_n + h[c_1 f(t_n, y_n) + c_2 f(t_n + ah, y_n + bhf(t_n, y_n))]$ egylépéses módszerben a c_1, c_2, a, b paraméterek értékeit úgy, hogy a módszer rendje minél magasabb legyen! ⟶

8.10. Írjuk fel a 8.2. feladat explicit módszereinek Butcher-táblázatát! ⟹

8.11. Írjuk fel képlet alakban a Butcher-táblázattal megadott klasszikus negyedrendű Runge–Kutta-módszert! ⟹

8.12. Írjuk fel képlet alakban az alábbi Butcher-táblázat formában megadott Runge-Kutta-módszereket! ⟶

(a)

0	0	0	0
1/2	1/4	1/4	0
1	0	1	0
	1/6	2/3	1/6

(b)
0	0	0	0
1/2	1/2	0	0
1	-1	2	0
	1/6	2/3	1/6

(c)
0	0	0	0	0
1/2	1/2	0	0	0
1/2	1/4	1/4	0	0
1	0	-1	2	0
	1/6	0	2/3	1/6

(d)
1/3	1/3	0
1	1	0
	3/4	1/4

8.13. Írjuk fel képlet alakban az alábbi Butcher-táblázat formában megadott implicit Runge–Kutta-módszereket!

(a)
1	1
	1

(b)
1/2	1/2
	1/2

(c)
0	0	0
1	1/2	1/2
	1/2	1/2

$\longrightarrow$

8.14. Butcher-táblázat segítségével határozzuk meg a 8.10.-8.11. feladatokban szereplő módszerek rendjét! $\longrightarrow \Longrightarrow$

8.15. Írjuk fel az alábbi explicit Runge–Kutta-módszerek Butcher-táblázatát!

(a) $y_{n+1} = y_n + hf(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hf(t_n, y_n))$

(b) $y_{n+1} = y_n + h[(1 - \frac{1}{2\alpha})f(t_n, y_n) + \frac{1}{2\alpha}f(t_n + \alpha h, y_n + \alpha hf(t_n, y_n))]$

(c) $y_{n+1} = y_n + h\left[\frac{1}{4}f(t_n, y_n) + \frac{3}{4}f\left(t_n + \frac{2}{3}h, y_n + \frac{2}{3}f(t_n + \frac{1}{3}h, y_n + \frac{1}{3}f(t_n, y_n))\right)\right]$

(d) $y_{n+1} = y_n + h\left[\frac{1}{4}f(t_n, y_n) + \frac{3}{4}f\left(t_n + \frac{2}{3}h, y_n + \frac{2}{3}f(t_n, y_n))\right)\right]$

(e) $y_{n+1} = y_n + h\left[(1-\theta)\, f\left(t_n, y_n\right) + \theta f\left(t_{n+1}, y_{n+1}\right)\right]$

$\longrightarrow \Longrightarrow$

8.16. Határozzuk meg az alábbi módszerek stabilitási függvényét!

(a) explicit Euler

(b) implicit Euler

(c) Crank–Nicolson

(d) θ-módszer

(e) javított Euler

(f) Euler–Heun

(g) implicit középpontszabály

$\longrightarrow \Longrightarrow$

8.17. Határozzuk meg, hogy a 8.16. feladat módszerei közül melyek A-stabilak! $\longrightarrow \Longrightarrow$

8.18. (⊟) Írjunk olyan MATLAB programot, amely az RK1, RK2, RK3 és RK4 módszerek stabilitási tartományait ábrázolja! $\longrightarrow$

8.19. (⊟) Írjunk olyan MATLAB programot, amely a 8.18. feladat stabilitási tartományainak határvonalait egy ábrán jeleníti meg! $\Longrightarrow$

8.20. (⊟) Írjunk olyan MATLAB programot, amely az implicit Euler és Crank–Nicolson-módszerek stabilitási tartományait ábrázolja! $\Longrightarrow$

8.21. Tekintsük az alábbi tesztfeladatot:

$$\begin{cases} \dot{y}(t) = \lambda y(t), \quad t \in [0, \infty), \ \lambda \in \mathbb{R}, \\ y(0) = 1. \end{cases}$$

A 8.1 táblázatban a különböző λ értékekkel kitűzött tesztfeladat numerikus megoldásának hibáit láthatjuk a $t = 1$ pontban.
Adjunk magyarázatot arra, hogy miért viselkednek ennyire eltérően az explicit és implicit Euler-módszerek bizonyos h értékek esetén! $\Longrightarrow$

8.22. Válaszoljuk meg az alábbi Crank–Nicolson-módszerrel kapcsolatos kérdést! Hogyan viselkedik a módszer $h > 2/(-\lambda), \ \lambda \in \mathbb{R}^-$ esetén? $\Longrightarrow$

	$\lambda = -9$		$\lambda = -99$		$\lambda = -999$	
h	EE	IE	EE	IE	EE	IE
0.1	3.07e-01	1.20e-01	3.12e+09	9.17e-02	8.95e+19	9.93e-03
0.01	1.72e-02	1.60e-02	3.62e-01	1.31e-01	2.38e+95	9.09e-02
0.001	1.71e-03	1.60e-03	1.90e-02	1.75e-02	3.67e-01	1.32e-01
0.0001	1.66e-04	1.65e-04	1.78e-03	1.68e-03	1.92e-02	1.76e-02
0.00001	1.66e-05	1.65e-05	1.82e-04	1.82e-04	1.83e-03	1.83e-03

8.1. táblázat. Hibaértékek Euler-módszerek esetén adott h és λ értékek mellett.

	EE	IE
y_0	0	0
y_1	0.50000	0.08333
y_2	-1.50000	0.09722
y_3	6.50000	0.09936

8.2. táblázat. A $h = 1/2$ lépésközre számolt numerikus értékek.

8.23. Tekintsük a 8.2. feladatban szereplő kezdetiérték-feladatot. Adjuk meg azon h kritikus lépésközértéket, amely mellett a feladatra alkalmazott explicit Euler-módszerrel nyert közelítő megoldás oszcillál! $\Longrightarrow$

8.24. Tekintsük a 8.2. feladatban szereplő kezdetiérték-feladatot. A 8.2 táblázatban a $h = 1/2$ lépésközű explicit Euler és implicit Euler-módszerek eredményeit láthatjuk. Magyarázzuk meg, hogy ilyen h választása mellett az explicit Euler-módszer elszálló eredményt ad, míg az implicit Euler jól közelíti a feladat megoldását! $\longrightarrow$

8.2.2. Többlépéses módszerek

8.25. Taylor-sorfejtés útján határozzuk meg az alábbi kétlépéses módszerek konzisztenciarendjét!

(a) $y_n - \frac{4}{3}y_{n-1} + \frac{1}{3}y_{n-2} = \frac{2}{3}hf_n$

(b) $y_n - 4y_{n-1} + 3y_{n-2} = -2hf_{n-2}$

(c) $y_n + 4y_{n-1} - 5y_{n-2} = h\Big(4f_{n-1} + 2f_{n-2}\Big)$

$\longrightarrow \Longrightarrow$

8.26. Mennyi a konzisztenciarendje az alábbi többlépéses módszereknek?

(a) $y_n - \frac{4}{3}y_{n-1} + \frac{1}{3}y_{n-2} = \frac{2}{3}hf_n$

(b) $y_n - y_{n-1} = h\left(\frac{3}{2}f_{n-1} - \frac{1}{2}f_{n-2}\right)$

(c) $y_n - y_{n-2} = 2hf_{n-1}$

(d) $y_n - y_{n-1} = h\left(\frac{23}{12}f_{n-1} - \frac{4}{3}f_{n-2} + \frac{5}{12}f_{n-3}\right)$

$\longrightarrow \Longrightarrow$

8.27. Határozzuk meg az $y_n + a_1y_{n-1} + a_2y_{n-2} = h(b_1f_{n-1} + b_2f_{n-2})$ kétlépéses módszer együtthatóit úgy, hogy a konzisztenciarendje minél magasabb legyen! $\Longrightarrow$

8.28. Határozzuk meg az $y_n - y_{n-1} = h(b_1f_{n-1} + b_2f_{n-2} + b_3f_{n-3})$ háromlépéses módszer együtthatóit úgy, hogy a konzisztenciarendje minél magasabb legyen! $\longrightarrow$

8.29. Oldjuk meg az $y_n - 4y_{n-1} + 3y_{n-2} = -2hf_{n-2}$ módszerrel az

$$\begin{cases} \dot{y}(t) = -y(t), & t \in [0,1] \\ y(0) = 1 \end{cases}$$

egyenletet $h = 1/10$ választással! Nézzük meg minden egyes lépés után, hogy a hiba hogyan változik! Konvergens-e a módszer? $\Longrightarrow$

8.30. Az alábbi módszerek közül melyek teljesítik a gyökkritériumot?

(a) $y_n - 6y_{n-1} + 5y_{n-2} = h(4f_{n-1} + 2f_{n-2})$

(b) $y_n - y_{n-2} = \frac{h}{2}(f_n + 4f_{n-1} + f_{n-2})$

(c) $y_n + -\frac{4}{3}y_{n-1} + \frac{1}{3}y_{n-2} = \frac{2}{3}hf_n$

(d) $y_n - \frac{11}{6}y_{n-1} + y_{n-2} - \frac{1}{6}y_{n-3} = h(2f_{n-2} - 3f_{n-3})$

(e) $y_n - 2y_{n-2} + y_{n-4} = h(f_n + f_{n-3})$

$\longrightarrow \Longrightarrow$

8.31. Határozzuk meg, hogy a 8.25., 8.26. és 8.30. feladatokban szereplő többlépéses módszerek közül melyek lesznek erősen stabilak! $\longrightarrow \Longrightarrow$

8.32. Mutassuk meg, hogy az Adams-módszerek erősen stabilak! $\Longrightarrow$

9. fejezet

A közönséges differenciálegyenletek peremérték-feladatainak numerikus módszerei

9.1. Képletek, összefüggések

Ebben a fejezetben a közönséges differenciálegyenletek peremérték-feladatainak numerikus megoldásait tárgyaljuk. Tipikusan a másodrendű differenciálegyenleteket vizsgáljuk egy korlátos $[a, b]$ intervallumon, azaz az $u''(t) = f(t, u(t), u'(t))$ egyenletet, ahol a megoldás a két végpontban ismert, azaz adottak az $u(a) = \alpha$ és $u(b) = \beta$ peremfeltételek. Fontos megjegyezni, hogy a Cauchy-féle kezdetiérték-feladattól eltérően erre a feladatra az egyértelmű megoldás létezése nemcsak az f függvény alakjától függ, hanem a peremfeltétel megadásától is.

A legtipikusabb numerikus megoldási módszerek a *belövéses módszer* és a *véges differenciák módszere*.

A belövéses módszer lényege, hogy a másodrendű egyenlet peremérték-feladatának megoldását visszavezetjük elsőrendű Cauchy-féle kezdetiérték-feladatra, és ezek megoldására a korábban megismert numerikus módszerek valamelyikét alkalmazzuk. A visszavezetést az $u_1(t) = u(t)$ és az $u_2(t) = u'(t)$ új függvények bevezetésével hajtjuk végre, amelyek segítségével a feladatunk az

$$\begin{aligned} u_1'(t) =& u_2(t) \\ u_2'(t) =& f(t, u_1(t), u_2(t)) \end{aligned}$$

alakot ölti. A kezdeti feltétel az u_1 függvényre ismert az eredeti feladatból ($u_1(a) = \alpha$). Az $u_2(a)$ értéket úgy kell meghatározni, hogy a kezdetiérték-feladat megoldására az $u_1(b) = \beta$ egyenlőség teljesüljön. A belövéses módszer lényege az $u_2(a) = c$ feltételből az

ismeretlen c paraméter meghatározása. Ez a probléma visszavezet a nemlineáris egyenletek megoldásának problémájához. Tehát a belövéses módszer realizálása két numerikus eljárás alkalmas megválasztását jelenti: elsőrendű Cauchy-féle kezdetiérték-feladat megoldása valamely módszerrel, illetve a nemlineáris egyenletek megoldása numerikus módszerrel.

A másik tipikus módszer a véges differenciák módszere, amelynek során az $[a, b]$ intervallumon egy rácshálót generálunk, a rácsháló pontjaiban az $u(t)$ függvény első és második deriváltjait a szokásos véges differenciákkal közelítjük. Ezzel az $u''(t_i) = f(t_i, u(t_i), u'(t_i))$ egyenlet felhasználásával numerikus eljárást konstruálhatunk az $u(t_i)$ ismeretlen értékek y_i közelítésének meghatározására. Alapvető kérdés a konvergencia belátása, azaz annak kimutatása, hogy finomodó rácshálók ($h \to 0$) esetén a numerikus megoldás tart-e (ha igen, akkor milyen rendben) a pontos megoldáshoz.

Lényegesen egyszerűbb a lineáris eset, amikor az

$$u''(t) = p(t)u'(t) + q(t)u(t) + r(t)$$

egyenletet vizsgáljuk, ahol p, q és r adott függvények. Ilyenkor a véges differenciák módszere egy lineáris algebrai egyenlethez vezet. Ennek numerikus kezelése lényegesen könnyebb. Emellett a konvergencia, illetve annak rendjének kérdése is megválaszolható.

A fenti fogalmak részletesen megismerhetők a [4] könyv 10. fejezetéből. Ezen belül a folytonos feladat megoldhatóságával a 10.3, a belövéses módszerrel a 10.4 foglalkozik. A véges differenciás approximációt és annak konvergenciáját a 10.2 és a 10.5 szakaszok tárgyalják.

9.2. Feladatok

9.2.1. Peremérték-feladatok megoldhatósága

9.1. Állítsuk elő az

$$\begin{cases} u''(x) = u(x), \quad x \in (0,1) \\ u(0) = 2/3, \ u(1) = 3/8 \end{cases}$$

feladat megoldását! $\Longrightarrow$

9.2. Vizsgáljuk meg a 9.1. feladatot az $u(0) = 0, \ u(1) = 1$ peremfeltételekkel! $\longrightarrow$

9.3. Tekintsük az

$$\begin{cases} u''(x) = -4u(x), \quad x \in (0, \pi/2) \\ u(0) = 1, \ u(\pi/2) = -1 \end{cases}$$

peremérték-feladatot. Melyik állítás igaz az alábbiak közül?

(a) Nincs megoldása.

(b) Egyértelmű megoldása van.

(c) Az elemi függvények körében van megoldása.

$\Longrightarrow$

9.4. Adjuk meg a 9.3. feladat kérdéseire a helyes válaszokat, ha a feladat peremfeltételei $u(0) = 1,\ u(\pi/2) = 2$ alakúak! $\longrightarrow$

9.5. Határozzuk meg az

$$\begin{cases} u''(x) - 2u'(x) + u(x) = 0, & x \in (0,1) \\ u(0) = \alpha,\ u(1) = \beta \end{cases}$$

feladat megoldását! Van olyan (α, β) pár, amelyre a feladatnak nem létezik megoldása?
$\longrightarrow$

9.6. Tekintsük az

$$\begin{cases} u''(x) = -u(x), & x \in (a,b) \\ u(a) = \alpha,\ u(b) = \beta \end{cases}$$

peremérték-feladatot. Mit mondhatunk a feladat megoldásáról, ha a peremfeltételek a következőek:

(a) $a = 0,\ b = \pi/2,\ \alpha = 3,\ \beta = 7,$

(b) $a = 0,\ b = \pi,\ \alpha = 3,\ \beta = 7$?

$\longrightarrow$

9.7. Van-e az alábbi feladatoknak egyértelmű megoldása?

(a) $\begin{cases} u''(x) = \sin(x) + u(x), & x \in (1,4) \\ u(1) = 3,\ u(4) = 7 \end{cases}$

(b) $\begin{cases} u''(x) = \sin(x)u'(x) + 2u(x) + e^x, & x \in (1,2) \\ u(1) = 3,\ u(2) = 4 \end{cases}$

(c) $\begin{cases} u''(x) = \lambda u'(x) + \lambda^2 u(x),\ x \in [0,1],\ \lambda \in [0.5,1] \\ u(0) = 5,\ u(1) = 8 \end{cases}$

$\longrightarrow \Longrightarrow$

9.8. Írjuk fel a peremérték-feladatokat elsőrendű rendszer alakjában!

(a) $\begin{cases} u''(x) = u(x), \quad x \in (a,b) \\ u(a) = \alpha,\ u(b) = \beta \end{cases}$

(b) $\begin{cases} u''(x) = \lambda u'(x) + \lambda^2 u(x),\ x \in [0,1],\ \lambda \in [0.5,1] \\ u(0) = 5,\ u(1) = 8 \end{cases}$

(c) $\begin{cases} u'''(x) = -2\lambda^3 u(x) + \lambda^2 u'(x) + 2\lambda u''(x), \quad x \in (0,1) \\ u(0) = \beta_1,\ u(1) = \beta_2,\ u'(1) = \beta_3 \end{cases}$

$\longrightarrow \Longrightarrow$

9.9. Rendszerekre vonatkozó ismereteink birtokában vizsgáljuk meg az alábbi feladatok megoldhatóságát!

(a) $\begin{cases} u''(x) = -u(x), \quad x \in (0,b) \\ u(0) = \alpha,\ u(b) = \beta \end{cases}$

(b) $\begin{cases} u''(x) = u(x), \quad x \in (0,b) \\ u(0) = \alpha,\ u(b) = \beta \end{cases}$

$\longrightarrow \Longrightarrow$

9.2.2. Véges differenciák módszere és a belövéses módszer

9.10. Tekintsük a

$$\begin{cases} -u''(x) = f(x), \quad x \in (0,l) \\ u(0) = \mu_1,\ u(l) = \mu_2 \end{cases}$$

peremérték-feladatot. Alkalmazzunk egy véges differenciák módszerén alapuló diszkretizációt, majd írjuk fel a kapott lineáris egyenletrendszert! $\Longrightarrow$

9.11. Tekintsük a

$$\begin{cases} -u''(x) + c(x)u(x) = f(x), \quad x \in (0,l) \\ u(0) = \mu_1,\ u(l) = \mu_2 \end{cases}$$

peremérték-feladatot, ahol $c(x)$ egy $C[0,l]$-beli nemnegatív függvény. Írjuk fel az operátoregyenletes alakot! $\Longrightarrow$

9.12. Írjuk fel a 9.11. feladat véges differenciás közelítését és annak operátoregyenletes alakját! $\longrightarrow$

9.13. Igazoljuk, hogy a 9.11. feladat diszkretizációjából származó együtthatómátrix M-mátrix! $\Longrightarrow$

9.14. Mutassuk meg a 9.12. feladatban meghatározott közelítések konvergenciáját! $\longrightarrow$

9.15. Tekintsük a

$$\begin{cases} u''(x) + a(x)u'(x) + b(x)u(x) = f(x), \quad x \in (0, l) \\ u(0) = \mu_1, \ u(l) = \mu_2 \end{cases}$$

peremérték-feladatot, ahol $a(x)$ és $b(x)$ $C[a, b]$-beli függvények. Alkalmazzunk egy másodrendű véges differenciák módszerén alapuló diszkretizációt, majd írjuk fel a kapott lineáris egyenletrendszert! $\Longrightarrow$

9.16. (⊞) Határozzuk meg $h = 1/5$ lépésköz mellett véges differenciák módszerével az

$$\begin{cases} u''(x) + xu'(x) + x^2u(x) = 10x, \quad x \in (0, 1) \\ u(0) = 1, \ u(1) = 2 \end{cases}$$

peremérték-feladat megoldását az $x = 0.8$ pontban! $\Longrightarrow$

9.17. (⊟) Írjunk olyan MATLAB programot, amely az

$$\begin{cases} u''(x) + t\cos(x)u(x) = 0, \quad x \in (0, 1) \\ u(0) = 0, \ u(1) = 1 \end{cases}$$

feladatot véges differencia módszerrel megoldja! Adjuk meg $h = 10^{-2}$ lépésköz esetén a megoldás numerikus értékét az $x = 0.92$ pontban! $\Longrightarrow$

9.18. (⊞) Alkalmazzuk a kpep.m fájlt úgy, hogy megoldja az

$$\begin{cases} u''(x) = 5x^3, \quad x \in (-4, 4) \\ u(-4) = -256, \ u(4) = 256 \end{cases}$$

feladatot véges differencia módszerrel! Módosítsuk a fájlt úgy, hogy ábrázolja a feladat pontos megoldását és a numerikus értékeket $h = 10^{-1}$ lépésköz esetén! $\longrightarrow$

9.19. (⊞) Oldjuk meg az

$$\begin{cases} u''(x) - 2u'(x) + u(x) = x + 2, \quad x \in (0, 1) \\ u(0) = 2, \ u(1) = e + \cos(1) \end{cases}$$

feladatot véges differencia módszerrel a kpep2.m fájl segítségével! Határozzuk meg, hogy mekkora a pontos megoldás és a numerikus értékek abszolútértékben vett maximuma a $[0, 1]$ intervallumon $h = 1/17$ lépésköz esetén! $\longrightarrow$

9.20. (⊞) Oldjuk meg az

$$\begin{cases} u''(x) = 2e^x - u(x), \quad x \in (0,1) \\ u(0) = 2,\ u(1) = e + \cos(1) \end{cases}$$

feladatot véges differencia módszerrel! Készítsünk táblázatot a lépésköz és a hiba kapcsolatáról $h = 2^{-1},\ 2^{-2},\ 2^{-3},\ 2^{-4}$ lépésközök esetén! Ezen értékek láttán mire következtethetünk a módszer rendjét illetően? $\longrightarrow$

9.21. (⊟) Tekintsük a klasszikus ágyúgolyó feladatát (a [4] könyv 10.1.1-es példája). Ismeretes, hogy az alábbi peremérték-feladathoz jutunk:

$$\begin{cases} Y''(x) = \dfrac{-g}{v^2}, \quad x \in (0,L) \\ Y(0) = 0,\ Y(L) = 0, \end{cases}$$

ahol g a gravitációs állandó, míg v a konstans sebesség. Írjunk olyan MATLAB programot, amely a fenti feladatot a belövéses módszer segítségével oldja meg! Alkalmazzuk a $h = 0.1,\ h = 0.01,\ h = 0.001$ lépésközű explicit Euler-módszert a kezdetiérték-feladatok megoldására! Vessük össze a kilövési szögeket meghatározó első deriváltak különbségének abszolút értékét a numerikus módszer eredménye és a pontos eredmény ismeretében! $\Longrightarrow$

9.22. (⊟) Módosítsuk a 9.21. feladat megoldására írt agyu.m és belovesesmodszer.m fájlokat úgy, hogy a kezdetiérték-feladat megoldására negyedrendű módszert használjon! $\longrightarrow$

10. fejezet

Parciális differenciálegyenletek

10.1. Képletek, összefüggések

A parciális differenciálegyenletek numerikus módszerei azokat a numerikus megoldási módszereket tárgyalja, amelyekkel a parciális differenciálegyenletek peremérték-feladatát és/vagy kezdetiérték-feladatát numerikusan meg lehet oldani. A vizsgált egyenletek alakja

$$a(x,y)\frac{\partial^2 u(x,y)}{\partial x^2}+2b(x,y)\frac{\partial^2 u(x,y)}{\partial x\partial y}+c(x,y)\frac{\partial^2 u(x,y)}{\partial y^2}+f\left(x,y,u,\frac{\partial u(x,y)}{\partial x},\frac{\partial u(x,y)}{\partial y}\right)=0.$$

Az adott a, b és c függvények határozzák meg az egyenlet típusát, amely lehet elliptikus, parabolikus vagy hiperbolikus. A megfelelő kiegészítő feltételekkel a feladat korrekt kitűzésű és a megoldás speciális esetekben a változók szétválasztásával előállítható.

A numerikus megoldást a véges differenciák módszerével adhatjuk meg. Ennek során a folytonos feladat értelmezési tartományán rácshálót generálunk, a rácsháló pontjaiban az $u(x, y)$ függvény első és második deriváltjait a szokásos véges differenciákkal közelítjük. Így az egyenlet felhasználásával numerikus eljárást konstruálhatunk az adott csomópontbeli ismeretlen értékek közelítésének meghatározására. Alapvető kérdés a konvergencia belátása, azaz annak kimutatása, hogy finomodó rácshálók ($h \to 0$) esetén a numerikus megoldás tart-e (ha igen, akkor milyen rendben) a pontos megoldáshoz. Lineáris feladatok esetén a konvergencia a konzisztencia és a stabilitás segítségével megmutatható.

A feladatok analitikus és numerikus megoldásai eltérően vizsgálhatók az elliptikus és a parabolikus esetekben. Ugyanakkor mindkét esetben a konvergencia belátásához az M-mátrixok tulajdonságait használjuk fel.

A fenti fogalmak részletesen megismerhetők a [4] könyv 11. fejezetéből. Ezen belül az osztályozással a 11.1 szakasz foglalkozik. A folytonos feladat megoldásával elliptikus peremérték-feladatokra téglalap tartomány esetén a 11.2 szakasz, a parabolikus esetre a 11.3

szakasz foglalkozik. Az alaptétellel, amely a konvergenciát bizonyítja, a 11.2.3 szakasz, míg az elliptikus feladatok véges differenciás megoldásának elméletét, illetve realizálását a 11.2.4 és 11.2.5 szakaszok ismertetik. Parabolikus feladatokra a numerikus elmélet és realizálása a 11.3.2-11.3.5 szakaszokban találhatóak meg.

10.2. Feladatok

10.2.1. Elméleti feladatok

10.1. Határozzuk meg, hogy $\mathbb{R}^2$ egyes részein milyen típusú az alábbi differenciáloperátor:

$$(Lu)(x,y) = x\frac{\partial^2 u(x,y)}{\partial x^2} + y\frac{\partial^2 u(x,y)}{\partial y^2}!$$

$\Longrightarrow$

10.2. Határozzuk meg, hogy $\mathbb{R}^2$ egyes részein milyen típusú az alábbi differenciáloperátor:

$$(Lu)(x,y) = (x+y)\frac{\partial^2 u(x,y)}{\partial x^2} + 2\sqrt{xy}\frac{\partial^2 u(x,y)}{\partial x \partial y} + (x+y)\frac{\partial^2 u(x,y)}{\partial y^2}!$$

$\Longrightarrow$

10.3. Határozzuk meg, hogy a Laplace-, Poisson-, hővezetési és hullámegyenletek milyen típusúak $\mathbb{R}^2$ egyes részein! $\longrightarrow$

10.4. Határozzuk meg a

$$\frac{\partial u(x,y)}{\partial y} - \frac{\partial u(x,y)}{\partial x} = 0,\ (x,y) \in \mathbb{R}^2$$

egyenlet klasszikus megoldását! $\Longrightarrow$

10.5. Határozzuk meg a

$$\frac{\partial^2 u(x,y)}{\partial y^2} - \frac{\partial^2 u(x,y)}{\partial x^2} = 0,\ (x,y) \in \mathbb{R}^2$$

egyenlet klasszikus megoldását! $\longrightarrow$

10.6. Oldjuk meg a változók szétválasztásának módszerével a

$$\frac{\partial^2 u(x,y)}{\partial x^2} = \frac{\partial u(x,y)}{\partial y}$$

egyenletet! $\Longrightarrow$

10.2.2. Elliptikus és parabolikus feladatok megoldása véges differenciákkal

10.7. Tekintsük az egységnégyzeten a

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = x^2 + y^2$$

egyenletet az $u(x,0) = 0$, $u(x,1) = x^2/2$, $u(0,y) = \sin(\pi y)$ és $u(1,y) = e^{\pi}\sin(\pi y) + y^2/2$ peremfeltétellel. Írjuk fel a feladat véges differenciás approximációját jelentő lineáris algebrai egyenletrendszer együtthatómátrixát, amikor $N_x = 3$ és $N_y = 2$ osztásrészt veszünk! $\Longrightarrow$

10.8. (⊟) Írjunk olyan MATLAB programot, amely megoldja tetszőleges $N_x = N_y$ osztásrész mellett a 10.7. feladatot! Készítsünk táblázatot, amely az osztásrészek száma és a maximumnormabeli pontosság közötti kapcsolatot mutatja, ha a feladat pontos megoldása $u(x,y) = e^{\pi x}\sin(\pi y) + 0.5x^2y^2$! Ábrázoljuk ezt a kapcsolatot a MATLAB segítségével! $\Longrightarrow$

10.9. (⊟) Tekintsük az egységnégyzeten a

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = e^y(x^2+2)$$

egyenletet az $u(x,0) = x^2$, $u(x,1) = ex^2$, $u(0,y) = 0$ és $u(1,y) = e^y$ peremfeltétellel. Írjunk olyan MATLAB programot, amely tetszőleges osztásrész mellett megoldja a feladatot! Készítsünk táblázatot, amely az osztásrészek száma és a maximumnormabeli pontosság közötti kapcsolatot mutatja, ha a feladat pontos megoldása $u(x,y) = e^y x^2$! Ábrázoljuk ezt a kapcsolatot a MATLAB segítségével! $\Longrightarrow$

10.10. (⊞) Tekintsük a $(0,1)\times(0,1)$ tartományon a

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \quad (x,t) \in (0,1)\times(0,1]$$

egyenletet az $u(x,0) = e^x$, $\quad x \in [0,1]$ kezdeti és az $u(0,t) = e^t$, $\quad u(1,t) = e^{1+t}$, $\quad t \in (0,1]$ peremfeltétellel. Módosítsuk a heatexp.m fájlt úgy, hogy a fenti feladatot megoldja! A hibafüggvény meghatározásához számoljuk ki a feladat pontos megoldását is! $\Longrightarrow$

10.11. (⊟) Tekintsük a

$$\frac{\partial u(t,x)}{\partial t} = \frac{\partial^2 u(t,x)}{\partial x^2}, \quad t \in [0,T],\ x \in [0,\pi]$$

egyenletet az $u(x,0) = \sin(x)$ kezdeti feltétellel és Dirichlet-peremfeltétellel. Írjunk olyan MATLAB programot, amely a fenti feladatot az alábbi bemenő paraméterekkel oldja meg:

n: az osztópontok száma,

T: az időintervallum végpontja,

r: a δ/h^2 rácsparaméterek hányadosa,

θ: a θ-módszer értéke!

A program ábrázolja az egyes $t \in [0, T]$ időpillanatban a megoldást a $[0, \pi]$ intervallumon!
$\Longrightarrow$

Útmutatások, végeredmények

Előismeretek

Nevezetes mátrixtípusok

1.1. Nem diagonalizálható. Nincs három lineárisan független sajátvektora.

1.4. A $\det(\mathbf{A}-\mathbf{E})$ értékről lássuk be, hogy nulla. A zárójelben emeljünk ki $\mathbf{A}$-t, majd alkalmazzuk a determinánsok szorzási szabályát ill. a feladatban szereplő feltételeket!

1.5. A $\overline{\mathbf{v}}$ vektor továbbra is sajátvektor lesz $\lambda(1-\overline{\mathbf{v}}^T\overline{\mathbf{v}})$ sajátértékkel. A többi sajátvektor és sajátérték nem változik.

1.6. A $\overline{\mathbf{g}} = [u(h), u(2h), \ldots, u(nh)]^T$, ahol $u : [0,1] \to \mathbb{R}$, $u(x) = x(1-x)$ és $h = 1/(n+1)$ választással megmutatható, hogy $\mathbf{A}\overline{\mathbf{g}} > \mathbf{0}$, ami már mutatja, hogy M-mátrixról van szó.

1.7. Alkalmazzuk a Gersgorin-tételt!

1.8. Írjuk fel az $\mathbf{M}$ mátrixot $\mathbf{M} = \mu\mathbf{E} - \mathbf{H}$ alakban, ahol μ megfelelő pozitív szám és $\mathbf{H}$ megfelelő nemnegatív mátrix! Mutassuk meg, hogy ez a felbontás reguláris, majd használjuk ki, hogy nemnegatív inverzű mátrixok reguláris felbontásából származó iterációs mátrixok konvergensek, azaz spektrálsugaruk kisebb 1-nél!

1.9. Igazoljuk, hogy minden bal felső sarokdeterminánsok pozitív!

1.10. Próbáljuk ki $n \times n$-es mátrix esetén sajátvektornak a $\overline{\mathbf{v}}_k = \sin(ik\pi h)$ alakú vektorokat, ahol $h = 1/(n+1)$, továbbá $k, i = 1, \ldots, n$!

1.11. A transzponáltjával szorozva, majd egyszerűsítve az egységmátrixot kapjuk.

1.12. Gondoljuk végig, hogy két felső háromszögmátrix szorzása során a szorzatmátrix főátló alatti elemei hogy állíthatók elő! Az inverz esetén használjunk indirekt bizonyítást (tegyük fel, hogy az inverzben van a főátló alatt nemnulla elem)!

1.13. Írjuk fel az egyenlőség két oldalán álló mátrixok főátlóinak elemeit!

Normált és euklideszi terek

1.17. Nullvektorokra triviálisan igaz az állítás. Egyébként pedig vizsgáljuk a $\phi(t) = \langle \mathbf{x} + t\mathbf{y}, \mathbf{x} + t\mathbf{y} \rangle$ nyilvánvalóan nemnegatív függvényt $t \in \mathbb{R}$ esetén!

Banach-féle fixponttétel

1.19. A T leképezésnek van egyértelmű fixpontja (Banach-féle fixponttétel). Ebből megmutatható, hogy F-nek maximum egy fixpontja lehet. Ezen kívül mutassuk meg még azt, hogy T fixpontja F-nek is fixpontja!

1.20. A kontrakciós tulajdonság a Lagrange-féle középértéktétel segítségével mutatható meg. Ennek segítségével látható a legkisebb választható kontrakciós tényező értéke is. A fixpont meghatározásához az $F(x^\star) = x^\star$ egyenletet kell megoldani.

1.21. Az egyértelműséget úgy kell igazolni, mint a Banach-féle fixponttétel bizonyításában. Annak megmutatására, hogy nem feltétlenül van fixpont vizsgáljuk az $F : [1, \infty) \to [1, \infty),\ F(x) = x + 1/x$ függvényt!

1.22. Igazoljuk hogy az $\mathbf{x} \mapsto \mathbf{T}(\mathbf{x}) + \mathbf{y}$ leképezés kontrakció, majd alkalmazzuk a Banach-féle fixponttételt!

1.23. Először igazoljuk, hogy van olyan $0 \leq q < 1$ szám, melyre $|f'(c)| \leq q$ minden $c \in [a, b]$ esetén, majd alkalmazzuk a Lagrange-féle középértéktételt!

Vektornormák

1.24. $\|\overline{\mathbf{x}}\|_1 = 6,\ \|\overline{\mathbf{x}}\|_2 = \sqrt{14},\ \|\overline{\mathbf{x}}\|_\infty = 3.$

1.25. $\|\overline{\mathbf{x}}\|_1 = 5050,\ \|\overline{\mathbf{x}}\|_2 = 581.6786,\ \|\overline{\mathbf{x}}\|_\infty = 100.$

1.28. Mutassuk meg, hogy ezek a normák nem teljesítik a háromszög-egyenlőtlenséget!

1.30. A Young-egyenlőtlenség igazolásához elemi függvényvizsgálati eszközökkel lássuk be, hogy a

$$f(a) = \frac{a^p}{p} + \frac{b^q}{q} - ab$$

függvény nemnegatív a $[0, \infty)$ intervallumon! Ebből már következik az egyenlőtlenség.

1.31. Az első két normaaxióma teljesülése triviális, a harmadik pedig a Minkowski-egyenlőtlenség segítségével igazolható.

Mátrixnormák

1.33. Mutassuk meg pl. hogy ez a norma nem szubmultiplikatív!

1.36. Számítsuk ki az $\mathbf{A}^T\mathbf{A}$ mátrix i-edik sorának főátlóbeli elemét! Mekkora az egységmátrix Frobenius-normája?

1.39. Induljunk ki az $\mathbf{A}\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}$ egyenlőségből, majd szorozzuk ezt jobbról $\overline{\mathbf{v}}^T$-tal!

1.45. Definiáljunk egy vektornormát egy tetszőleges $\overline{\mathbf{y}} \neq \mathbf{0}$ vektor segítségével az alábbi módon: $\|\overline{\mathbf{x}}\| = \|\overline{\mathbf{x}}\overline{\mathbf{y}}^T\|$! Ezzel a vektornormával konzisztens a mátrixnorma.

1.47. Induljunk ki abból, hogy van olyan $\overline{\mathbf{x}} \neq \mathbf{0}$ vektor, melyre $\mathbf{B}\overline{\mathbf{x}} = \mathbf{0}$. Erre az $\overline{\mathbf{x}}$ vektorra:

$$\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\overline{\mathbf{x}} = \overline{\mathbf{x}}!$$

1.48. Azt igazoljuk, hogy tetszőleges pozitív ε számhoz van olyan n_0 index, hogy minden $k > n_0$ esetén

$$\varrho(\mathbf{A}) \leq \|\mathbf{A}^k\|^{1/k} \leq \varrho(\mathbf{A}) + \varepsilon.$$

Ebből ugyanis az állítás már következik.

1.50. A mátrix M-mátrix, így használhatjuk az M-mátrixok inverzére vonatkozó becslést.

Modellalkotás és hibaforrásai

Feladatok kondicionáltsága

2.1. A feladat $d \neq \pm 2$ esetén korrekt kitűzésű. A kondíciószám $2 < |d| < \sqrt{40000/9999}$ esetén lesz 100-nál nagyobb.

2.2. Az első esetben 98.5, a másodikban 0.4975 a kondíciószám.

A gépi számábrázolás

2.12. Nem kapnánk meg. Az adott számrendszerben számoló számítógépen 2.9 lenne az eredmény.

2.13. $-5e - 6$ lenne az eredmény, melynek relatív hibája 0.3612. Elkerülhetjük a nagy realtív hibájú számolást a $\cos(2x)$-re vonatkozó formula használatával.

2.15. Az eltérés $\pi^2/6 - 1.6447253 = 2.0877 \times 10^{-4}$. Jobb eredményt kaphatunk, ha fordított sorrendben adjuk össze a számokat.

2.16.

$$\frac{x - fl(x)}{x} = -\frac{1}{4}\mathtt{u}.$$

Lineáris egyenletrendszerek megoldása

Kondicionáltság

3.1. $\kappa_\infty(\mathbf{A}) \geq 201$, a keresett kondíciószám 404.01.

3.2. $\kappa_1(\mathbf{A}) = \kappa_\infty(\mathbf{A}) = 1.5 \cdot 18 = 27$, és $\kappa_2(\mathbf{A}) = 1.2676/0.0657 = 19.3$.

$$\|\delta\overline{\mathbf{x}}\|_\infty = 0.01\|\mathbf{A}^{-1}\overline{\mathbf{b}}\|_\infty \leq 0.18\|\overline{\mathbf{b}}\|_\infty.$$

3.3. Ha ortogonális, akkor a kondíciószáma 1, de az állítás megfordítása nem igaz. Keressünk rá ellenpéldát!

3.10. Az egyenlőtlenség következik a kondíciószám egyik tulajdonságából, az egyenlőséghez pedig először lássuk be, hogy egy mátrix és transzponáltjának 2-es normája megegyezik!

Direkt módszerek

3.12.

$$\|\delta\overline{\mathbf{x}}\|_\infty/\|\overline{\mathbf{x}}\|_\infty \leq 0.00153.$$

3.13. 0.1035.

3.19. $n^3 + n^2 - 4n + 3$.

3.32. A feltételek mellett a szereplő $\mathbf{Q}$ és $\mathbf{R}$ mátrixok nemszingulárisak. A $\mathbf{Q}_1\mathbf{R}_1 = \mathbf{Q}_2\mathbf{R}_2$ egyenlőségből $\mathbf{R}_1\mathbf{R}_2^{-1} = \mathbf{Q}_1^T\mathbf{Q}_2$ következik. Vizsgáljuk meg az egyenlőség két oldalán álló mátrixok szerkezetét!

Iterációs módszerek

3.34. Az ω paraméter értékének a (0,2) intervallumba kell esnie. Az $\omega = 1$ választás esetén lesz a leggyorsabb a konvergencia.

3.35. Legfeljebb 20 lépés kell az adott pontosság eléréséhez.

3.43. Rendezzük át az egyenletrendszer sorait úgy, hogy diagonálisan domináns mátrixú egyenletrendszert kapjunk!

Sajátérték-feladatok numerikus megoldása

Sajátértékbecslések

4.1. Használjuk közvetlenül a Gersgorin-tételeket!

4.2. Alkalmazzuk a Bauer–Fike-tételt, vagy számítsuk ki az $\mathbf{S}^{-1}(\mathbf{A}+\varepsilon\mathbf{B})\mathbf{S}$ mátrixot, ahol $\mathbf{S}$ az $\mathbf{A}$ mátrixot diagonalizáló mátrix, majd alkalmazzuk a Gersgorin-tételt!

4.3. Alkalmazzuk közvetlenül a Gersgorin-tételt!

4.4. Permutációs mátrixszal végzett hasonlósági transzformáció segítségével hozzuk a mátrixot blokkdiagonális alakra! Ekkor a mátrix sajátértékei a főátlóban álló négyzetes mátrixok sajátértékei lesznek.

4.6. A Rayleigh-hányadossal kell megszorozni, hogy legközelebb legyen hozzá.

A hatványmódszer és változatai

4.12. A legjobb választás kb. 12.5.

Nemlineáris egyenletek és egyenletrendszerek megoldása

Sorozatok konvergenciarendje, hibabecslése

5.1. Vizsgáljuk meg, hogy az a_{k+1}/a_k^r hányados milyen r esetén marad korlátos! Mindkét sorozat konvergenciarendje 1.

5.2. Az elsőnek 2, a másodiknak 1.

5.3. A konvergencia negyedrendű.

5.5. Alkalmazzuk a Lagrange-féle középértéktételt az x és $x^\star$ pontokban!

Zérushelyek lokalizációja

5.6. Az intervallum két végpontjában ellenkező a függvény előjele, deriváltja pedig pozitív.

5.7. Egy zérushely van a [0,2] intervallumban.

5.8. Három zérushely van rendre a [0,1/3], [1/3,1] és [1,2] intervallumok belsejében.

5.10. Használjuk az 5.2. tételt!

Intervallumfelezési módszer

5.11. Alkalmazzuk az 5.3. tételt! 8 lépés elég, $x_8 = 1.3828125$.

5.12. Alkalmazzuk az 5.3. tételt! 3 lépés elég. $x_3 = 2.9375$.

Newton-módszer

5.15. Alkalmazzuk az 5.6. tételben szereplő hibabecslést!

5.22. Az ok az, hogy a függvény zérushelye kétszeres zérushely, azaz a deriváltja is nulla a zérushelynél. A módszer másodrendűvé tehető a

$$x_{k+1} = x_k - 2\frac{f(x_k)}{f'(x_k)}$$

módosítással. A másodrend pl. úgy igazolható, hogy a fenti iteráció minkét oldalából $\mathbf{x}^\star$-t kivonunk, majd mindkét oldalt $f'(x_k)$-val szorozzuk. Ezután a bal oldalon $f'(x_k)$-t az elsőrendű tagig, a jobb oldalon pedig magát az egész jobb oldalt a harmadrendű tagig sorbafejtjük $\mathbf{x}^\star$ körül.

5.23. A másodrend pl. úgy igazolható, hogy a fenti iteráció minkét oldalából $\mathbf{x}^\star$-t kivonunk, majd mindkét oldalt $f'(x_k)$-val szorozzuk. Ezután a bal oldalon $f'(x_k)$-t az $(m-1)$-edrendű tagig, a jobb oldalon pedig magát az egész jobb oldalt az $(m+1)$-ed rendű tagig sorbafejtjük $\mathbf{x}^\star$ körül.

5.24. Írjuk fel $f(x)$-et $f(x) = (x - x^\star)^m h(x)$ alakban és $f'(x)$-et $f'(x) = (x - x^\star)^{m-1} k(x)$ alakban!

5.25. A Newton-módszer az adott pontból nem használható, mert ciklikusan ismétlődő sorozatot állít elő.

Fixpont iterációk

5.30. Az iteráció indítható pl. a [-0.5,0.5] intervallumból. A konvergencia harmadrendű.

5.31. Az $A = 1/4$, $B = -5/8$ választással a konvergencia harmadrendű lesz.

5.35. Az első elsőrendben, a második másodrendben konvergál, a harmadik pedig nem konvergens.

Nemlineáris egyenletrendszerek megoldása

5.37. Alkalmazzuk az 5.9. tételt!

5.41. Alkalmazzuk az 5.10. tételt!

Interpoláció és approximáció

Polinominterpoláció

6.1.

$$\frac{5}{6}x^3 - \frac{9}{2}x^2 + \frac{8}{3}x + 10.$$

6.5. Vezessük le az ún. baricentrikus interpolációs formulát!

6.8. Használjuk ki, hogy $s \leq n$ esetén az (x_k, x_k^s) $(k = 0, \ldots, n)$ pontokra illesztett polinom éppen az $L_n(x) = x^s$ polinom lesz!

6.12.

$$c_k = \frac{\sum_{i=0}^{k} \binom{k}{i} (-1)^i f_{k-i}}{h^k k!}.$$

6.25. Alkalmazzuk a 6.6. tételt!

Trigonometrikus interpoláció

6.33.

$$t(x) = 1 + \frac{2}{\sqrt{3}} \sin x.$$

Approximáció polinomokkal és trigonometrikus polinomokkal

6.37. $y = -0.5x + 1.75$.

6.38. $y = -x^2/2 + x + 3/2$.

6.41. Az interpolációs polinom legfeljebb elsőfokú részletösszege lesz a legjobban közelítő polinom.

Numerikus deriválás és numerikus integrálás

Numerikus deriválás

7.2. A feladatban szereplő kifejezés az első deriváltat negyedrendben approximálja és hibája:

$$-\frac{h^4}{30}f^{(5)}(x_0)+O(h^5).$$

7.3. A feladatban szereplő kifejezés a második deriváltat negyedrendben approximálja és hibája:

$$-\frac{h^4}{90}f^{(6)}(x_0)+O(h^6).$$

7.6. A kifejezés felső határoló függvénye ϵ pontosságú adatok esetén:

$$g(h)=\frac{h^4}{30}M_5+\frac{3\epsilon}{2h},$$

ahol $M_5=\sup|f^{(5)}(x)|$. Az optimális lépésköz:

$$h_{\text{opt}}=\sqrt[5]{\frac{45\epsilon}{4M_5}}.$$

7.7. A centrális differencia felső határoló függvénye ϵ pontosságú adatok esetén:

$$g(h)=\frac{h^2}{12}M_4+\frac{4\epsilon}{h^2},$$

ahol $M_4=\sup|f^{(4)}(x)|$. Az optimális lépésköz:

$$h_{\text{opt}}=\sqrt[4]{\frac{48\epsilon}{M_4}}.$$

7.10. A feladatban szereplő módszerekről az alábbiak mondhatóak el:

(a) A módszer az $f'(1)$-et közelíti a 0.494 értékkel.

(b) A módszer egyetlen derivált értéket sem közelít.

(c) A módszer egyetlen derivált értéket sem közelít.

(d) A módszer az $f''(1)$-et közelíti a -0.236 értékkel.

(e) A módszer az $f'''(1)$-et közelíti a 0.24 értékkel.

(f) A módszer egyetlen derivált értéket sem közelít.

Numerikus integrálás

7.14. Az integrál pontos értéke:

$$I(f) = \int_0^1 \frac{1}{1+x^2}\,\mathrm{d}x = \frac{\pi}{4} \approx 0.7853981634.$$

A MATLAB programcsomag a

$$\texttt{quad}('1./(1+\text{x.}\hat{}2)', 0, 1)$$

parancs esetén is ezt az értéket adja. A MATLAB programcsomag segítségével az alábbi eredményeket adják a tanult összetett szabályok:

n	$\lvert I(f) - I_{\mathrm{E}}(f)\rvert$	$\lvert I(f) - I_{\mathrm{Tr}}(f)\rvert$	$\lvert I(f) - I_{\mathrm{Simp}}(f)\rvert$
32	$2.0345051636 \cdot 10^{-5}$	$4.0690103704 \cdot 10^{-5}$	$9.2391649886 \cdot 10^{-12}$
64	$5.0862630135 \cdot 10^{-6}$	$1.0172526034 \cdot 10^{-5}$	$1.4421797089 \cdot 10^{-13}$
128	$1.2715657553 \cdot 10^{-6}$	$2.5431315102 \cdot 10^{-6}$	$2.5535129566 \cdot 10^{-15}$
256	$3.1789143839 \cdot 10^{-7}$	$6.3578287790 \cdot 10^{-7}$	$1.1102230246 \cdot 10^{-16}$

10.1. táblázat. Hibaértékek adott n és adott módszer esetén.

A táblázat eredményeiből leolvasható, hogy az összetett érintő- és trapézformula az intervallumszám duplázásával (avagy ennek megfelelően a lépésköz felezésével) a hiba negyedelődik. Ezt azt jelenti, hogy a két módszer konvergenciarendje 2, amely megfelel az elméletből ismert ténynek.

Az összetett Simpson-formula esetében a hiba az intervallumszám duplázásával tizenhatod részére csökken, azaz a módszer megfelel a korábban ismert ténynek, miszerint a módszer negyedrendben konvergens.

7.15. Az összetett trapézformula 23 részre történő osztás esetén a 7.15. feladatban szereplő integrált a 4 értékkel közelíti. Ehhez a MATLAB-ban az alábbi parancsokat kell beírni:

```
>> x = linspace(-2,2,23);
>> y=x.^5-3*x.^3+2*x+1;
>> trapz(x,y)

ans =

   4.000000000000000
```

7.17. A módosítás eredménye az alábbi osszerinto.m forráskódhoz vezet:

```
function osszerinto(a,b,n,fv)

format long
h=(b-a)/n;
 fprintf('\n');
 disp('A feladat megoldása összetett érint■formulával.')

 x=[a:h/2:b];
 y=eval(fv);
 ((b-a)/n)*sum(y(2:2:2*n))
```

7.19. A zárt Newton–Cotes-formulák esetében tudjuk, hogy a formula súlyai a Lagrange-féle alappolinomok $[a,b]$ intervallumon vett integráljai lesznek, azaz

$$a_k = \int_a^b l_k(x)\,\mathrm{d}x,\ \ k = 0,\dots,n.$$

A súlyokat kézzel is meghatározhatjuk, de használhatjuk a MATLAB `int` parancsát a polinomok integráláshoz. Ekkor a [0,1] intervallum esetén az alábbi együtthatókat kapjuk:

$N^{4,k}$	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
$n=4$	$\frac{7}{90}$	$\frac{32}{90}$	$\frac{12}{90}$	$\frac{32}{90}$	$\frac{7}{90}$

10.2. táblázat. A zárt $N^{4,k}$ Newton–Cotes együtthatói.

A módszer (ún. Boole-formula) az [a,b] intervallumon az alábbi módon realizálódik:

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{(b-a)}{90}\Big(7f(a)+32f(x_1)+12f(x_2)+32f(x_3)+7f(b)\Big),$$

ahol $x_i = a + i(b-a)/n,\ i = 1, \ldots, 3$.

7.22. A 7.21. feladatban használt gondolatmenet alapján hasonlóan meghatározható a Gauss–Csebisev-kvadratúra képlete. Ekkor a formula nem más, mint

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,\mathrm{d}x \approx \frac{\pi}{3}\Big(f(-\sqrt{3}/2) + f(0) + f(\sqrt{3}/2)\Big),$$

amely pontos lesz minden legalább ötödfokú polinomra.

7.27. A Gauss-függvény integrálja a $[0, 1]$ intervallumon 0.842700793. Ekkor a Romberg-módszerrel számított közelítő integrál értékei az alábbiak:

0.77174333				
0.82526296	0.84310283			
0.83836778	0.84273605	0.84271160		
0.84161922	0.84270304	0.84270083	0.84270666	
0.84243051	0.84270093	0.84270079	0.84270079	0.84270079

10.3. táblázat. A Romberg-módszer értékei.

7.28. Útmutatás: A MATLAB-ban két `for` ciklus segítségével a program előállítható. Előbbiben a Crank–Nicolson-módszert, utóbbiban a Richardson-extrapolációt állítjuk elő.

A közönséges differenciálegyenletek kezdetiérték-feladatainak numerikus módszerei

Egylépéses módszerek

8.1. A feladatban szereplő módszerek konzisztenciarendjei az alábbiak:

(a) Az explicit Euler-módszer elsőrendben konzisztens.

(b) Az implicit Euler-módszer elsőrendben konzisztens.

(c) A Crank–Nicolson-módszer másodrendben konzisztens.

(d) A θ-módszer $\theta \neq 1/2$ esetén elsőrendben, míg $\theta = 1/2$ esetén másodrendben konzisztens.

8.2. A számítás eredményeit az alábbi táblázatban foglaltuk össze:

h	EE	IE	CN	JE	EH
1/2	-25.50000000	0.09992283	0.09662640	-5.21906250e+002	-5.21906250e+002
1/4	-2.46289062	0.09999555	0.09999999	-4.76213398	-4.76213398
1/8	0.099999999	0.09999976	0.09999999	0.09999597	0.09999597
1/16	0.099999999	0.09999998	0.09999999	0.09999999	0.09999999

10.4. táblázat. A numerikus értékek adott módszer és lépésköz mellett.

8.3. A feladatban szereplő módszerek adott lépésközű eredményei az alábbi táblázatban foglalható össze:

h	EE	IE	CN	JE	EH
1/2	3.33333333	6.00000000	4.00000000	3.87619047	3.79166666
1/4	3.60000000	4.66666666	4.00000000	3.96179918	3.93042304
1/8	3.77777777	4.28571428	4.00000000	3.98940000	3.97979547
1/16	3.88235294	4.13333333	4.00000000	3.99721170	3.99455697

10.5. táblázat. A numerikus értékek adott módszer és lépésköz mellett.

h	EE	IE	CN	JE	EH
1/2	5.49401855	7.71404151	6.42957783	6.29456039	6.34371731
1/4	5.88458254	6.94662094	6.37475602	6.33984573	6.35441025
1/8	6.10866555	6.63405317	6.36130951	6.35241232	6.35636471
1/16	6.22946604	6.49146658	6.35796378	6.35571693	6.35674557

10.6. táblázat. A numerikus értékek adott módszer és lépésköz mellett.

8.4. A feladatban szereplő módszerek adott lépésközű eredményei az alábbi táblázatban foglalhatóak össze:

8.5. A feladatban szereplő módszerek adott lépésközű eredményei az alábbi táblázatban foglalhatóak össze:

h	EE	JE	EH
1/2	2.95715863	2.73260420	2.77630626
1/4	2.85177621	2.75627855	2.76661978
1/8	2.80544980	2.76142191	2.76394025
1/16	2.78375834	2.76262203	2.76324360

10.7. táblázat. A numerikus értékek adott módszer és lépésköz mellett.

8.6. Programozzuk le a tanult módszereket (expliciteuler.m, eulerheun.m, javitotteuler.m)! Segítségképpen megadjuk az eulerheun.m fájl forráskódját, amely magától értetődő módon módosítható a másik két módszerre.

```
function eulerheun(a,b,t0,y0,N)

%% Bemen■ paraméterek listája

% a           az intervallum kezdete
```

```
% b            az intervallum vége
% t0           a kezdeti id■pont
% y0           a kezdeti érték
% N            a lépésközök száma


%% Kimen■ paraméter

% y            a numerikus megoldás vektora
% y(N+1)       a numerikus megoldás


%% El■készületek

h=(b-a)/N;                      %lépésköz
x=linspace(a,b,N+1);            % az intervallum felosztása
y=zeros(1,N+1);                 % numerikus megoldas vektora


%% Az Euler-Heun-módszer algoritmusa

y(1)=y0;
t(1)=t0;
for j=1:N
        y(j+1)=y(j)+h/2*[f(a+(j-1)*h, y(j))]
        +h/2*[f(a+j*h,y(j)+h*f(a+(j-1)*h, y(j)))];
end

%y;
y


%% Az f, vagyis az y'(t)=f(t,y(t)) egyenlet jobboldala
function ered=f(t,y)

ered=y+t*cos(t);
```

8.7. Útmutatás: használjuk a MATLAB help funkcióját (`help ODE45`) az ODE45 módszer alkalmazásához és tanulmányozzuk a függvény működését! A numerikus megoldás az alábbi módon határozható meg a 8.2. feladatra:

Először elkészítünk egy `odefun.m` m-fájlt, amely definiálja a differenciálegyenlet jobb

oldalát:

```
function F=odefun(t,y)
F=1-10*y;
```

Ezek után a parancsorból az alábbi módon futtatható a módszer:

```
[t,y]=ode45('odefun',[0,2],0)
```

Ha kimenő paraméterek nélkül futtatjuk a parancsot, akkor a megoldásfüggvény grafikonjának közelítését kapjuk vissza. A parancs első argumentuma definiálja a differenciálegyenletet, a második a megoldási intervallum, és a harmadik a kezdeti feltétel.

8.9. Egy- és többdimenziós Taylor-sorfejtést alkalmazva az alábbi egyenletrendszerhez juthatunk:

$$\begin{array}{ll} I. & 1 - c_1 - c_2 = 0, \\ II./a & 1/2 - ac_1 = 0, \\ II./b & 1/2 - bc_2 = 0. \end{array}$$

Az I.-es egyenlet az első, míg a II./a és II./b egyenletek a másodrendű konzisztencia szükséges feltételei. A sorfejtés alkalmazása után kapott hibatag esetén látható, hogy a módszer nem lehet harmadrendű. A fenti egyenletrendszernek eleget tesznek például a javított Euler ($c_1 = 0, c_2 = 1, a = 1/2, b = 1/2$) és az Euler–Heun-módszerek ($c_1 = 1/2$, $c_2 = 1/2, a = 1, b = 1$).

8.12. A megadott Butcher-táblázatokból felírt módszerek az alábbiak:

(a) $k_1 = f(t_n, y_n)$
$k_2 = f(t_n + \frac{h}{2}, y_n + h(\frac{1}{4}k_1 + \frac{1}{4}k_2))$
$k_3 = f(t_n + h, y_n + hk_2)$

Azaz a módszer alakja: $y_{n+1} = y_n + h(1/6k_1 + 2/3k_2 + 1/6k_3)$.

(b) $k_1 = f(t_n, y_n)$
$k_2 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1))$
$k_3 = f(t_n + h, y_n + h(-k_1 + 2k_2))$

Azaz a módszer alakja: $y_{n+1} = y_n + h(1/6k_1 + 2/3k_2 + 1/6k_3)$.

(c) $k_1 = f(t_n, y_n)$
$k_2 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1))$
$k_3 = f(t_n + \frac{h}{2}, y_n + h(\frac{1}{4}k_1 + \frac{1}{4}k_2))$
$k_4 = f(t_n + h, y_n + h(-k_2 + 2k_3))$

Azaz a módszer alakja: $y_{n+1} = y_n + h(1/6k_1 + 2/3k_3 + 1/6k_4)$.

(d) $k_1 = f(t_n + \frac{h}{3}, y_n + \frac{h}{3}k_1)$
$k_2 = f(t_n + h, y_n + hk_1)$

Azaz a módszer alakja: $y_{n+1} = y_n + h(3/4k_1 + 1/4k_2)$.

8.13. A megadott Butcher-táblázatokból felírt módszerek az alábbiak:

(a) $k_1 = f(t_n + h, y_n + h)$

Azaz a módszer alakja: $y_{n+1} = y_n + hk_1$.

(b) $k_1 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2})$

Azaz a módszer alakja: $y_{n+1} = y_n + hk_1$.

(c) $k_1 = f(t_n, y_n)$
$k_2 = f(t_n + h, y_n + h(\frac{1}{2}k_1 + \frac{1}{2}k_2))$

Azaz a módszer alakja: $y_{n+1} = y_n + h(1/2k_1 + 1/2k_2)$.

8.14. A 8.10. feladat módszereinek konzisztenciarendje:

(a) A módszer elsőrendben konzisztens.

(b) A módszer másodrendben konzisztens.

(c) A módszer másodrendben konzisztens.

A 8.11. feladat negyedrendben konzisztens.

8.15. A feladatban szereplő módszerek Butcher-táblázatai:

8.16. Az adott módszerek stabilitásfüggvényei az alábbiak:

(a) explicit Euler: $R(z) = 1 + z$,

(b) implicit Euler: $R(z) = \dfrac{1}{1 - z}$.

(c) Crank–Nicolson: $R(z) = \dfrac{2 + z}{2 - z}$.

(d) θ-módszer: $R(z) = \dfrac{1 + z + \theta}{1 - z\theta}$.

(e) javított Euler: $R(z) = 1 + z + \dfrac{z^2}{2}$.

(a)
$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

(b)
$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}$$

(c)
$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 \\ 2/3 & 0 & 2/3 & 0 \\ \hline & 1/4 & 0 & 1/4 \end{array}$$

(d)
$$\begin{array}{c|cc} 0 & 0 & 0 \\ 2/3 & 2/3 & 0 \\ \hline & 1/4 & 3/4 \end{array}$$

(e)
$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1-\theta & \theta \\ \hline & 1-\theta & \theta \end{array}$$

(f) Euler–Heun: $R(z) = 1 + z + \dfrac{z^2}{2}$.

(g) implicit középpontszabály: $R(z) = \dfrac{2+z}{2-z}$.

8.17. A 8.16. feladat stabilitásfüggvényeinek segítségével meghatározhatjuk, hogy mely módszerek A-stabilak. Ezek figyelembevételével az alábbiakat mondhatjuk:

A-stabilak: Implicit Euler, Crank–Nicolson, θ-módszer $\theta \in [1/2, 1]$, implicit középpontszabály.

Nem A-stabilak: Explicit Euler, θ-módszer $\theta \in [0, 1/2)$, javított Euler, Euler–Heun.

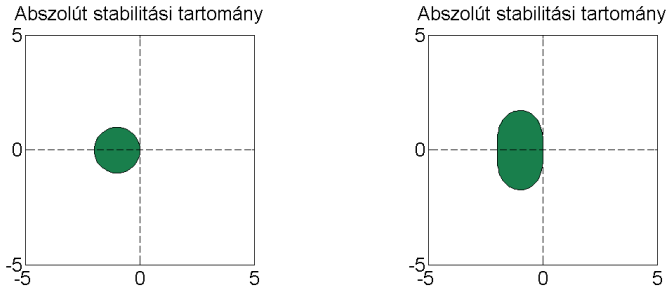
8.18. A stabilitási tartományt a stabilitási függvények segítségével határozhatjuk meg. Ezek eredményei a 8.16. feladathoz tartozó Útmutatások, végeredmények fejezetben megtalálhatóak.

Ekkor feladatunk nem lesz más, mint az egyes stabilitási függvények beprogramozása. A feladatot MATLAB-ban megoldó fájl: Astabilitas.m.

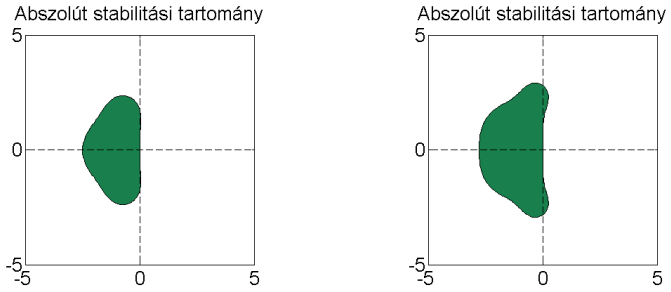
A futtatás eredményeként a $[-5, 5] \times [-5, 5]$ négyzeten ábrázolja a program a stabilitási tartományt. A program tanulmányozása során könnyen észrevehető módon a szükséges módszer kivételével a többit kommentelve az alábbi parancsot írjuk be a futtatáshoz:

```
>> Astabilitas
```

Ekkor az egyes Runge–Kutta-módszerekre az alábbi stabilitási tartományokat nyerjük vissza, melyből rögtön leolvasható az elméletből ismeretes tény, nevezetesen az, hogy explicit Runge–Kutta-módszer sosem A-stabil.



10.1. ábra. Az ERK1 és ERK2 módszerek abszolút stabilitási tartományai.



10.2. ábra. Az ERK3 és ERK4 módszerek abszolút stabilitási tartományai.

8.24. Útmutatás: alkalmazzuk a feladatra a módszerek stabilitási tartományaira vonatkozó ismereteinket.

Többlépéses módszerek

8.25. Taylor-sorfejtés után az alábbi konzisztenciarendek állapíthatóak meg:

(a) A módszer másodrendben konzisztens.

(b) A módszer másodrendben konzisztens.

(c) A módszer harmadrendben konzisztens.

8.26. A konzisztencia feltételek ellenőrzése után az alábbi rendek állapíthatóak meg:

(a) A módszer másodrendben konzisztens.

(b) A módszer másodrendben konzisztens.

(c) A módszer másodrendben konzisztens.

(d) A módszer harmadrendben konzisztens.

8.28. A módszer maximális konzisztenciarendje 3. Az ismeretlen együtthatók az alábbiak:

$$b_0 = 23/12,\ b_1 = -4/3,\ b_2 = 5/12.$$

8.30. A gyökkritériumhoz szükséges feltételek vizsgálata után az alábbiak mondhatóak el:

(a) A módszer nem teljesíti a gyökkritériumot.

(b) A módszer teljesíti a gyökkritériumot.

(c) A módszer teljesíti a gyökkritériumot.

(d) A módszer teljesíti a gyökkritériumot.

(e) A módszer nem teljesíti a gyökkritériumot.

8.31. Az erős stabilitáshoz szükséges feltételek vizsgálata után az alábbiak állapíthatók meg:

A 8.25. feladat eredményei:

(a) A módszer erősen stabil.

(b) A módszer erősen stabil.

(c) A módszer nem erősen stabil.

A 8.26. feladat eredményei:

(a) A módszer erősen stabil.

(b) A módszer erősen stabil.

(c) A módszer nem erősen stabil.

(d) A módszer erősen stabil.

A 8.30. feladat eredményei:

(a) A módszer nem erősen stabil.

(b) A módszer nem erősen stabil.

(c) A módszer erősen stabil.

(d) A módszer erősen stabil.

(e) A módszer nem erősen stabil.

A közönséges differenciálegyenletek peremérték-feladatainak numerikus módszerei

Peremérték-feladatok megoldhatósága

9.2. A kétpontos peremérték-feladat megoldása az $u(x) = \frac{-e}{1-e^2}e^x + \frac{e}{1-e^2}e^{-x}$.

9.4. A peremérték-feladatra az alábbi válaszok jelenthetőek ki:

(a) Igaz.

(b) Hamis.

(c) Hamis.

9.5. A feladatnak tetszőleges (α, β) pár mellett létezik egyértelmű megoldása. Nevezetesen:

$$u(x) = \alpha e^x + \frac{\beta - \alpha e}{e} x e^x.$$

9.6. Az egyértelműségi kérdésre adott válaszok:

(a) Van egyértelmű megoldása.

(b) Nincs megoldása, így nincs egyértelmű megoldása sem.

9.7. A kérdésre adott válaszok:

(a) Van egyértelmű megoldása.

(b) Van egyértelmű megoldása.

(c) Van egyértelmű megoldása.

9.8. A peremérték-feladat elsőrendű rendszerének és peremfeltételeit tartalmazó alakjai a kitűzött feladatok esetén az alábbiak:

(a) Az elsőrendű rendszer alakja:

$$\mathbf{u}'(x) = A\mathbf{u}(x) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u_1(x) \\ u_2(x) \end{pmatrix}.$$

A feladat peremfeltétele:

$$B_a\mathbf{u}(a) + B_b\mathbf{u}(b) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1(a) \\ \mathbf{u}_2(a) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1(b) \\ \mathbf{u}_2(b) \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \mathbf{v}.$$

(b) Az elsőrendű rendszer alakja:

$$\mathbf{u}'(x) = A\mathbf{u}(x) = \begin{pmatrix} 0 & 1 \\ \lambda^2 & \lambda \end{pmatrix} \begin{pmatrix} u_1(x) \\ u_2(x) \end{pmatrix}.$$

A feladat peremfeltétele:

$$B_a\mathbf{u}(0) + B_b\mathbf{u}(1) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1(0) \\ \mathbf{u}_2(0) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1(1) \\ \mathbf{u}_2(1) \end{pmatrix} = \begin{pmatrix} 5 \\ 8 \end{pmatrix} = \mathbf{v}.$$

(c) Az elsőrendű rendszer alakja:

$$\mathbf{u}'(x) = A\mathbf{u}(x) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2\lambda^3 & \lambda^2 & 2\lambda \end{pmatrix} \begin{pmatrix} u_1(x) \\ u_2(x) \\ u_3(x) \end{pmatrix}.$$

A feladat $B_0\mathbf{u}(0) + B_1\mathbf{u}(1) = \mathbf{v}$ peremfeltétele:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1(0) \\ \mathbf{u}_2(0) \\ \mathbf{u}_3(0) \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1(1) \\ \mathbf{u}_2(1) \\ \mathbf{u}_3(1) \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

9.9. A peremérték-feladatok megoldhatóságára az alábbi állítások érvényesek:

(a) A feladat pontosan akkor oldható meg egyértelműen, ha $b \neq k\pi$, $k \in \mathbb{Z}$.

(b) A feladat pontosan akkor oldható meg egyértelműen, ha $b \neq 0$.

Véges differenciák módszere és a belövéses módszer

9.12. A feladatra alkalmazott standard véges differenciás közelítés után az alábbi alakot kapjuk:

$$-\frac{y_h(x_i+h)-2y_h(x_i)+y_h(x_i-h)}{h^2}+c(x_i)y_h(x_i)=f(x_i),\ x_i\in\omega_h$$

$$y_h(x_0)=\mu_1,\ y_h(x_N)=\mu_2.$$

A fenti alakból az $L_hw_h=b_h$ operátoregyenletes alak származtatható, ahol az $L_h:\mathbb{F}(\overline{\omega}_h)\to\mathbb{F}(\overline{\omega}_h)$ operátor egy tetszőleges $w_h\in\mathbb{F}(\overline{\omega}_h)$ rácsfüggvény esetén az alábbi módon hat:

$$(L_hw_h)(x_i)=\begin{cases}-\dfrac{w_h(x_{i+1})-2w_h(x_i)+w_h(x_{i-1})}{h^2}+c(x_i)w_h(x_i), & x_i\in\omega_h\\ w_h(x_0), & x_0=0\\ w_h(x_N), & x_N=l.\end{cases}$$

A $b_h\in\mathbb{F}(\overline{\omega}_h)$ (jobb oldal és a peremértékek) az alábbi alakban írható:

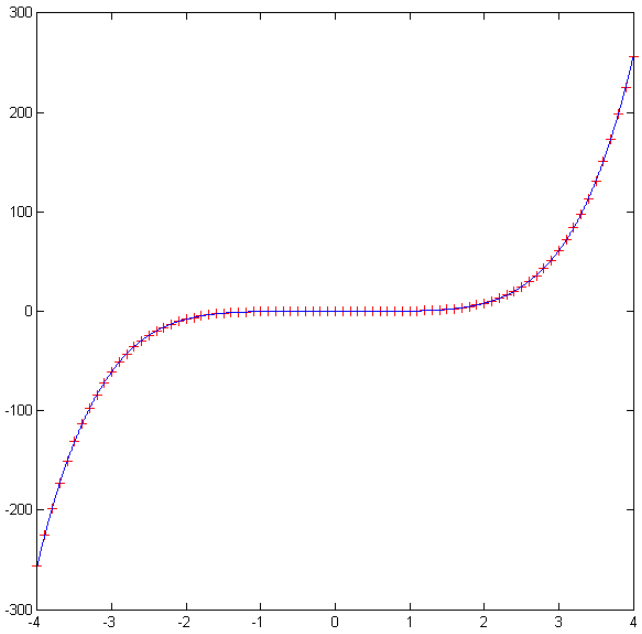
$$\tilde{b_h}(x_i)=\begin{cases}f(x), & x_i\in\omega_h\\ \mu_1, & x_i=x_0\\ \mu_2, & x_i=x_N.\end{cases}$$

9.14. Mutassuk meg, hogy a numerikus feladat diszkretizáló L_h operátorának (amely ekvivalens az A_h mátrixszal) inverze maximum normában korlátos! Ekkor a kívánt eredmény definíció alapján könnyen igazolható.

9.18. A feladat pontos megoldása $u(x)=x^5/4$. Ekkor a kpep2.m fájl módosítása után a pontos megoldás és a numerikus eredmények a 10.3 ábrán látható módon viszonyulnak egymáshoz.

9.19. A feladat pontos megoldása $u(x)=e^x+xe^x+x+2$. A kpep2.m fájl módosítása után a feladat pontos megoldásának és a numerikus megoldás különbségének abszolút értékben vett maximuma a $[0,1]$ intervallumon $h=1/17$ lépésköz mellett 0.26358552.

9.20. A feladat pontos megoldása $u(x)=2e^x+\cos(1)$. A kpep2.m fájl módosítása után az alábbi táblázatban a feladat pontos megoldásának és a numerikus megoldás különbségének abszolút értékben vett maximumát láthatjuk a $[0,1]$ intervallumon a megadott h lépésközök mellett. Azaz a kívánt közelítésnek megfelelően a hiba is másodrendben csökken.



10.3. ábra. A feladat pontos és véges differenciás megoldása $h = 0.1$ esetén a $[-4, 4]$ intervallumon.

h	A hiba értéke
2^{-1}	1.13531898
2^{-2}	0.00185654
2^{-3}	0.00046245
2^{-4}	0.00011551

9.22. Útmutatás: Nézzük meg az agyu.m és a belovesesmodszer.m fájlok forráskódjait!

Ekkor arra a következtetésre juthatunk, hogy az agyu.m fájl oldja meg a kezdetiérték-feladatot. Ennek negyedrendű megoldására programozzuk be az RK4 módszert vagy használhatjuk a MATLAB beépített ODE45 megoldóját is!

Parciális differenciálegyenletek

Elméleti feladatok

10.3. A feladatban szereplő operátorok $\mathbb{R}^2$ egyes részein az alábbi típusúak:

(a) A Laplace-egyenlet $\mathbb{R}^2$-en elliptikus típusú.

(b) A Poisson-egyenlet $\mathbb{R}^2$-en elliptikus típusú.

(c) A hővezetési egyenlet $\mathbb{R}^2$-en parabolikus típusú.

(d) A hullámegyenlet $\mathbb{R}^2$-en hiperbolikus típusú.

10.5. A 10.4. feladatban bevezetett gondolatot használva nyerjük, hogy

$$\frac{\partial^2 u(x,y)}{\partial x^2} = \frac{\partial^2 U(\xi,\eta)}{\partial \xi^2} + 2\frac{\partial^2 U(\xi,\eta)}{\partial \xi \partial \eta} + \frac{\partial^2 U(\xi,\eta)}{\partial \eta^2},$$

$$\frac{\partial^2 u(x,y)}{\partial y^2} = \frac{\partial^2 U(\xi,\eta)}{\partial \xi^2} - 2\frac{\partial^2 U(\xi,\eta)}{\partial \xi \partial \eta} + \frac{\partial^2 U(\xi,\eta)}{\partial \eta^2}.$$

Így feladatunk alakja:

$$\frac{\partial^2 U(\xi,\eta)}{\partial \xi \partial \eta} = 0.$$

Ennek megoldása $U(\xi,\eta) = C(\xi) + D(\eta)$. Azaz az eredeti feladat megoldása:

$$u(x,y) = C(x+y) + D(x-y), \quad C, D \in C^2(\mathbb{R}).$$

Megoldások

Előismeretek

Nevezetes mátrixtípusok

1.1. A mátrix mindhárom sajátértéke 3. A

$$\begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

sajátértékegyenletből a sajátvektorok elemeire azt kapjuk, hogy $x = 0$, $y = 0$, $z \neq 0$ tetszőleges. Tehát nincs 3 lineárisan független sajátvektor, így a mátrix nem diagonalizálható.

Máshogy: Ha diagonalizálható lenne, akkor a $3\mathbf{E}$ mátrixszal lenne hasonló, de akkor $\mathbf{A} = \mathbf{S}(3\mathbf{E})\mathbf{S}^{-1} = 3\mathbf{E}$, ami nyilvánvalóan ellentmondás.

1.2. Az

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

mátrix nem diagonalizálható, hiszen akkor az egységmátrixszal lenne hasonló, így a mátrixnak meg kellene egyezni az egységmátrixszal. Ez pedig nem teljesül.

A

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

mátrix pedig diagonalizálható (sajátértékei különbözőek), de könnyen ellenőrizhetően nem normális.

1.3. Az $\mathbf{A}$ mátrixnak a -2 háromszoros sajátértéke, a hozzá tartozó sajátvektorok a $[-1, 2, -4]^T$ vektor számszorosai. Emiatt a mátrix nem diagonalizálható.

A $\mathbf{B}$ mátrixnak az 1 egyszeres, a 2 kétszeres sajátértéke. Az 1-hez tartozó sajátvektor pl. $[1, 1, -1]^T$, a 2-höz tartozó két lineárisan független sajátvektor pl. $[-4, 0, 1]^T$ és $[-2, 1, 0]^T$. Emiatt a mátrix a

$$\begin{bmatrix} 1 & -4 & -2 \\ 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix}^{-1} \mathbf{B} \begin{bmatrix} 1 & -4 & -2 \\ 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

módon diagonalizálható.

A $\mathbf{C}$ mátrixnak három különböző sajátértéke van, így biztosan diagonalizálható: 2,3 és 6. A hozzájuk tartozó sajátvektorok pl. $[0,1,1]^T$, $[1,-1,1]^T$ és $[-2,-1,1]^T$. Így a mátrix a

$$\begin{bmatrix} 0 & 1 & -2 \\ 1 & -1 & -1 \\ 1 & 1 & 1 \end{bmatrix}^{-1} \mathbf{C} \begin{bmatrix} 0 & 1 & -2 \\ 1 & -1 & -1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

módon diagonalizálható.

1.4. Azt kell megmutatni, hogy $\det(\mathbf{A}-\mathbf{E})=0$, mert ez pontosan azt jelenti, hogy 1 sajátértéke a mátrixnak. A determinánsok szorzási szabályát, valamint a mátrixok transzponáltjának és konstansszorosának determinánsára vonatkozó szabályt használva kapjuk, hogy

$$\det(\mathbf{A}-\mathbf{E}) = \det(\mathbf{A}-\mathbf{A}\mathbf{A}^T) = \det(\mathbf{A})\det(\mathbf{E}-\mathbf{A}^T)$$
$$= \det(\mathbf{E}-\mathbf{A}) = \det(-(\mathbf{A}-\mathbf{E})) = -\det(\mathbf{A}-\mathbf{E}),$$

amiből nyilvánvalóan következik már az állítás.

1.5. A $\overline{\mathbf{v}}$ vektor továbbra is sajátvektor lesz $\lambda(1-\overline{\mathbf{v}}^T\overline{\mathbf{v}})$ sajátértékkel, ugyanis

$$(\mathbf{A}-\lambda\overline{\mathbf{v}}\overline{\mathbf{v}}^T)\overline{\mathbf{v}} = \mathbf{A}\overline{\mathbf{v}} - \lambda\overline{\mathbf{v}}\overline{\mathbf{v}}^T\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}} - \lambda(\overline{\mathbf{v}}^T\overline{\mathbf{v}})\overline{\mathbf{v}} = \lambda(1-\overline{\mathbf{v}}^T\overline{\mathbf{v}})\overline{\mathbf{v}}.$$

Mivel a mátrix szimmetrikus, így a többi sajátirány már mind merőleges lesz az $\overline{\mathbf{v}}$ vektorra. Emiatt tehát egy tetszőleges $\overline{\mathbf{w}}$ sajátvektorra és a hozzá tatozó μ sajátértékre igaz, hogy

$$(\mathbf{A}-\lambda\overline{\mathbf{v}}\overline{\mathbf{v}}^T)\overline{\mathbf{w}} = \mathbf{A}\overline{\mathbf{w}} - \lambda\overline{\mathbf{v}}\overline{\mathbf{v}}^T\overline{\mathbf{w}} = \mu\overline{\mathbf{w}} - 0 = \mu\overline{\mathbf{w}},$$

azaz $\overline{\mathbf{w}}$ az új mátrixnak is sajátvektora lesz ugyanakkora sajátértékkel.

1.6. Tegyük fel, hogy a mátrix $n \times n$-es. A főátlón kívül nincsenek pozitív elemek, így elegendő olyan $\overline{\mathbf{g}}$ pozitív vektort mutatni, melyre $\mathbf{M}\overline{\mathbf{g}}$ pozitív. Azt állítjuk, hogy a $\overline{\mathbf{g}} = [u(h), u(2h), \ldots, u(nh)]^T$ vektor megfelelő lesz, ahol $u : [0,1] \to \mathbb{R}$, $u(x) = x(1-x)$, és $h = 1/(n+1)$. A $\overline{\mathbf{g}}$ vektor pozitivitása nyilvánvaló, továbbá $\mathbf{M}\overline{\mathbf{g}}$ i-edik eleme

$$\frac{-u(h(i-1)) + 2u(hi) - u(h(i+1))}{h^2} = 2,$$

hiszen a fenti képlet pontosan az u függvény ih pontbeli második deriváltjának -1-szeresét adja (lásd numerikus deriválás témakör). Igazából most az is elég lenne, hogy az érték pozitív, ami könnyen látszik az u függvény konkávitásából, de a pontos értéket egy későbbi feladatban használni fogjuk.

1.7. M-mátrixoknak a főátlóiban pozitív elemek állnak. Mivel a főátlóban pozitív elemek, azon kívül pedig nempozitív elemek állnak, így a szigorú dominancia miatt a mátrixra érvényes az $\mathbf{A}\bar{\mathbf{e}} > \mathbf{0}$ becslés. Ez viszont azt jelenti a Gersgorin-tétel szerint, hogy mindegyik sajátértéknek pozitívnak kell lennie, azaz a mátrix pozitív definit.

1.8. Ha μ olyan valós szám, amely nagyobb $\mathbf{M}$ minden főátlóbeli eleménél, akkor a $\mathbf{H} = \mu\mathbf{E} - \mathbf{M}$ mátrix nemnegatív mátrix lesz, hiszen egy M-mátrix főátlóján kívül nem áll pozitív elem, ill. a főátlójában nincs μ-nél nagyobb elem. Így az $\mathbf{M} = \mu\mathbf{E} - \mathbf{H}$ felírás már egy reguláris felbontás, hiszen az előbb láttuk, hogy $\mathbf{H} \geq 0$, másrészt $\mu\mathbf{E}$ invertálható és az inverze is nemnegatív. Mivel ez egy nemnegatív inverzű (M-mátrixról lévén szó) mátrix reguláris felbontása, így $\varrho((1/\mu)\mathbf{EH}) = \varrho((1/\mu)\mathbf{H}) < 1$, azaz $\varrho(\mathbf{H}) < \mu$. Mivel szimmetrikus mátrixok sajátértékei valósak, és $\mathbf{M}$ sajátértékei μ − ($\mathbf{H}$ sajátértékei) alakúak, így $\mathbf{M}$ minden sajátértéke szükségképpen pozitív. Ez mutatja hogy a mátrix pozitív definit.

1.9. Mivel a mátrix szimmetrikus, azt kell igazolni pl., hogy minden bal felső sarokdeterminánsa pozitív. Jelöljük ezeket D_k-val, ahol k a determinánsok méretét jelenti. Látható, hogy $D_1 = 2$, $D_2 = 3$, továbbá a determinánsok kifejtési tétele miatt igaz, hogy $D_{n+1} = 2D_n - D_{n-1}$, ahonnét a $D_n = n + 1$ összefüggést nyerjük, ami nyilvánvalóan pozitív értéket ad minden pozitív egész n-re.

1.10. Korábban láttuk (1.9. feladat), hogy a mátrix éppen a második derivált -1-szeresének közelítését adja. Innét jöhet az ötlet, hogy kipróbáljuk sajátvektornak az $\overline{\mathbf{v}}_k = \sin(ik\pi h)$ alakú vektorokat, ahol $h = 1/(n+1)$, ha a mátrix $n \times n$-es, továbbá $k, i = 1, \ldots, n$.

Ekkor

$$\begin{aligned}(\mathbf{M}\overline{\mathbf{v}}_k)_i &= -\sin((i-1)k\pi h) + 2\sin(ik\pi h) - \sin((i+1)k\pi h)\\ &= -(\sin(ik\pi h)\cos(k\pi h) - \cos(ik\pi h)\sin(k\pi h)) + 2\sin(ik\pi h)\\ &\quad - (\sin(ik\pi h)\cos(k\pi h) + \cos(ik\pi h)\sin(k\pi h))\\ &= 2(1-\cos(k\pi h))\sin(ik\pi h),\end{aligned}$$

ami mutatja, hogy a megadott vektorok valóban sajátvektorok és a hozzájuk tartozó sajátértékek $\lambda_k = 2(1-\cos(k\pi h))$. Megjegyezzük, hogy mivel minden sajátérték pozitív, ez is mutatja, hogy a mátrix pozitív definit (1.9. feladat).

1.11. Mivel a mátrix ferdén szimmetrikus, így $\mathbf{A}^T = -\mathbf{A}$, továbbá a szereplő mátrixok kommutálása miatt igaz, hogy

$$\begin{aligned}(\mathbf{E}+\mathbf{A})^{-1}(\mathbf{E}-\mathbf{A})((\mathbf{E}+\mathbf{A})^{-1}(\mathbf{E}-\mathbf{A}))^T &= (\mathbf{E}+\mathbf{A})^{-1}(\mathbf{E}-\mathbf{A})(\mathbf{E}-\mathbf{A})^T((\mathbf{E}+\mathbf{A})^{-1})^T\\ &= (\mathbf{E}+\mathbf{A})^{-1}(\mathbf{E}-\mathbf{A})(\mathbf{E}+\mathbf{A})(\mathbf{E}-\mathbf{A})^{-1} = \mathbf{E},\end{aligned}$$

azaz a transzponáltja lesz az inverze, így a mátrix valóban ortogonális.

1.12. Ha egy $\mathbf{C}$ mátrix felső háromszögmátrix, akkor $i > j$ esetén $c_{ij} = 0$.

Annak igazolásához, hogy két felső háromszögmátrix szorzata is felső háromszögmátrix, tegyük fel, hogy $\mathbf{A}$ és $\mathbf{B}$ is felső háromszögmátrixok, és számoljuk ki a szorzat i-edik sorának j-edik elemét a főátló alatt $(i > j)$

$$(\mathbf{AB})_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

Itt $a_{ik} = 0$, ha $k < i$, és $b_{kj} = 0$, ha $k > j$, azaz az $i > j$ egyenlőség miatt a fenti összeg minden tagjában valamelyik tényező nulla lesz, így $(\mathbf{AB})_{ij} = 0$.

Most igazoljuk, hogy felső háromszögmátrix inverze is felső háromszögmátrix. Jelölje az inverz mátrixot $\mathbf{B}$, és tegyük fel indirekt, hogy a j-edik oszlopban a főátló alatt van a $\mathbf{B}$ mátrixban nemnulla elem. Válasszuk ki a j-edik oszlopban a főátló alatt a legnagyobb sorindexű nemnulla elemet. Legyen az a b_{ij} elem. Tehát $i > j$ és $b_{kj} = 0$, ha $k > i$. Ekkor

$$(\mathbf{AB})_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = \sum_{k=1}^{i-1} a_{ik} b_{kj} + a_{ii} b_{ij} + \sum_{k=i+1}^{n} a_{ik} b_{kj}.$$

Itt az első tag nulla, hiszen az $\mathbf{A}$ mátrix felső háromszög, a második tag nem nulla, mert $\mathbf{A}$ invertálható és $b_{ij} \neq 0$, és a harmadik tag szintén nulla (ha van egyáltalán), mert $k > i$.

Megegyezzük, hogy az állítás igazolható az inverz mátrix Gauss-eliminációs meghatározási módszerének felhasználásával is úgy, hogy végiggondoljuk, hogy hol lesznek nemnulla elemek az inverz mátrixban.

1.13. Jelöljük a $\mathbf{T}$ mátrix elemeit t_{ij}-vel. A mátrixegyenlőség két oldalán lévő mátrixok első sorának első elemét kiszámítva a

$$t_{11}^2 = t_{11}^2 + t_{12}^2 + \ldots + t_{1n}^2$$

egyenlőséghez jutunk, ami csak úgy teljesülhet, ha $\mathbf{T}$ első sorában a főátlón kívül nullák állnak. Hasonlóan okoskodhatunk a többi főátlóbeli elem esetén, amiből már következik, hogy $\mathbf{T}$ diagonális mátrix.

Normált és euklideszi terek

1.14.

$$\begin{aligned}
\frac{1}{4}\left(\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2\right) &= \frac{1}{4}\left(\langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}\rangle - \langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y}\rangle\right) \\
&= \frac{1}{4}\left(\|\mathbf{x}\|^2 + 2\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2 - \left(\|\mathbf{x}\|^2 - 2\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2\right)\right) \\
&= \langle \mathbf{x}, \mathbf{y}\rangle.
\end{aligned}$$

1.15.

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2+\|\mathbf{x}-\mathbf{y}\|^2 &= \langle \mathbf{x}+\mathbf{y},\mathbf{x}+\mathbf{y}\rangle + \langle \mathbf{x}-\mathbf{y},\mathbf{x}-\mathbf{y}\rangle \\ &= \|\mathbf{x}\|^2+2\langle \mathbf{x},\mathbf{y}\rangle+\|\mathbf{y}\|^2+(\|\mathbf{x}\|^2-2\langle \mathbf{x},\mathbf{y}\rangle+\|\mathbf{y}\|^2) \\ &= 2\|\mathbf{x}\|^2+2\|\mathbf{y}\|^2.\end{aligned}$$

1.16. Az állítás közvetlen következménye a polarizációs egyenlőségnek (1.14. feladat).

1.17. Az állítás triviálisan igaz, ha $\mathbf{x}$ vagy $\mathbf{y}$ nullvektor. Tegyük fel, hogy egyik sem nullvektor, és tekintsük a

$$\phi(t)=\langle \mathbf{x}+t\mathbf{y},\mathbf{x}+t\mathbf{y}\rangle$$

valós függvényt. Ez a függvény nyilvánvalóan nem vehet fel negatív értéket, továbbá érvényes, hogy

$$\phi(t)=\|\mathbf{x}\|^2+2t\langle \mathbf{x},\mathbf{y}\rangle+t^2\|\mathbf{y}\|^2.$$

Ez csak úgy lehet, ha a t-ben másodfokú kifejezés diszkriminánsa nempozitív, azaz

$$4\langle \mathbf{x},\mathbf{y}\rangle^2-4\|\mathbf{x}\|^2\|\mathbf{y}\|^2\le 0,$$

ami éppen az igazolandó egyenlőtlenséget adja.

1.18.

$$\|\mathbf{x}+\mathbf{y}\|^2=\langle \mathbf{x}+\mathbf{y},\mathbf{x}+\mathbf{y}\rangle=\|\mathbf{x}\|^2+2\langle \mathbf{x},\mathbf{y}\rangle+\|\mathbf{y}\|^2,$$

ahonnét egyszerre látszik, hogy az állítás pontosan akkor teljesül csak, ha $\langle \mathbf{x},\mathbf{y}\rangle=0$, azaz a vektorok ortogonálisak.

Banach-féle fixponttétel

1.19. A T leképezésnek a Banach-féle fixponttétel szerint van egyértelműen létező fixpontja (T a zárt $[a,b]$ intervallumból ugyanebbe az intervallumba képez és kontrakció). Jelöljük ezt $x^\star$-gal. Meg kell mutatni, hogy $x^\star$ F-nek is fixpontja, és hogy a fixpont egyértelmű. Az

$$\|F(x^\star)-x^\star\|=\|F(T(x^\star))-T(x^\star)\|=\|T(F(x^\star))-T(x^\star)\|\le q\|F(x^\star)-x^\star\|$$

egyenlőtlenség csak úgy teljesülhet, ha $\|F(x^\star)-x^\star\|=0$, azaz ha $F(x^\star)=x^\star$, ami azt jelenti, hogy $x^\star$ F-nek is fixpontja. A második lépésben felhasználtuk, hogy T és F felcserélhető, a harmadikban pedig azt, hogy T kontrakció.

Az egyértelműséghez elég arra hivatkozni, hogy a T leképezésnek a Banach-féle fixponttétel miatt pontosan egy fixpontja van, így F-nek sem lehet egynél több, hiszen F fixpontjai egyúttal T-nek is fixpontjai.

1.20. A Lagrange-féle középértéktétel miatt tetszőleges $x, y \in [0, \infty)$ számokra

$$|F(x) - F(y)| = |F'(\xi)| \cdot |x - y|,$$

ahol ξ egy az x és y értékek közé eső megfelelő szám. Mivel $F'(\xi) = 1/2 - 1/\xi^2$ és ennek abszolút értéke nem lehet $1/2$-nél nagyobb az $[1, \infty)$ intervallumon, ezért írhatjuk, hogy

$$|F(x) - F(y)| = |1/2 - 1/\xi^2| \cdot |x - y| \leq 1/2|x - y|.$$

Tehát F valóban kontrakció, és a kontrakciós tényező választható $1/2$-ednek. Ez a lehetséges legkisebb kontrakciós tényező, hiszen ha x és y elegendően nagyok, akkor $|1/2 - 1/\xi^2|$ tetszőlegesen közel kerülhet (alulról) $1/2$-hez. Teljesülnek tehát a Banach-féle fixponttétel feltételei, így F-nek egyértelműen létezik fixpontja. F fixpontjának meghatározásához az $F(x^\star) = x^\star/2 + 1/x^\star = x^\star$ egyenletet kell megoldani, melynek megoldásai $x^\star_{1,2} = \pm\sqrt{2}$. Ezek közül csak az $x^\star = \sqrt{2}$ érték esik a $[0, \infty)$ intervallumba, így az a fixpont.

1.21. Mutassuk meg először, hogy nem lehet egynél több fixpont! Tegyük fel indirekt, hogy van két fixpont. Jelöljük ezeket $x^\star$-gal és $y^\star$-gal! Ekkor, kihasználva a feladatbeli, a kontrakciós tulajdonságot helyettesítő feltételt, érvényes az alábbi becslés:

$$\|x^\star - y^\star\| = \|F(x^\star) - F(y^\star)\| < \|x^\star - y^\star\|,$$

ami nyilvánvaló ellentmondás. Így nem lehet egynél több fixpont.

Azt, hogy a feltételek mellett nem feltétlenül van fixpont mutatja az $F : [1, \infty) \to [1, \infty)$, $F(x) = x + 1/x$ függvény. Ennek a függvénynek nyilvánvalóan nincs fixpontja a $[0, \infty)$ intervallumon, viszont a feladatban szereplő feltételt kielégíti, ugyanis tetszőleges $x, y \in [1, \infty)$ esetén a Lagrange-féle középértéktételt használva

$$|F(x) - F(y)| = |x + 1/x - (y + 1/y)| = |1 - 1/\xi^2| \cdot |x - y| < |x - y|.$$

1.22. Mivel

$$\|T(\mathbf{x}_1) + \mathbf{y} - (T(\mathbf{x}_2) + \mathbf{y})\| = \|T(\mathbf{x}_1) - T(\mathbf{x}_2)\| \leq q\|\mathbf{x}_1 - \mathbf{x}_2\|,$$

ezért az $\mathbf{x} \mapsto T(\mathbf{x}) + \mathbf{y}$ leképezés is kontrakció V-n, így pontosan egy fixpontja van. Ezzel igazoltuk, hogy az egyenletnek mindig pontosan egy megoldása lesz. Legyen a megoldófüggvény, azaz az a függvény, amely az $\mathbf{y}$ elemhez hozzárendeli az egyenlet $\mathbf{x}$ megoldását, u. Azt kell megmutatnunk, hogy u folytonos. Tegyük fel tehát, hogy egy $\{\mathbf{y}_n\}$ sorozat $\mathbf{y}$-hoz tart! Igazoljuk, hogy ekkor $u(\mathbf{y}_n) \to u(\mathbf{y})$! Mivel

$$\begin{aligned}
\|u(\mathbf{y}_n) - u(\mathbf{y})\| &= \|T(u(\mathbf{y}_n)) + \mathbf{y}_n - (T(u(\mathbf{y})) + \mathbf{y})\| \\
&\leq \|T(u(\mathbf{y}_n)) - (T(u(\mathbf{y})) + \mathbf{y}_n - \mathbf{y})\| \\
&\leq \|T(u(\mathbf{y}_n)) - (T(u(\mathbf{y}))\| + \|\mathbf{y}_n - \mathbf{y})\| \\
&\leq q\|u(\mathbf{y}_n) - u(\mathbf{y})\| + \|\mathbf{y}_n - \mathbf{y})\|,
\end{aligned}$$

így

$$\|u(\mathbf{y}_n) - u(\mathbf{y})\| \leq \frac{1}{1-q}\|\mathbf{y}_n - \mathbf{y})\|,$$

amiből már következik az állítás.

1.23. Először igazoljuk, hogy van olyan $0 \leq q < 1$ szám, melyre $|f'(c)| \leq q$ minden $c \in [a,b]$ esetén! Tegyük fel indirekt, hogy nincs ilyen q! Ekkor minden $n \in \mathbb{N}$ esetén létezik olyan $c_n \in [a,b]$, melyre $1 - 1/n < |f'(c_n)| < 1$. Mivel a $\{c_n\}$ sorozat korlátos, így van $\{c_{i_n}\}$ konvergens részsorozata. Legyen ennek határértéke $c^\star \in [a,b]$! Mivel a feladat feltétele szerint $f'(x)$ folytonos $[a,b]$-n, így $f'(c^\star) = 1$ lenne, ami ellentmond a feladat feltételének.

Ezek után a feladat állítása már a Lagrange-féle középértéktételből következik, ugyanis eszerint tetszőleges $x \neq y \in [a,b]$ esetén egy tőlük függő megfelelő $c \in (a,b)$ számra

$$\frac{|f(x) - f(y)|}{|x-y|} = |f'(c)|,$$

azaz

$$|f(x) - f(y)| \leq |f'(c)| \cdot |x-y| \leq q|x-y|,$$

ami azt jelenti, hogy f kontrakció.

Vektornormák

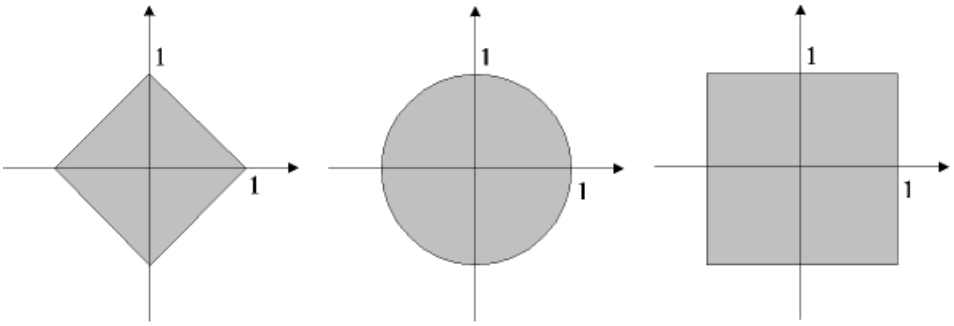
1.24. Az 1-es norma az elemek abszolút értékben vett összege, azaz $1 + |-2| + 3 = 6$. Az euklideszi norma az elemek négyzetösszegének gyöke, azaz $\sqrt{1^2 + (-2)^2 + 3^2} = \sqrt{14}$. A maximumnorma pedig a legnagyobb abszolút értékű elem abszolút értéke, azaz 3.

1.25. $\|\overline{\mathbf{x}}\|_1 = 1 + 2 + \ldots + 100 = 5050,$

$$\|\overline{\mathbf{x}}\|_2 = \sqrt{1^2 + 2^2 + \ldots, 100^2} = \sqrt{100 \cdot 101 \cdot 201/6} \approx 581.6786,$$

$1 + 2 + \ldots + 100 = 5050$, $\|\overline{\mathbf{x}}\|_\infty = 100.$

1.26. Kételemű $\overline{\mathbf{x}} = [x,y]^T$ oszlopvektorok esetén a nevezetes normák képletei a következők: $\|\overline{\mathbf{x}}\|_1 = |x| + |y|$, $\|\overline{\mathbf{x}}\|_2 = x^2 + y^2$, $\|\overline{\mathbf{x}}\|_\infty = \max\{|x|, |y|\}$. Emiatt a síkon rendre azon (x,y) pontok esnek az origó adott normában 1 sugarú környezetébe, melyek koordinátáira rendre igaz, hogy $|x| + |y| < 1$, $x^2 + y^2 < 1$ ill. $\max\{|x|, |y|\} < 1$. Ezek az alakzatok rendre a 10.4 ábrán látható tartományokat adják környezetként.



10.4. ábra. Az origó 1 sugarú környezete 1-es, 2-es és ∞ normában.

1.27. Legyen $\overline{\mathbf{x}}$ egy tetszőleges $\mathbb{R}^n$-beli vektor. Ekkor nyilvánvalóan $\|\overline{\mathbf{x}}\|_\infty \leq \|\overline{\mathbf{x}}\|_1$ és $\|\overline{\mathbf{x}}\|_\infty \leq \|\overline{\mathbf{x}}\|_2$. A normák ekvivalenciája az alábbi egyszerű becslésekből következik

$$\|\overline{\mathbf{x}}\|_\infty \leq \|\overline{\mathbf{x}}\|_2 \leq \sqrt{n}\|\overline{\mathbf{x}}\|_\infty \leq \sqrt{n}\|\overline{\mathbf{x}}\|_1 \leq n\sqrt{n}\|\overline{\mathbf{x}}\|_\infty = n^{3/2}\|\overline{\mathbf{x}}\|_\infty.$$

Megjegyezzük, hogy az 1-es és a 2-es normára vonatkozóan a fenti becsléseknél élesebb becslések is igazak. Nevezetesen

$$\|\overline{\mathbf{x}}\|_2 \leq \|\overline{\mathbf{x}}\|_1 \leq \sqrt{n}\|\overline{\mathbf{x}}\|_2.$$

A bal oldali reláció az

$$\|\overline{\mathbf{x}}\|_1 = |x_1| + \ldots + |x_n| \leq \sqrt{|x_1|^2 + \ldots + |x_n|^2 + 2|x_1||x_2| + \ldots} \geq \|\overline{\mathbf{x}}\|_2$$

becslésből, a jobb oldali pedig a számtani és kvadratikus közepek közti egyenlőtlenségből következik.

1.28. Ha a norma skaláris szorzatból származna, akkor teljesítené a parallelogramma egyenlőséget (1.15. feladat). Így ellenpéldát kell mutatnunk. Tekintsük pl. az $\overline{\mathbf{e}}_1$ és $\overline{\mathbf{e}}_2$ egységvektorokat! Ezekkel

$$2 = \|\overline{\mathbf{e}}_1 + \overline{\mathbf{e}}_2\|_\infty^2 + \|\overline{\mathbf{e}}_1 - \overline{\mathbf{e}}_2\|_\infty^2 \neq 2\|\overline{\mathbf{e}}_1\|_\infty^2 + 2\|\overline{\mathbf{e}}_2\|_\infty^2 = 4,$$

így a maximumnorma nem lehet indukált norma. Továbbá

$$8 = \|\overline{\mathbf{e}}_1 + \overline{\mathbf{e}}_2\|_1^2 + \|\overline{\mathbf{e}}_1 - \overline{\mathbf{e}}_2\|_1^2 \neq 2\|\overline{\mathbf{e}}_1\|_1^2 + 2\|\overline{\mathbf{e}}_2\|_1^2 = 4,$$

azaz az 1-es norma sem lehet indukált norma.

1.29. Az állítás az alábbi becslésből és a rendőr-elvből következik:

$$\|\overline{\mathbf{x}}\|_\infty = \sqrt[p]{\|\overline{\mathbf{x}}\|_\infty^p} \leq \|\overline{\mathbf{x}}\|_p = \sqrt[p]{|x_1|^p + \ldots + |x_n|^p} \leq \sqrt[p]{n\|\overline{\mathbf{x}}\|_\infty^p} \leq \sqrt[p]{n}\|\overline{\mathbf{x}}\|_\infty \to \|\overline{\mathbf{x}}\|_\infty,$$

ha $p \to \infty$.

1.30. Tekintsük az

$$f(a) = \frac{a^p}{p} + \frac{b^q}{q} - ab$$

valós függvényt a nemnegatív a számokon. Vizsgáljuk meg ezt a függvényt! Az

$$f'(a) = a^{p-1} - b = 0$$

egyenlőség csak az $a = b^{1/(p-1)}$ pontban teljesül, itt

$$f(b^{1/(p-1)}) = b^q \left(\frac{1}{p} + \frac{1}{q}\right) - b^q = 0,$$

továbbá $a = b^{1/(p-1)}$-től jobbra szigorúan monoton növő a függvény, balra pedig szigorúan monoton csökkenő, $f(0) = b^q/q \geq 0$. Ebből látszik, hogy $f(a) \geq 0$ a teljes $[0, \infty)$ intervallumon, ami átrendezve éppen a keresett egyenlőtlenséget adja.

Most térjünk át a Hölder-egyenlőtlenség igazolására! Az állítás nyilvánvalóan igaz, ha p vagy q értéke 1, vagy ha $\overline{\mathbf{x}}$ vagy $\overline{\mathbf{y}}$ nullvektorok. Tegyük fel, hogy $1 < p, q < \infty$ és hogy egyik vektor sem a nullvektor. Legyenek $\overline{\mathbf{x}}$ és $\overline{\mathbf{y}}$ olyanok, hogy $\|\overline{\mathbf{x}}\|_p = \|\overline{\mathbf{y}}\|_q = 1$. Ekkor a Young-egyenlőtlenség alkalmazásával kapjuk, hogy

$$|\langle \overline{\mathbf{x}}, \overline{\mathbf{y}} \rangle| = \left| \sum_{i=1}^n x_i y_i \right| \leq \sum_{i=1}^n |x_i|\,|y_i| \leq \sum_{i=1}^n \left(\frac{|x_i|^p}{p} + \frac{|y_i|^q}{q} \right) = \frac{\|\overline{\mathbf{x}}\|_p^p}{p} + \frac{\|\overline{\mathbf{y}}\|_q^q}{q} = \frac{1}{p} + \frac{1}{q} = 1.$$

Általános esetben ($\|\overline{\mathbf{x}}\|_p = \|\overline{\mathbf{y}}\|_q = 1$ valamelyike nem teljesül) alkalmazzuk a fent nyert becslést az $\overline{\mathbf{x}}/\|\overline{\mathbf{x}}\|_p$ és $\overline{\mathbf{y}}/\|\overline{\mathbf{y}}\|_q$ vektorokra, melyek p és q normája most már egységnyi:

$$|\langle \overline{\mathbf{x}}/\|\overline{\mathbf{x}}\|_p, \overline{\mathbf{y}}/\|\overline{\mathbf{y}}\|_q \rangle| \leq 1,$$

ahonnét $\|\overline{\mathbf{x}}\|_p \|\overline{\mathbf{y}}\|_q$-val való szorzás után éppen a Hölder-egyenlőtlenséget nyerjük.

1.31. A normaaxiómák közül az első kettő triviálisan teljesül. Csak a harmadik teljesülését (háromszög-egyenlőtlenség) kell megmutatni, azaz azt, hogy $1 \leq p \leq \infty$ esetén

$$\|\overline{\mathbf{x}} + \overline{\mathbf{y}}\|_p \leq \|\overline{\mathbf{x}}\|_p + \|\overline{\mathbf{y}}\|_p.$$

Ez éppen az ún. Minkowski-egyenlőtlenség, melyet az alábbi módon igazolhatunk. Ha

$\|\overline{\mathbf{x}}+\overline{\mathbf{y}}\|_p = 0$, akkor triviális az állítás, különben pedig az alábbi becsléseket tehetjük:

$$\begin{aligned}
\|\overline{\mathbf{x}}+\overline{\mathbf{y}}\|_p^p &= \sum_{i=1}^n |x_i+y_i|^p \le \sum_{i=1}^n |x_i+y_i|^{p-1}(|x_i|+|y_i|) \\
&= \sum_{i=1}^n |x_i||x_i+y_i|^{p-1} + \sum_{i=1}^n |y_i||x_i+y_i|^{p-1} \\
&\le (\|\overline{\mathbf{x}}\|_p + \|\overline{\mathbf{y}}\|_p) \sqrt[q]{\sum_{i=1}^n |x_i+y_i|^{q(p-1)}} \\
&= (\|\overline{\mathbf{x}}\|_p + \|\overline{\mathbf{y}}\|_p) \sqrt[q]{\sum_{i=1}^n |x_i+y_i|^{p}} \\
&= (\|\overline{\mathbf{x}}\|_p + \|\overline{\mathbf{y}}\|_p) \|\overline{\mathbf{x}}+\overline{\mathbf{y}}\|_p^{p/q},
\end{aligned}$$

ahol a második sorból a harmadikat a Hölder-egyenlőtlenség segítségével nyertük. Az első tagban azt pl. az $(|x_1|,\dots,|x_n|)^T$ és $(|x_1+y_1|^{p-1},\dots,|x_n+y_n|^{p-1})^T$ vektorokra alkalmaztuk.

Ezek után a keresett egyenlőtlenséget $\|\overline{\mathbf{x}}+\overline{\mathbf{y}}\|_p^{p/q}$-val való osztás után nyerjük, hiszen

$$\frac{\|\overline{\mathbf{x}}+\overline{\mathbf{y}}\|_p^p}{\|\overline{\mathbf{x}}+\overline{\mathbf{y}}\|_p^{p/q}} = \|\overline{\mathbf{x}}+\overline{\mathbf{y}}\|_p \le \|\overline{\mathbf{x}}\|_p + \|\overline{\mathbf{y}}\|_p,$$

amit igazolni szerettünk volna. Azaz a p-norma valóban normát ad meg.

1.32. A normaaxiómákat kell leellenőrizni kihasználva, hogy $\|\cdot\|$ normaként teljesíti a norma axiómáit. Nyilvánvalóan $\|\overline{\mathbf{x}}\|_A = \|\mathbf{A}\overline{\mathbf{x}}\|$ pontosan akkor ad nullát, ha $\mathbf{A}\overline{\mathbf{x}} = \mathbf{0}$, ez pedig pontosan akkor teljesül, ha $\overline{\mathbf{x}}$ a nullvektor, hiszen $\mathbf{A}$ invertálható.

Továbbá tetszőleges $\alpha \in \mathbb{C}$ esetén

$$\|\alpha\overline{\mathbf{x}}\|_A = \|\mathbf{A}(\alpha\overline{\mathbf{x}})\| = \|\alpha(\mathbf{A}\overline{\mathbf{x}})\| = |\alpha| \cdot \|\mathbf{A}\overline{\mathbf{x}}\| = |\alpha| \cdot \|\overline{\mathbf{x}}\|_A.$$

Így a második axióma is teljesül.

A harmadik axióma érvényessége az alábbi becslésből látható (két tetszőleges $\overline{\mathbf{x}}, \overline{\mathbf{y}} \in \mathbb{R}^n$ vektor esetén):

$$\|\overline{\mathbf{x}}+\overline{\mathbf{y}}\|_A = \|\mathbf{A}(\overline{\mathbf{x}}+\overline{\mathbf{y}})\| = \|\mathbf{A}\overline{\mathbf{x}}+\mathbf{A}\overline{\mathbf{y}}\| \le \|\mathbf{A}\overline{\mathbf{x}}\| + \|\mathbf{A}\overline{\mathbf{y}}\| = \|\overline{\mathbf{x}}\|_A + \|\overline{\mathbf{y}}\|_A.$$

Mátrixnormák

1.33. A norma axiómái közül az első kettő triviálisan teljesül. A harmadikhoz pedig az

$$\|\mathbf{A}+\mathbf{B}\| = \max_{i,j} |a_{ij}+b_{ij}| \le \max_{i,j}(|a_{ij}|+|b_{ij}|) \le \max_{i,j}|a_{ij}| + \max_{ij}|b_{ij}| = \|\mathbf{A}\| + \|\mathbf{B}\|$$

becslésből következik.

A norma nem származtatható vektornormából, ugyanis az 1.1. tételben nem teljesül a harmadig tulajdonság pl. ha $\mathbf{A}$-t is és $\mathbf{B}$-t is a csupa 1 mátrixnak választjuk. Ekkor $\|\mathbf{A}\| = \|\mathbf{B}\| = 1$, de $\|\mathbf{AB}\| = n$.

1.34. A becslések egyszerűen következnek az 1.27. feladat eredményéből.

1.35. A norma axiómái közül az első kettő triviálisan teljesül. A harmadik becslés pedig a Minkowski-egyenlőtlenség következménye (lásd 1.31. feladat).

A Frobenius-norma amiatt nem lehet indukált norma, mert akkor az egységmátrix normájának 1-nek kelle lennie, viszont $\|\mathbf{E}\|_F = \sqrt{n}$.

1.36.

$$(\mathbf{A}^T\mathbf{A})_{ii} = \sum_{k=1}^{n}(\mathbf{A}^T)_{ik}(\mathbf{A})_{ki} = \sum_{k=1}^{n}(\mathbf{A})_{ki}(\mathbf{A})_{ki} = \sum_{k=1}^{n}((\mathbf{A})_{ki})^2,$$

azaz az i-edik oszlop négyzetösszege. Azaz a főátlóbeli elemek összege megadja az összes oszlop négyzetösszegét, azaz a Frobenius-normát.

Legyen $\mathbf{B} = \mathbf{S}^T\mathbf{AS}$, ahol $\mathbf{S}$ ortogonális mátrix. Hasonló mátrixok sajátértékei megegyeznek, így a trace-ük is megegyezik, mert az a sajátértékek összege.

$$\|\mathbf{B}\|_F^2 = \operatorname{trace}(\mathbf{B}^T\mathbf{B}) = \operatorname{trace}(\mathbf{S}^T\mathbf{A}^T(\mathbf{SS}^T)\mathbf{AS})$$

$$= \operatorname{trace}(\mathbf{S}^T(\mathbf{A}^T\mathbf{A})\mathbf{S}) = \operatorname{trace}(\mathbf{A}^T\mathbf{A}) = \|\mathbf{A}\|_F^2.$$

(Itt azt is kihasználhattuk volna, hogy a mátrixok ciklikus permutációja során a trace nem változik.)

1.37. Jelölje $\overline{\mathbf{a}}_{i\star}$ az $\mathbf{A}$ mátrix i-edik sorvektorát! Ekkor a Cauchy–Schwarz–Bunyakovszkij-egyenlőtlenséget használva

$$\|\mathbf{A}\overline{\mathbf{x}}\|_2^2 = \sum_{i=1}^{n}(\overline{\mathbf{a}}_{i\star}\overline{\mathbf{x}})^2 \le \sum_{i=1}^{n}\|\overline{\mathbf{a}}_{i\star}\|_2^2\|\overline{\mathbf{x}}\|_2^2 = \|\overline{\mathbf{x}}\|_2^2\|\mathbf{A}\|_F^2.$$

Ezt szerettük volna megmutatni.

1.38. Jelölje általánosan egy $\mathbf{C}$ mátrix j-edik oszlopát $(\mathbf{C})_{\star j}$. Ekkor

$$\|\mathbf{AB}\|_F^2 = \sum_{j=1}^{n}\|(\mathbf{AB})_{\star j}\|_2^2 = \sum_{j=1}^{n}\|\mathbf{A}(\mathbf{B})_{\star j}\|_2^2 \le \sum_{j=1}^{n}\|\mathbf{A}\|_F^2\|\mathbf{B}_{\star j}\|_2^2 = \|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2,$$

ahol felhasználtuk az 1.37. feladat eredményét.

1.39. Legyen $\overline{\mathbf{v}}$ egy $\mathbf{A}$ négyzetes mátrix egy sajátvektora és λ a hozzá tartozó sajátérték. Ekkor igaz, hogy $\mathbf{A}\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}$. Jobbról szorozzunk $\overline{\mathbf{v}}^T$-tal, majd vegyük mindkét oldal normáját:

$$\|\mathbf{A}\overline{\mathbf{v}}\overline{\mathbf{v}}^T\| = \|\lambda\overline{\mathbf{v}}\overline{\mathbf{v}}^T\|!$$

A bal oldalt becsüljük a szubmultiplikatív tulajdonság alapján, a jobb oldalon pedig a norma egyik axiómáját használva:

$$\|\mathbf{A}\| \cdot \|\overline{\mathbf{v}}\overline{\mathbf{v}}^T\| \geq \|\mathbf{A}\overline{\mathbf{v}}\overline{\mathbf{v}}^T\| = |\lambda| \cdot \|\overline{\mathbf{v}}\overline{\mathbf{v}}^T\|.$$

Mivel $\overline{\mathbf{v}} \neq 0$, így $\|\overline{\mathbf{v}}\overline{\mathbf{v}}^T\| \neq 0$, és ezzel a tényezővel osztva adódik, hogy a sajátérték abszolút értéke nem lehet nagyobb, mint a norma. Így ez igaz a spektrálsugárra is.

Mivel $\|\mathbf{A}\|_1 = \|\mathbf{A}\|_\infty = 1.1$ és $\|\mathbf{A}\|_F = \sqrt{0.5^2 + 0.5^2 + 0.6^2 + 0.1^2} \approx 0.93$, ezért csak a Frobenius-norma értékéből következik, hogy a spektrálsugár kisebb, mint 1.

1.40. Mindegyik esetben a főátlóbeli elemek legnagyobb abszolút értékét kapjuk eredményül: $\|\mathbf{D}\| = \max_i |d_{ii}|$. Jelöljük az egyszerűség kedvéért ezt az értéket D-vel!

1-es norma esetén: mivel

$$\|\mathbf{D}\overline{\mathbf{x}}\|_1 = \sum_{i=1}^{n} |d_{ii}x_i| \leq D \sum_i |x_i|,$$

így $\|\mathbf{D}\|_1 \leq D$. Ha az $\overline{\mathbf{x}}$ vektort $\overline{\mathbf{e}}_j$-nek választjuk, ahol j az az index, melyre $|d_{jj}| = D$, akkor $\|\mathbf{D}\overline{\mathbf{x}}\|_1 = D$, ami mutatja, hogy $\|\mathbf{D}\|_1 = D$.

Az állítás maximum- és euklideszi normára is hasonlóan igazolható.

1.41. Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy adott mátrix és $\overline{\mathbf{x}} \in \mathbb{R}^n$ egy tetszőleges nemnulla vektor. Ekkor

$$\|\mathbf{A}\overline{\mathbf{x}}\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \leq \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}||x_j| = \sum_{j=1}^{n}\sum_{i=1}^{n} |a_{ij}||x_j| = \sum_{j=1}^{n} \left(|x_j| \sum_{i=1}^{n} |a_{ij}| \right) \leq$$

$$\leq \left(\max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \right) \sum_{j=1}^{n} |x_j| = \left(\max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \right) \|\overline{\mathbf{x}}\|_1,$$

ami mutatja, hogy $\|\mathbf{A}\|_1 \leq \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|$. Az egyenlőséghez azt kell megmutatni, hogy van olyan $\overline{\mathbf{x}}_0 \in \mathbb{R}^n$ vektor, mellyel a fenti becslésekben egyenlőségek szerepelnek. Tegyük fel, hogy a $\sum_{i=1}^{n} |a_{ij}|$ összeg a j_0 oszlopban a legnagyobb. Ekkor az $\overline{\mathbf{x}}_0 = \overline{\mathbf{e}}_{j_0} \sum_{i=1}^{m} |a_{ij_0}|$ választás megfelelő, ugyanis

$$\|\mathbf{A}\overline{\mathbf{x}}_0\|_1 = \left(\sum_{i=1}^{n} |a_{ij_0}| \right) \sum_{i=1}^{n} |a_{ij_0}| = \left(\max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \right) \|\overline{\mathbf{x}}_0\|_1.$$

Itt $\overline{\mathbf{e}}_{j_0}$ a j_0-adik egységvektort jelöli, azaz azt az n elemű vektort, melynek j_0-adik eleme 1, a többi pedig nulla.

1.42. Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy adott mátrix és $\overline{\mathbf{x}} \in \mathbb{R}^n$ egy tetszőleges nemnulla vektor. Ekkor

$$\|\mathbf{A}\overline{\mathbf{x}}\|_\infty = \max_i \left|\sum_{j=1}^n a_{ij}x_j\right| \le \max_i \sum_{j=1}^n |a_{ij}|\,|x_j| \le \max_i \sum_{j=1}^n |a_{ij}| \max_k |x_k|$$

$$= \left(\max_k |x_k|\right) \max_i \sum_{j=1}^n |a_{ij}|\,,$$

ami mutatja, hogy $\|\mathbf{A}\|_\infty \le \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|$. Az egyenlőséghez azt kell megmutatni, hogy van olyan $\overline{\mathbf{x}}_0 \in \mathbb{R}^n$ vektor, mellyel a fenti becslésekben egyenlőségek szerepelnek. Legyen i_0 annak a sornak az indexe, melynek abszolút értékben vett összege éppen $\max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|$. Válasszuk $\overline{\mathbf{x}}_0$-nak ezek után a $\operatorname{sgn}([a_{i_0 1}, \dots, a_{i_0 n}]^T)$ vektort! Ezzel

$$\|\mathbf{A}\overline{\mathbf{x}}_0\|_\infty = \max_i \left|\sum_{j=1}^n a_{ij}(x_0)_j\right| = \max_i \sum_{j=1}^n |a_{ij}|\,.$$

Ezzel igazoltuk az állítást.

1.43. Az $\mathbf{A}^H\mathbf{A}$ mátrix hermitikus és pozitív szemidefinit. Az hermitikusság nyilvánvaló, a pozitív szemidefinitség következik az $\overline{\mathbf{x}}^H\mathbf{A}^H\mathbf{A}\overline{\mathbf{x}} = \|\mathbf{A}\overline{\mathbf{x}}\|_2^2 \ge 0$ egyenlőtlenségből. Az hermitikusság miatt a mátrix diagonalizálható, azaz $\mathbf{A}^H\mathbf{A}$ felírható $\mathbf{A}^H\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$ alakban, ahol $\mathbf{V}$ megfelelő unitér mátrix, $\mathbf{\Lambda}$ pedig a nemnegatív valós sajátértékeket tartalmazó diagonális mátrix. Így

$$\frac{\|\mathbf{A}\overline{\mathbf{x}}\|_2^2}{\|\overline{\mathbf{x}}\|_2^2} = \frac{\overline{\mathbf{x}}^H\mathbf{A}^H\mathbf{A}\overline{\mathbf{x}}}{\|\overline{\mathbf{x}}\|_2^2} = \frac{\overline{\mathbf{x}}^H\mathbf{V}\mathbf{\Lambda}\mathbf{V}^H\overline{\mathbf{x}}}{\|\overline{\mathbf{x}}\|_2^2} = \frac{\|\sqrt{\mathbf{\Lambda}}\mathbf{V}^H\overline{\mathbf{x}}\|_2^2}{\|\overline{\mathbf{x}}\|_2^2}$$
$$\le \frac{\|\sqrt{\mathbf{\Lambda}}\mathbf{V}^H\|_2^2\|\overline{\mathbf{x}}\|_2^2}{\|\overline{\mathbf{x}}\|_2^2} = \|\sqrt{\mathbf{\Lambda}}\|_2^2 = \varrho(\mathbf{A}^H\mathbf{A}).$$

A $\sqrt{\mathbf{\Lambda}}$ mátrix az a diagonális mátrix, melynek főátlóbeli elemei $\mathbf{\Lambda}$ megfelelő elemeinek gyökei. Az $\mathbf{A}^H\mathbf{A}$ mátrix legnagyobb abszolút értékű sajátértékéhez tartozó sajátvektort választva $\overline{\mathbf{x}}$-nek pont egyenlőség van. Így az állítás valóban igaz.

1.44. Az, hogy a hozzárendelés normát ad meg következik az 1.33. feladat eredményéből. A szubmultiplikativitás pedig az alábbi módon látható be.

$$\|\mathbf{AB}\| = n \max_{i,j} |(\mathbf{AB})_{ij}| \le n \max_{i,j} \sum_{k=1}^n |a_{ik}b_{kj}| \le n \max_{i,j} \sum_{k=1}^n |a_{ik}||b_{kj}|$$

$$\le n \cdot n \cdot \max_{i,j} |a_{ij}| \max_{i,j} |b_{ij}| = \|\mathbf{A}\|\|\mathbf{B}\|.$$

1.45. Jelölje a mátrixnormát $\|\cdot\|$! Definiáljunk egy vektornormát egy tetszőleges $\overline{\mathbf{y}} \neq \mathbf{0}$ vektor segítségével az alábbi módon: $\|\overline{\mathbf{x}}\| = \|\overline{\mathbf{x}}\overline{\mathbf{y}}^T\|$! Látható, hogy ezzel a vektornormával konzisztens a mátrixnorma, ugyanis

$$\|\mathbf{A}\overline{\mathbf{x}}\| = \|\mathbf{A}\overline{\mathbf{x}}\overline{\mathbf{y}}^T\| \leq \|\mathbf{A}\| \cdot \|\overline{\mathbf{x}}\overline{\mathbf{y}}^T\| = \|\mathbf{A}\|\|\overline{\mathbf{x}}\|.$$

1.46.

$$\|\mathbf{AB}\| = \sup_{\overline{\mathbf{x}}\neq\mathbf{0}} \frac{\|\mathbf{AB}\overline{\mathbf{x}}\|}{\|\overline{\mathbf{x}}\|} \leq \sup_{\overline{\mathbf{x}}\neq\mathbf{0}} \frac{\|\mathbf{A}\|\|\mathbf{B}\overline{\mathbf{x}}\|}{\|\overline{\mathbf{x}}\|} = \|\mathbf{A}\| \sup_{\overline{\mathbf{x}}\neq\mathbf{0}} \frac{\|\mathbf{B}\overline{\mathbf{x}}\|}{\|\overline{\mathbf{x}}\|} = \|\mathbf{A}\|\|\mathbf{B}\| = \|\mathbf{A}\|.$$

1.47. Mivel $\mathbf{B}$ szinguláris, így van olyan $\overline{\mathbf{x}} \neq \mathbf{0}$, melyre $\mathbf{B}\overline{\mathbf{x}} = \mathbf{0}$. Erre az $\overline{\mathbf{x}}$ vektorra:

$$\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\overline{\mathbf{x}} = \overline{\mathbf{x}},$$

azaz

$$\|\mathbf{A}^{-1}\|\|\mathbf{A} - \mathbf{B}\|\|\overline{\mathbf{x}}\| \geq \|\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\overline{\mathbf{x}}\| = \|\overline{\mathbf{x}}\|,$$

majd $\|\overline{\mathbf{x}}\|$-szel osztva és átrendezve kapjuk a bizonyítandó állítást.

1.48. Azt igazoljuk, hogy tetszőleges pozitív ε számhoz van olyan n_0 index, hogy minden $k > n_0$ esetén

$$\varrho(\mathbf{A}) \leq \|\mathbf{A}^k\|^{1/k} \leq \varrho(\mathbf{A}) + \varepsilon.$$

Ebből ugyanis az állítás már következik.

A bal oldali egyenlőtlenség igazolása: $\|\mathbf{A}^k\| \geq \varrho(\mathbf{A}^k) = (\varrho(\mathbf{A}))^k$, azaz $\varrho(\mathbf{A}) \leq \|\mathbf{A}^k\|^{1/k}$.

A jobb oldali egyenlőtlenség igazolása:

$$\varrho\left(\frac{\mathbf{A}}{\varrho(\mathbf{A}) + \varepsilon}\right) = \left(\frac{\varrho(\mathbf{A})}{\varrho(\mathbf{A}) + \varepsilon}\right) < 1.$$

Emiatt

$$\left(\frac{\mathbf{A}}{\varrho(\mathbf{A}) + \varepsilon}\right)^k$$

elemenként nullához tart, azaz bármilyen normában is nullához tart. Így elegendően nagy k-ra a mátrix normája kisebb lesz 1-nél. Azaz ilyen k értékekre

$$\left\|\left(\frac{\mathbf{A}}{\varrho(\mathbf{A}) + \varepsilon}\right)^k\right\| \leq 1,$$

amit átrendezve a jobb oldali egyenlőtlenség következik.

1.49. Legyen $\overline{\mathbf{y}} \in \mathbb{R}^k$ egy tetszőleges nemnulla vektor! Ekkor

$$\|\mathbf{A}\|_2 = \sup_{\overline{\mathbf{x}} \in \mathbb{R}^n \neq 0} \frac{\|\mathbf{A}\overline{\mathbf{x}}\|_2}{\|\overline{\mathbf{x}}\|_2} \geq \sup_{\begin{bmatrix} \overline{\mathbf{y}} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^n} \frac{\left\| \mathbf{A} \begin{bmatrix} \overline{\mathbf{y}} \\ \mathbf{0} \end{bmatrix} \right\|_2}{\left\| \begin{bmatrix} \overline{\mathbf{y}} \\ \mathbf{0} \end{bmatrix} \right\|_2} \geq \sup_{\overline{\mathbf{y}} \in \mathbb{R}^k} \frac{\|\mathbf{A}^{(k)}\overline{\mathbf{y}}\|_2}{\|\overline{\mathbf{y}}\|_2} = \|\mathbf{A}^{(k)}\|_2.$$

1.50. Megoldás I: A $\mathbf{C}$ mátrix egy M-mátrix, hiszen a főátlón kívül nincs pozitív eleme és a $\overline{\mathbf{g}} = [1, 1, 1]^T$ vektorral beszorozva a $[0.7, 0.8, 0.7]^T$ pozitív vektor adódik. Az M-mátrixok invertálhatók, továbbá a szimmetria miatt az 1-es norma megegyezik a maximumnormával. Így az M-mátrixok inverzének becsléséről szóló tétel alapján:

$$\|\mathbf{C}^{-1}\|_1 = \|\mathbf{C}^{-1}\|_\infty \leq \frac{1}{0.7} = 1.43.$$

Megoldás II: A $\mathbf{C}$ mátrix tulajdonképpen az egységmátrix

$$\mathbf{R} = \begin{bmatrix} 0 & -0.1 & -0.2 \\ -0.1 & 0 & -0.1 \\ -0.2 & -0.1 & 0 \end{bmatrix}$$

mátrixszal való perturbációja. A 3.2. tétel alapján, mivel $\|\mathbf{R}\|_1 = 0.3 < 1$, azért $\mathbf{E} + \mathbf{R}$ invertálható és

$$\|\mathbf{C}^{-1}\|_1 \leq \frac{1}{1 - 0.3} = 1.43.$$

1.51. Az eredményeket MATLAB-bal számítva az alábbi táblázat adódik.

```
 n  Az inverz normája
 1  1.0000e+000
 2  1.8000e+001
 3  4.0800e+002
 4  1.3620e+004
 5  4.1328e+005
 6  1.1865e+007
 7  3.7996e+008
 8  1.2463e+010
 9  3.8871e+011
10  1.2070e+013
```

1.52. $\|\mathbf{H}\|_2 = \varrho(\mathbf{H}) = 1.5671$, $\|\mathbf{H}\|_1 = \|\mathbf{H}\|_\infty = 2.2833$.

Modellalkotás és hibaforrásai

Feladatok kondicionáltsága

2.1. A feladat akkor korrekt kitűzésű, ha $|d| > 2$. Ez a szereplő függvények folytonosságából következik. A $d = \pm 2$ értékek sem megfelelők, mert ezeknek nincs olyan környezetük, melyben egyértelmű megoldás lenne. A kondíciószám:

$$\kappa(d) = \frac{|-1 + \frac{1}{2\sqrt{d^2-4}} 2d| \cdot |d|}{|-d + \sqrt{d^2-4}|} = \frac{|d|}{\sqrt{d^2-4}}.$$

Ez akkor lesz 100-nál nagyobb, ha $2 < |d| < \sqrt{40000/9999}$. Ilyen d pl. a $d = 2.0001$. A feladat jól kondicionált, ha $|d|$ nagy és rosszul, ha értéke közel van 2-höz.

2.2. Az x megoldás az $x = 1/(1-d^2)$ alakban írható. Így a relatív kondíciószám

$$\kappa_1(d) = \frac{2d^2}{|1-d^2|},$$

ami mutatja, hogy 1 közeli d értékekre a feladat rosszul kondicionált. A $d = 0.99$ értékre $\kappa_1(d) = 98.5$. Mivel $x + y = 1/(1+d) = G_2(d)$, így

$$\kappa_2(d) = \frac{d}{1+d},$$

azaz a megoldások összegének kiszámítása jól kondicionált. A $d = 0.99$ értékre $\kappa_2(d) = 0.4975$. (Megjegyezzük, hogy ha d közel van 1-hez, akkor x és y két abszolút értékben nagy szám, ellentétes előjellel, melyek összege kb. 2, így $x + y$ kiszámításakor kiegyszerűsödés léphet fel. Ez viszont már a numerikus számítás tulajdonsága és nem az eredeti feladaté.)

2.3. A képlet alapján $x = \sqrt{d+1} - \sqrt{d}$, azaz a megoldófüggvény $G(d) = \sqrt{d+1} - \sqrt{d}$. Innét látható, hogy minden $d > 0$ esetén a feladat korrekt kitűzésű. A relatív kondíciószám a $\kappa(d) = |d \cdot G'(d)/G(d)|$ képlettel számítható. Erre, egyszerűsítések után, a

$$\kappa(d) = \frac{1}{2}\sqrt{\frac{d}{d+1}}$$

eredményt kapjuk, ami minden pozitív d esetén legfeljebb 1/2 lehet. Ez mutatja, hogy a feladat minden $d > 1$ esetén jól kondicionált (hiszen maximum fele akkora százalékot változik x, mint d).

2.4. Azokban a d pontokban, melyekben a kondíciószámok értelmezhetők

$$\kappa_{fg}(d) = \frac{|d \cdot (f \cdot g)'(d)|}{|(f \cdot g)(d)|} = \frac{|d \cdot (f'(d) \cdot g(d) + f(d) \cdot g'(d))|}{|f(d) \cdot g(d)|} \leq \kappa_f(d) + \kappa_g(d),$$

azaz a szorzat kondíciószáma a két kondíciószám összegével becsülhető felülről.

2.5. Az egyenletrendszer megoldása $d \neq \pm 1$ esetén

$$x = \frac{1}{1-d^2}, \; y = \frac{-d}{1-d^2},$$

így a feladat $d \neq \pm 1$ esetén korrekt kitűzésű. A megoldófüggvény tehát

$$G(d) = \left[\frac{1}{1-d^2}, \frac{-d}{1-d^2}\right].$$

A kondíciószám a (2.1) képlettel számítható.

$$\kappa(d) = \frac{\| \, [2d/(1-d^2)^2, -(1+d^2)/(1-d^2)^2] \, \|_\infty \cdot |d|}{\| \, [1/(1-d^2), -d/(1-d^2)] \, \|_\infty}$$

$$= \begin{cases} 1 + d^2/|1-d^2|, & \text{ha } |d| > 1, \\ (1+d^2)|d|/|1-d^2|, & \text{ha } 0 < |d| < 1, \end{cases}$$

ha pedig $d = 0$, akkor az abszolút kondíciószám számítható: $\kappa_{abs}(0) = 1$.

A gépi számábrázolás

2.6. A pontosan ábrázolható számok a következők: 0, 0.01, ..., 0.09, 0.1,..., 0.9, 1, ..., 9, 10, 20, ..., 90, 100, 200, ...,900, és ezen számok -1-szeresei. Így tehát a legnagyobb ábrázolható szám a 900, a legkisebb pozitív ábrázolható szám a 0.01, a gépi epszilon pedig 0.01.

2.7. $fl(1/3) = 0.3$, $fl(1/900) = 0$, $fl(20 \cdot 200) = Inf$, $fl(((2+0.1)+0.1)+\cdots+0.1) = 2$, $fl((((0.1+0.1)+0.1)+\cdots+0.1)+2) = 3$.

2.8. a) F(1,0,3) , b) F(3,0,0), c) F(1,-3,0), d) F(4,0,3).

2.9. a) $2.2 \cdot 3.45 = 7.59$ ábrázolásához az F(3,0,0) számrendszer kell. b) az 1/80=0.0125 ábrázolásához az F(3,-2,1) számrendszer szükséges. c) $2 \times 10^2 \cdot 7 \times 10^2 = 1.4 \times 10^5$ számításához az F(2,2,5) számrendszerre van szükség.

2.10. A számábrázolás hibáit vesszük figyelembe. Így

$$\hat{z} = \frac{x(1+\delta_x)}{y(1+\delta_y)}(1+\delta),$$

ahol $|\delta_x|, |\delta_y|, |\delta| \leq \mathtt{u}$. Ekkor az abszolút hiba

$$|\hat{z} - z| = \left| \frac{x(1+\delta_x)}{y(1+\delta_y)}(1+\delta) - \frac{x}{y} \right|$$

$$\leq \frac{x}{y} \left| \frac{(1+\mathtt{u})^2}{1-\mathtt{u}} - 1 \right| = \frac{x}{y} \left| (\mathtt{u}^2 + 2\mathtt{u} + 1)(1 + \mathtt{u} + \mathtt{u}^2 + \ldots) - 1 \right| \approx \frac{x}{y} 3\mathtt{u},$$

ahol $\mathtt{u}$ magasabb hatványait elhagytuk. Az abszolút hiba jelentős lehet, ha x jóval nagyobb y-nál. A relatív hiba $3\mathtt{u}$, ami a gépi pontosság nagyságrendje.

2.11. A kiszámolt érték (mindig hatjegyű mantisszára kerekítve): -2×10^{-3}. A hiba a kiegyszerűsödés miatt lép fel, két közeli szám kivonása miatt. Ez elkerülhető közös nevezőre hozással és egyszerűsítéssel. Így

$$A = \frac{-2a}{1-2a},$$

melynek eredménye -2.00401×10^{-3}.

2.12. Természetesen nem. Valahányadik tagtól az $1/n$ értékek már kisebbek lesznek ε_0-nál, így a számítógép ezeket már nullának fogja tekinteni. Az adott esetben - mivel csak normálalakban lévő számokat tudunk ábrázolni - a legkisebb ábrázolható pozitív szám 0.1, így csak az $1 + 1/2 + \ldots + 1/10$ összeget kell kiszámolnunk. Mindig figyelembe véve a használható mantisszahosszt (a törtek számításakor és az összegzéskor is), összegnek 2.9-et kapunk.

2.13. A cos(0.7854) érték 6-jegyű mantisszára kerekítve: 7.07105×10^{-1}. Ennek négyzete 4.99997×10^{-1}. Hasonlóan $\sin^2(0.7854) \approx 5.00002 \times 10^{-1}$. Így $f(0.7854) \approx -5 \times 10^{-6}$. A pontos $f(0.7854)$ érték -3.67321×10^{-6}. A relatív hiba tehát 0.3612. Ez nagy relatív hiba. Oka az, hogy két közeli számot vontunk ki egymásból a számolás során. Ez elkerülhető az $f(x) = \cos(2x)$ formula alkalmazásával. Ezzel, szintén hatjegyű mantisszára kerekítve, -3.67321×10^{-6} adódik a számolás során. Ennek sokkal kisebb a relatív hibája.

2.14. A gyökjel alatt az $a^2 - 4b = 2.5 \times 10^{17} - 4$ értéket kellene kiszámítani, amire a MATLAB 2.5×10^{17}-ent fog adni ($\varepsilon_g \approx 2 \times 10^{-16}$). Ezért a két gyök $x_1 = 5 \times 10^8$ és $x_2 = 0$. Nyilvánvaló, hogy x_1 relatív hibája kicsi, míg az x_2 gyöké nagy. Jobb eredményt érhetünk el, ha észrevesszük, hogy a két gyök szorzata (Viéte-formula) 1, így x_2 jobban számolható úgy, hogy x_1 reciprokát vesszük. Így $x_2 = 2 \times 10^{-9}$ adódik. (Hasonlóan jó megoldás a számláló gyöktelenítése is a konjugálttal való szorzással és osztással.)

2.15. A szimpla pontosságú kettes számrendszerbeli számok esetén a mantissza úgy néz ki, hogy 1-es szerepel a „kettedespont" előtt, utána 23 biten szerepelhetnek 1-esek ill. nullák. Az $a + b$ összeg számítógépen számolt értéke akkor marad a, ha b kisebb, mint 2^{-24}. (Ekkor már kerekítve sem változtat a mantisszán.) Azaz $1/(k+1)^2$ értéke akkor nem adódik hozzá s_k-hoz, ha k legalább 4096. Így a megadott érték, azaz 1.6447253, lesz a számítógépen számolt határérték, azaz az eltérés $\pi^2/6 - 1.6447253 = 2.0877 \times 10^{-4}$.

Jobb eredményt kapunk, ha fordított sorrendben adjuk össze a sor tagjait. Pl.

$$\sum_{i=nmax:-1:1} \frac{1}{i^2},$$

ahol $nmax$ lehet jóval nagyobb, mint 4096.

2.16. A 0.1 szám értéke kettes számrendszerben

$$x = 1.10011001100110011001100...\cdot 2^{-4},$$

szimpla pontosságú lebegőpontos számként (melynél a gépi pontosság $\mathtt{u} = 2^{-24}$) pedig

$$fl(x) = 1.10011001100110011001101 \cdot 2^{-4}.$$

(Itt az utolsó számjegy kerekített lett.) Vonjuk ki $fl(x)$-ből x-et. Azt kapjuk, hogy

$$(2^{-23} - 2^{-24} - 2^{-25} - 2^{-28} - 2^{-29} - \ldots)\cdot 2^{-4} = 2^{-24}(2^{-3} - 2^{-4} - 2^{-5} - 2^{-8} - 2^{-9} - \ldots)$$

$$= 2^{-24}(1/8 - 1/10) = 2^{-24}\frac{1}{40}.$$

Emiatt

$$\frac{x - fl(x)}{x} = \frac{-2^{-24}/40}{0.1} = -\frac{1}{4}\mathtt{u}.$$

2.17.

A feladat megoldható pl. az

```
y=2*sqrt(2);
for k=2:31
y=2^(k+1)*sqrt((1-sqrt(1-((2^-k)*y)^2))/2);
fprintf('%16d   %13.12f\n',2^(k+1),y);
end
```

szkript segítségével. Az eredményt az alábbi táblázatban adjuk meg.

Csúcsszám	Félkerület
4	2.828427124746
8	3.061467458921
16	3.121445152258
32	3.136548490546
64	3.140331156955
128	3.141277250933
256	3.141513801144
512	3.141572940368
1024	3.141587725280
2048	3.141591421505
4096	3.141592345611
8192	3.141592576545
16384	3.141592633463
32768	3.141592654808
65536	3.141592645321
131072	3.141592607376
262144	3.141592910940
524288	3.141594125195
1048576	3.141596553705
2097152	3.141596553705
4194304	3.141674265022
8388608	3.141829681889
16777216	3.142451272494
33554432	3.142451272494
67108864	3.162277660168
134217728	3.162277660168
268435456	3.464101615138
536870912	4.000000000000
1073741824	0.000000000000
2.147484e+009	0.000000000000
4.294967e+009	0.000000000000

Innét jól látszik, hogy bár a sorozat az elején π-hez konvergálónak tűnik, néhány lépés után a sorozat nullává válik, azaz teljesen hibás határértéket ad. Ennek oka a kiegyszerűsödés, ugyanis az iteráció képletében két 1-hez közeli számot vonunk ki egymásból. A módosított iteráció már π-hez konvergáló sorozatot állít elő.

2.18. Azt kapjuk eredményül, hogy a kiegyszerűsödés miatt (pozitív és negatív számokat adunk össze úgy, hogy az összeg nagyon kicsi lesz) több nagyságrendnyi eltérés van a

pontos érték és a számított érték között. Az n érték növelésével az összeg a MATLAB-bal számítva $8.0866e-007$-hoz konvergálónak tűnik (majd egy adott n-től a MATLAB már NaN értéket ad). Ez az érték nagyon messze van a tényleges $1.3888e-011$ értéktől. Ezt az értéket úgy kaphatjuk meg pontosabban, hogy e^{25}-öt számoljuk ki. Ebben nincs kiegyszerűsödés, majd pedig vesszük a kiszámolt szám reciprokát.

2.19. Minden számolást úgy végzünk el, hogy kiszámoljuk pontosan, majd az eredményt 4-jegyű mantisszára kerekítjük (tulajdonképpeni normálalak 3 tizedesjeggyel). A megoldások

$$x_{1,2} = \frac{1634 \pm \sqrt{1634^2 - 4 \cdot 2}}{2}.$$

Az x_2 megoldás értéke amiatt nulla, mert a megoldóképletben szereplő gyökjel alatti kifejezésben nagy számból vonunk ki kicsit, ami nem fog változtatni az eredmény mantisszáján. Így a gyök értéke éppen 1634 marad, és kivonás után nullát kapunk a számlálóban (kiegyszerűsödés). A képlet alapján x_1 értéke 1634-nek adódik. Mivel a gyökök és együtthatók közötti összefüggésből $x_1 x_2 = 2$, így x_1 értékéből $x_2 = 2/x_1$ adódik, amire az 1.224×10^{-3} értéket kapjuk (a korábbi nulla helyett).

Lineáris egyenletrendszerek megoldása

Kondicionáltság

3.1. Válasszuk $\mathbf{B}$-nek a

$$\mathbf{B} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

mátrixot. Ekkor $\|\mathbf{A}\|_\infty = 2.01$ és az 1.47. feladat eredménye miatt $\|\mathbf{A}^{-1}\|_\infty \geq 100$, így a $\kappa_\infty(\mathbf{A}) \geq 201$ becslést nyerjük. Mivel $\|\mathbf{A}^{-1}\|_\infty = 201$, így a keresett kondíciószám 404.01.

3.2.

$$\mathbf{A}^{-1} = \begin{bmatrix} 4 & -6 \\ -6 & 12 \end{bmatrix},$$

így $\kappa_1(\mathbf{A}) = \kappa_\infty(\mathbf{A}) = 1.5 \cdot 18 = 27$, és $\kappa_2(\mathbf{A}) = 1.2676/0.0657 = 19.3$ ($\mathbf{A}$ legnagyobb és legkisebb sajátértékének hányadosa).

Legyen $\overline{\mathbf{x}}^\star$ az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszer megoldása. Ekkor $\overline{\mathbf{b}}$-t 1%-kal megnövelve az $\mathbf{A}(\overline{\mathbf{x}}^\star + \delta\overline{\mathbf{x}}) = 1.01\overline{\mathbf{b}}$ egyenlőséghez jutunk, ahol $\delta\overline{\mathbf{x}}$ becsülendő maximumnormában. A fenti egyenlőségből $\mathbf{A}\delta\overline{\mathbf{x}} = 0.01\overline{\mathbf{b}}$ adódik, azaz

$$\|\delta\overline{\mathbf{x}}\|_\infty = 0.01\|\mathbf{A}^{-1}\overline{\mathbf{b}}\|_\infty \leq 0.01\|\mathbf{A}^{-1}\|_\infty\|\overline{\mathbf{b}}\|_\infty = 0.01 \cdot 18\|\overline{\mathbf{b}}\|_\infty = 0.18\|\overline{\mathbf{b}}\|_\infty.$$

Ezzel a kívánt becslést kaptuk.

3.3. Mivel ortogonális mátrixok 2-es normája 1, így a kondíciószámuk is nyilván 1 2-es normában. A másik irány pedig nem igaz. Pl. az $\mathbf{A} = 2\mathbf{E}$ mátrixra $\kappa_2(\mathbf{A}) = 1$, de nem ortogonális, hiszen $\mathbf{A}^{-1} = (1/2)\mathbf{E} \neq \mathbf{A}^T$.

3.4.

$$\frac{\|\overline{\mathbf{v}} - \overline{\mathbf{u}}\|}{\|\overline{\mathbf{v}}\|} = \frac{\|\mathbf{A}^{-1}\overline{\mathbf{b}} - \mathbf{A}^{-1}\overline{\mathbf{b}}/(1+c)\|}{\|\overline{\mathbf{v}}\|} = \frac{\|\mathbf{A}^{-1}\overline{\mathbf{b}}(1 - 1/(1+c)\|}{\|\overline{\mathbf{v}}\|}$$
$$= \frac{\|\overline{\mathbf{v}}(c/(1+c))\|}{\|\overline{\mathbf{v}}\|} = \frac{|c/(1+c)|\|\overline{\mathbf{v}}\|}{\|\overline{\mathbf{v}}\|} = \left|\frac{c}{1+c}\right|.$$

3.5. Az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ és $\mathbf{A}\overline{\mathbf{x}}^\star = \overline{\mathbf{b}} + \delta\overline{\mathbf{b}}$ egyenlőségekből kivonás után kapjuk, hogy

$$\mathbf{A}(\overline{\mathbf{x}}^\star - \overline{\mathbf{x}}) = \delta\overline{\mathbf{b}},$$

amiből

$$\overline{\mathbf{x}}^\star - \overline{\mathbf{x}} = \mathbf{A}^{-1}\delta\overline{\mathbf{b}}$$

adódik. Tehát

$$\|\overline{\mathbf{x}}^\star - \overline{\mathbf{x}}\| = \|\mathbf{A}^{-1}\delta\overline{\mathbf{b}}\| \leq \|\mathbf{A}^{-1}\| \cdot \|\delta\overline{\mathbf{b}}\|.$$

Ezen becslés alapján, ha az

$$\begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix} \overline{\mathbf{x}} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$$

lineáris egyenletrendszer jobb oldalához hozzáadjuk az $[\varepsilon_1, \varepsilon_2]^T$ vektort, akkor a megoldás megváltozása 2-es normában maximum

$$\left\| \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}^{-1} \right\|_2 \left\| \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \right\|_2$$

lehet. Mivel az

$$\begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}^{-1} = \frac{1}{-5} \begin{bmatrix} -1 & -2 \\ -2 & 1 \end{bmatrix}$$

mátrix szimmetrikus, így 2-es normája megegyezik a spektrálsugarával. A

$$(1/5 - \lambda)(-1/5 - \lambda) - 4/25 = 0$$

egyenletet kell megoldani. Ennek megoldásai $\pm 1/\sqrt{5}$. Így a spektrálsugár $1/\sqrt{5}$.

A $\delta\overline{\mathbf{b}}$ vektor kettes normája $\sqrt{\varepsilon_1^2 + \varepsilon_2^2}$, melyre $\sqrt{2} \cdot 10^{-4}$ egy megfelelő felső becslés.

A megoldás megváltozására vonatkozó megfelelő felső becslés tehát

$$\frac{1}{\sqrt{5}} \cdot \sqrt{2} \cdot 10^{-4} \approx 6.3246 \cdot 10^{-5}.$$

3.6. Induljunk ki egy tetszőleges $\overline{\mathbf{x}}$ vektor esetén az $\mathbf{A}(\overline{\mathbf{x}} - \overline{\mathbf{x}}^\star)$ egyenlőségből, ahol $\overline{\mathbf{x}}^\star$ az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszer megoldása. Ekkor $\mathbf{A}(\overline{\mathbf{x}} - \overline{\mathbf{x}}^\star) = \mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}} = -\overline{\mathbf{r}}$, azaz $\|\mathbf{A}\|\|\overline{\mathbf{x}} - \overline{\mathbf{x}}^\star\| \geq \|\overline{\mathbf{r}}\|$ valamilyen vektornormában és az általa indukált mátrixnormában. Másrészt $\overline{\mathbf{x}} - \overline{\mathbf{x}}\star = -\mathbf{A}^{-1}\overline{\mathbf{r}}$, azaz $\|\overline{\mathbf{x}} - \overline{\mathbf{x}}^\star\| \leq \|\mathbf{A}^{-1}\|\|\overline{\mathbf{r}}\|$. A fenti két egyenlőtlenségből az alábbi becsléseket kapjuk:

$$\frac{\|\mathbf{A}\|}{\|\overline{\mathbf{r}}\|} \leq \|\overline{\mathbf{x}} - \overline{\mathbf{x}}^\star\| \leq \|\mathbf{A}^{-1}\|\|\overline{\mathbf{r}}\|.$$

Ez mutatja, hogy abból, hogy a maradékvektor kicsi normájú, csak akkor következik, hogy az $\overline{\mathbf{x}}$ vektor közel van az egyenletrendszer megoldásához, ha $\|\mathbf{A}^{-1}\|$ kicsi.

A példában $\|\mathbf{A}\|_\infty = \|\mathbf{A}^{-1}\|_\infty = 144$, így az első $\overline{\mathbf{x}}$ vektorra vonatkozó felső becslés $144 \cdot 0.01 = 1.44$, a másodikra vonatkozó pedig $144 \cdot 1.44 = 207.36$. Ennek ellenére a második vektor van közelebb a pontos megoldáshoz, ami $\overline{\mathbf{x}}^\star = [-1, 1]^T$.

3.7.

$$\|\mathbf{A}\|_2^2 = \varrho(\mathbf{A}^T\mathbf{A}) \le \|\mathbf{A}^T\mathbf{A}\|_\infty \le \|\mathbf{A}^T\|_\infty\|\mathbf{A}\|_\infty \le \|\mathbf{A}\|_1\|\mathbf{A}\|_\infty,$$

ahol felhasználtuk, hogy egy mátrix maximumnormája megegyezik a transzponáltjának 1-es normájával. A második állítás az első állítás segítségével

$$\kappa_2^2(\mathbf{A}) = \|\mathbf{A}\|_2^2\|\mathbf{A}^{-1}\|_2^2 \le \|\mathbf{A}\|_1\|\mathbf{A}\|_\infty\|\mathbf{A}^{-1}\|_1\|\mathbf{A}^{-1}\|_\infty = \kappa_1(\mathbf{A})\kappa_\infty(\mathbf{A})$$

módon adódik.

3.8. A becslések közvetlenül következnek az 1.34. feladatban igazolt becslésekből.

3.9. A mátrix determinánsa minden n esetén 1, maximumnormája pedig n. Mivel a mátrix inverze

$$\begin{bmatrix} 1 & 2^0 & 2^1 & \dots & 2^{n-2} \\ 0 & 1 & 2^0 & \dots & 2^{n-3} \\ \vdots & 0 & \ddots & \ddots & 2^1 \\ \vdots & \ddots & \ddots & 1 & 2^0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

alakú, így inverzének maximumnormája 2^{n-1}, azaz a mátrix kondíciószáma $n2^{n-1}$. Látható, hogy míg a determináns mindig 1, a kondíciószám n növelésével exponenciálisan növekszik.

3.10. Az egyenlőtlenség nyilvánvalóan következik abból, hogy indukált mátrixnormákban a kondíciószám nem lehet kisebb 1-nél.

Az egyenlőség igazolásához először lássuk be, hogy egy négyzetes, invertálható mátrixra $\|\mathbf{A}\|_2 = \|\mathbf{A}^T\|_2$. Ez egyszerűen következik abból, hogy az $\mathbf{A}\mathbf{A}^T$ és $\mathbf{A}^T\mathbf{A}$ mátrixok karakterisztikus polinomjai megegyeznek, így spektrálsugaruk is azonos:

$$\det(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{E}) = \det(\mathbf{A}^T)\det(\mathbf{A} - \lambda(\mathbf{A}^T)^{-1})$$
$$= \det(\mathbf{A} - \lambda(\mathbf{A}^T)^{-1})\det(\mathbf{A}^T) = \det(\mathbf{A}\mathbf{A}^T - \lambda\mathbf{E}).$$

Ezek után a feladatban szereplő egyenlőség az alábbi módon igazolható:

$$\kappa_2^2(\mathbf{A}^T\mathbf{A}) = \|\mathbf{A}^T\mathbf{A}\|_2^2\|(\mathbf{A}^T\mathbf{A})^{-1}\|_2^2 = \varrho((\mathbf{A}^T\mathbf{A})\mathbf{A}^T\mathbf{A})\varrho((\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{A})^{-1})$$
$$= \varrho^2(\mathbf{A}^T\mathbf{A})\varrho^2((\mathbf{A}^T\mathbf{A})^{-1}) = \|\mathbf{A}\|_2^4\|(\mathbf{A}^{-1})^T\|_2^4 = \|\mathbf{A}\|_2^4\|\mathbf{A}^{-1}\|_2^4 = \kappa_2^4(\mathbf{A}).$$

3.11. Tegyük fel, hogy $\mathbf{A}$ és $\mathbf{B}$ ortogonálisan hasonlók, azaz létezik olyan $\mathbf{S}$ ortogonális mátrix, mellyel $\mathbf{B} = \mathbf{S}^T\mathbf{A}\mathbf{S}$. Ekkor

$$\|\mathbf{B}\|_2^2 = \varrho(\mathbf{B}^T\mathbf{B}) = \varrho(\mathbf{S}^T\mathbf{A}^T\mathbf{S}\mathbf{S}^T\mathbf{A}\mathbf{S}) = \varrho(\mathbf{S}^T\mathbf{A}^T\mathbf{A}\mathbf{S}) = \|\mathbf{A}\mathbf{S}\|_2^2 = \|\mathbf{A}\|_2^2,$$

ahol kihasználtuk a 2-es norma képletét, ill. hogy ortogonális mátrixszal való szorzás nem változtatja meg a 2-es normát.

A második egyenlőség az előbb igazolt egyenlőség következménye.

Direkt módszerek

3.12. Mivel $|\delta a_{ij}/a_{ij}| \leq 0.01\% = 10^{-4}$, azaz $|\delta a_{ij}| \leq 10^{-4}|a_{ij}|$, ezért $\|\delta\mathbf{A}\|_\infty/\|\mathbf{A}\|_\infty \leq 10^{-4}$. Hasonlóan kapjuk, hogy $\|\delta\overline{\mathbf{b}}\|_\infty/\|\overline{\mathbf{b}}\|_\infty \leq 10^{-4}$. A mátrixokról leolvasható, hogy $\|\mathbf{A}\|_\infty = 4$ és $\|\mathbf{A}^{-1}\|_\infty = 1.9$. Ebből a 3.1. tétel alapján kapjuk, hogy

$$\|\delta\overline{\mathbf{x}}\|_\infty/\|\overline{\mathbf{x}}\|_\infty \leq \frac{4 \cdot 1.91}{1 - 4 \cdot 1.91 \cdot 10^{-4}}(10^{-4} + 10^{-4}) = 0.00153.$$

3.13. Mivel a mátrix elemei nem hibával terheltek, így $\|\delta\mathbf{A}\| = 0$, valamint a szövegből kiderül, hogy $\|\delta\overline{\mathbf{b}}\|_\infty = 0.1$, $\kappa_\infty(\mathbf{A}) = 11 \cdot 0.2824$. A 3.1. tételt alkalmazva a relatív hibára 0.1035 adódik.

3.14. A mátrix LU-felbontását a Gauss-eliminációs eljárás során kaphatjuk meg, így lényegében az egyenletrendszert kell csak megoldanunk.

Az induló eliminációs táblázat a következő alakú:

$$\left.\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 2 \\ 1 & 4 & 9 & 16 & 10 \\ 1 & 8 & 27 & 64 & 44 \\ 1 & 16 & 81 & 256 & 190 \end{array}\right. .$$

Először az első sor első elemével nullázzuk le az első oszlop főátló alatti elemeit. Ehhez az első sor 1-szeresét kell kivonni a második, a harmadik és a negyedik sorokból. Ezek a szorzók megmondják, hogy mik lesznek az $\mathbf{L}$ mátrix első oszlopának elemei: $l_{21} = l_{31} = l_{41} = 1$. Ezzel az alábbi táblázatot nyertük:

$$\left.\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 2 \\ 0 & 2 & 6 & 12 & 8 \\ 0 & 6 & 24 & 60 & 42 \\ 0 & 14 & 78 & 252 & 188 \end{array}\right. .$$

A következő lépésben a második sor második elemével nullázzuk le a második oszlop főátló alatti részét. A harmadik sorból a második háromszorosát, a negyedikből pedig a hétszeresét kell kivonni. Így $l_{32} = 3$ és $l_{42} = 7$. Az újabb táblázat:

$$\left.\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 2 \\ 0 & 2 & 3 & 8 & 6 \\ 0 & 0 & 6 & 24 & 18 \\ 0 & 0 & 36 & 168 & 132 \end{array}\right. .$$

Hasonlóan járunk el a harmadik oszloppal is. Így $l_{43} = 6$ és az új táblázat:

$$\left[\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 2 \\ 0 & 2 & 3 & 8 & 6 \\ 0 & 0 & 6 & 24 & 18 \\ 0 & 0 & 0 & 248 & 24 \end{array}\right]. \tag{10.1}$$

Az itt szereplő mátrix első négy oszlopa adja az LU-felbontás $\mathbf{U}$ mátrixát. Az $\mathbf{L}$ mátrix pedig a fent meghatározott elemekből és abból határozható meg, hogy a főátlójában csupa 1-esek állnak. Így tehát az alábbi LU-felbontást kapjuk:

$$\mathbf{A} = \mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 1 & 7 & 6 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 0 & 24 \end{bmatrix}.$$

Az egyenletrendszer megoldásához ezek után úgy jutunk el, hogy a 10.1 alakú egyenletrendszert visszahelyettesítéssel megoldjuk. Itt először x_4 határozható meg a negyedik egyenletből: $x_4 = 1$. Ezek után x_3-at határozzuk meg a harmadik egyenletből: $x_3 = -1$. Hasonlóan kapjuk, hogy $x_2 = 1$ és $x_1 = -1$.

Az $\mathbf{A}$ mátrix determinánsa megegyezik az $\mathbf{U}$ mátrix determinánsával, ami pedig a főátlóbeli elemek szorzata. Tehát a determináns 288.

3.15.
$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ -1/3 & 1 & 0 \\ 0 & -1/3 & 1 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

3.16. Az, hogy van LU-felbontás, következik abból, hogy az $(n-1)$-edik rendig bezáróan a bal felső sarokdeterminánsok nullától különbözőek (mindegyik 1).

3.17. Az első oszlop eliminálása során a főátló alatti nullák helyére nemnulla elemek kerülnek (ezt a jelenséget feltöltődésnek hívjuk), így a többi oszloppal is végre kell hajtanunk az eliminációt. A jelenség ebben a feladatban úgy küszöbölhető ki, ha felcseréljük az első és utolsó oszlopot, azaz változócserét hajtunk végre az x_1 és x_n ismeretlenekkel.

3.18. Olvassuk ki a mátrixból az $\mathbf{L}$ és $\mathbf{U}$ mátrixokat! Mivel $\mathbf{U}$ főátlójának minden eleme pozitív, így az eredeti mátrix minden főminorja pozitív, azaz a mátrix szimmetrikus (ez a feladat szövegéből derül ki) és pozitív definit. Legyen $\mathbf{D}$ az $\mathbf{U}$ mátrix főátlómátrixa. Ekkor $\mathbf{G} = \mathbf{L}\sqrt{\mathbf{D}}$ ($\sqrt{\mathbf{D}}$-t úgy kapjuk, hogy $\mathbf{D}$ minden eleméből gyököt vonunk), azaz

$$\mathbf{G} = \begin{bmatrix} \sqrt{2} & 0 & 0 & 0 \\ 3\sqrt{2}/2 & \sqrt{6}/2 & 0 & 0 \\ \sqrt{2} & 2\sqrt{6}/3 & \sqrt{21}/3 & 0 \\ 2\sqrt{2} & \sqrt{6} & 3\sqrt{21}/7 & \sqrt{7}/7 \end{bmatrix}.$$

Az adott egyenletrendszert úgy oldhatjuk meg a leggyorsabban, ha először megoldjuk az $\mathbf{L}\overline{\mathbf{y}} = [1,0,0,0]^T$ egyenletrendszert, amely egyszerű visszahelyettesítéssel megoldható. Megoldásnak az $\overline{\mathbf{y}} = [3,-3/2,1,-2/7]$ vektor adódik, majd pedig az $\mathbf{U}\overline{\mathbf{x}} = \overline{\mathbf{y}}$ egyenlet megoldásával (szintén egyszerű visszahelyettesítéssel) adódik az $\overline{\mathbf{x}} = [3,-1,3,-2]^T$ megoldás.

3.19. Legyen az együtthatómátrix $n \times n$-es, a $\overline{\mathbf{b}}$ vektorral bővítve pedig $n \times (n+1)$-es. A k. oszlop eliminálásakor $n-1$ szorzótényezőt kell kiszámolnunk, majd a k. sor $k+1,\dots,n+1$ elemeit ezen szorzótényezőkkel rendre beszorozva az $1,\dots,n$ sorokból (kivéve a k.-at) kivonni. Ez $2(n-k+1)$ művelet. Az elimináció tehát összesen

$$\sum_{k=1}^{n-1}(n-1+(n-1)(2(n-k+1))) = \sum_{k=1}^{n-1}(n-1+2(n^2-1)-2(n-1)k)$$

$$= (n-1)(n-1+2(n^2-1))-2(n-1)\sum_{k=1}^{n-1}k = (n-1)(n-1+2(n^2-1))-(n-1)(n-1)n$$

$$= n^3+n^2-5n+3$$

műveletet jelent. Azután már csak egy olyan egyenletrendszert kell megoldani, melynek együtthatómátrixa diagonális mátrix. Ehhez n osztásra van szükség. Így az összes műveletigény n^3+n^2-4n+3.

3.20. Négy módszert fogunk vizsgálni.

Az első módszer a Gauss-módszer. Ebben az esetben az $\mathbf{A}(\mathbf{A}^{-1})_j = \overline{\mathbf{e}}_j$ $(j=1,\dots,n)$ egyenletrendszereket oldjuk meg egyszerre az $\mathbf{A}^{-1}$ mátrix $(\mathbf{A}^{-1})_j$ oszlopvektoraira a Gauss-módszer segítségével. Az $[\mathbf{A}|\mathbf{E}]$ mátrixra végrehajtjuk először az eliminációt. Itt a k. lépésben ki kell számolni az eliminációs szorzókat $n-k$ sorhoz, majd ezekkel a k. sor n darab elemét kell megszoroznunk és a megfelelő sorokból ($n-k$ darab) kivonnunk. Ennek műveletszáma

$$\sum_{k=1}^{n-1}(n-k+2n(n-k)) = n^3-\frac{1}{2}n^2-\frac{1}{2}n = n^3+O(n^2).$$

Ezek után még n darab felső háromszögmátrixú lineáris egyenletrendszert kell megoldani visszahelyettesítéssel. Ez $n\cdot n^2$ művelet. Tehát ez a módszer összesen

$$2n^3+O(n^2)$$

műveletet igényel.

A második a Gauss–Jordan-módszer. Ebben az esetben az $\mathbf{A}(\mathbf{A}^{-1})_j = \overline{\mathbf{e}}_j$ $(j=1,\dots,n)$ egyenletrendszereket a Gauss–Jordan-módszerrel oldjuk meg (lásd 3.19. feladat). Itt a k. lépésben ki kell számolni az eliminációs szorzókat $n-1$ sorhoz, majd

ezekkel a k. sor n darab elemét kell megszoroznunk és a megfelelő sorokból kivonnunk. Végül a főátló elemeivel le kell osztanunk a sorokat. Ez összesen

$$\sum_{k=1}^{n} (n - 1 + 2n(n-1)) + n^2 = 2n^3 - n = 2n^3 + O(n^2)$$

művelet.

A harmadik lehetőség, hogy elkészítjük az LU-felbontást, ennek ismeretében oldjuk meg az $\mathbf{A}(\mathbf{A}^{-1})_j = \overline{\mathbf{e}}_j$ $(j = 1, \ldots, n)$ egyenletrendszereket darabonként $2n^2$ művelettel. Ez összesen

$$\frac{2}{3}n^3 + O(n^2) + 2nn^2 = \frac{8}{3}n^3 + O(n^2)$$

művelet.

A negyedik módszerben szintén az LU-felbontást állítjuk elő először, majd abból az inverz mátrixot az $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$ módon számítjuk ki.

Egy $\mathbf{T}$ alsó háromszögmátrix $\mathbf{V}$ inverzét (ami szintén alsó háromszögmátrix lesz) az alábbi módon határozhatjuk meg:

$$v_{ii} = \frac{1}{t_{ii}}, \quad v_{ij} = \frac{-1}{t_{ii}} \left(\sum_{k=j}^{i-1} t_{ik} v_{kj} \right) \quad (i > j).$$

Ennek műveletszáma

$$\frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n = \frac{1}{3}n^3 + O(n^2).$$

Felső háromszögmátrixra a műveletszám ugyanekkora, majd pedig egy felső és egy alsó háromszögmátrixot kell összeszoroznunk, melynek műveletszáma

$$\sum_{k=1}^{n} (2k-1)(2n - (2k-1)) = \frac{1}{3}n + \frac{2}{3}n^3 = \frac{2}{3}n^3 + O(n).$$

Így az összes műveletszám

$$\frac{4}{3}n^3 + O(n^2).$$

A számolások mutatják, hogy a harmadik módszer a leglassabb, a másik három pedig nagyjából ugyanannyi idő alatt állítja elő egy mátrix inverzét. Az is látszik, hogy bár egyenletrendszer megoldásra a Gauss–Jordan-módszer használata nem célszerű a Gauss-módszerrel szemben, inverz mátrix számolása esetén a két módszernek lényegében ugyanakkora a műveletszáma.

3.21. Először meghatározzuk az LU-felbontást. Az $\mathbf{L}$ mátrix az LU-felbontás $\mathbf{L}$ mátrixa lesz, $\mathbf{D}$ az LU-felbontás $\mathbf{U}$ mátrixának diagonálisa és $\mathbf{M}^T$ a $\mathbf{D}^{-1}\mathbf{U}$ mátrix lesz. Így azt kapjuk, hogy

$$\mathbf{B} = \begin{bmatrix} 1 & -2 & 1 \\ 2 & -2 & -4 \\ 2 & 2 & -13 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 3 & 1 \end{bmatrix}}_{\mathbf{L}} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} 1 & -2 & 1 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{M}^T}.$$

3.22. A mátrix szimmetrikus, pozitív definit. Emiatt van Cholesky-felbontása. Praktikus először ezt meghatározni.

$$\mathbf{B} = \begin{bmatrix} \sqrt{2} & 0 \\ 1/\sqrt{2} & \sqrt{6}/2 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 1/\sqrt{2} \\ 0 & \sqrt{6}/2 \end{bmatrix}.$$

Ezek után az LDL^T felbontás úgy állítható elő, hogy az első tényezőből jobbra, a másodikból balra kiemeljük a diagonálisukat tartalmazó diagonális mátrixot.

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1/2 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{6}/2 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{6}/2 \end{bmatrix} \begin{bmatrix} 1 & 1/2 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ 1/2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3/2 \end{bmatrix} \begin{bmatrix} 1 & 1/2 \\ 0 & 1 \end{bmatrix}.$$

Ez az LDL^T felbontás.

3.23. A Cholesky-felbontások $\mathbf{G}$ mátrixai

$$\mathbf{G}_1 = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ (-1/3)\sqrt{3} & (2/3)\sqrt{6} & 0 \\ 0 & (-1/4)\sqrt{6} & (1/4)\sqrt{42} \end{bmatrix},$$

$$\mathbf{G}_2 = \begin{bmatrix} 2 & 0 & 0 \\ -1/2 & (1/2)\sqrt{15} & 0 \\ 0 & (-2/15)\sqrt{15} & (2/15)\sqrt{210} \end{bmatrix}.$$

3.24. Az alábbi mátrixokat kapjuk a felbontásokban:

$$\mathbf{U} = \begin{bmatrix} 6 & 4 & 4 \\ 0 & 28/3 & 16/3 \\ 0 & 0 & 2/7 \end{bmatrix},$$

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 2/3 & 1 & 0 \\ 2/3 & 4/7 & 1 \end{bmatrix},$$

$$\mathbf{G} = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 2\sqrt{6}/3 & 2\sqrt{21}/3 & 0 \\ 2\sqrt{6}/3 & 8\sqrt{21}/21 & \sqrt{14}/7 \end{bmatrix}.$$

3.25. Minden művelet után az eredményt 4-jegyű mantisszára kerekítjük. Pl.

$$5.291/0.003 = 1763.\mathbf{666}\ldots,$$

ami négyjegyű mantisszára kerekítve 1764. Így a második sor második elemére -104300 adódik:

$$\underbrace{-6.13 - \overbrace{59.14 \cdot 1764}^{1043\mathbf{22.96}\to 104300 = 1.043\times 10^5}}_{-1043\mathbf{06.13}\to -104300}.$$

Tehát

$$\begin{array}{cc|c} 0.003 & 59.14 & 59.17 \\ 5.291 & -6.13 & 46.78 \end{array} \to \begin{array}{cc|c} 0.003 & 59.14 & 59.17 \\ 0 & -104300 & -104400 \end{array}.$$

Innét $x_1 = -10$ és $x_2 = 1.001$. Teljes főelemkiválasztáshoz az első két oszlopot kell felcserélni (a változók felcserélődnek).

$$\begin{array}{cc|c} 59.14 & 0.003 & 59.17 \\ -6.13 & 5.291 & 46.78 \end{array} \to \begin{array}{cc|c} 59.14 & 0.003 & 59.17 \\ 0 & 5.291 & 52.92 \end{array}$$

Innét $x_1 = 10$ és $x_2 = 1$. (Ez a pontos megoldás.)

$$\left\| \begin{bmatrix} -10 \\ 1.001 \end{bmatrix} - \begin{bmatrix} 10 \\ 1 \end{bmatrix} \right\|_\infty = 20.$$

3.26. A feladat szerint tehát minden számot $x.xxxxx \cdot 10^k$ alakra kerekítünk, ahol a tizedespont előtti x nullától különböző. A kiindulási egyenletrendszer tehát

$$\begin{array}{ccc|c} 0.00001 & 2 & 3 & 5.00001 \\ 1 & 2 & 3 & 6 \\ 10 & 3 & 4 & 17 \end{array}.$$

Először kicseréljük az első és harmadik sorokat a részleges főelemkiválasztás miatt:

$$\begin{array}{ccc|c} 10 & 3 & 4 & 17 \\ 1 & 2 & 3 & 6 \\ 0.00001 & 2 & 3 & 5.00001 \end{array}.$$

Az első oszlopot eliminájuk

$$\left.\begin{array}{ccc|c} 10 & 3 & 4 & 17 \\ 0 & 1.7 & 2.6 & 4.3 \\ 0 & 2 & 3 & 4.99999 \end{array}\right. .$$

Itt pl. $2-3\cdot 10^{-6} = 1.99999700000$, kerekítve 2, de pl. $5.00001-17\cdot 10^{-6} = 4.99999300000$, kerekítve 4.99999. Most a második és harmadik sorokat cseréljük:

$$\left.\begin{array}{ccc|c} 10 & 3 & 4 & 17 \\ 0 & 2 & 3 & 4.99999 \\ 0 & 1.7 & 2.6 & 4.3 \end{array}\right. ,$$

majd eliminálunk

$$\left.\begin{array}{ccc|c} 10 & 3 & 4 & 17 \\ 0 & 2 & 3 & 4.99999 \\ 0 & 0 & 0.05 & 5.001 \times 10^{-2} \end{array}\right. .$$

Ezután visszahelyettesítéssel $x = 1.00001$, $y = 0.999695$ és $z = 1.0002$ adódik.

3.27. Legyen $\overline{\mathbf{x}} = [2, 1, 2]^T$. Ekkor $\overline{\mathbf{v}} = \overline{\mathbf{x}} \pm \|\overline{\mathbf{x}}\|_2 \overline{\mathbf{e}}_1$, majd a $\overline{\mathbf{v}}$ vektorral meghatározzuk a tükrözési mátrixot a $\mathbf{H} = \mathbf{E} - 2\overline{\mathbf{v}}\overline{\mathbf{v}}^T/(\overline{\mathbf{v}}^T\overline{\mathbf{v}})$ képlet segítségével. A $\pm$ előjelnek megfelelően a lehetséges két tükrözés

$$\mathbf{H} = \begin{bmatrix} -2/3 & -1/3 & -2/3 \\ -1/3 & 14/15 & -2/15 \\ -2/3 & -2/15 & 11/15 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 2/3 & 1/3 & 2/3 \\ 1/3 & 2/3 & -2/3 \\ 2/3 & -2/3 & -1/3 \end{bmatrix}.$$

3.28. Az első oszlophoz tartozó $\overline{\mathbf{v}}$ vektor (a képletben + jellel számolva) $\overline{\mathbf{v}} = [1, 0, 1]^T$. Így a $\mathbf{H}_1$ mátrix

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

azaz

$$\mathbf{H}_1\mathbf{A} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 0 & -1 \end{bmatrix}.$$

A második oszlop 2. és 3. eleméhez, mint kételemű vektorhoz tartozó $\overline{\mathbf{v}}$ vektor $[1, -1]^T$ (+ jellel számolva), így

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Így

$$\mathbf{R} = \mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{bmatrix} -1 & -1 \\ 0 & -1 \\ 0 & 0 \end{bmatrix}$$

és

$$\mathbf{Q} = \mathbf{H}_1^T\mathbf{H}_2^T = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix}.$$

Természetesen a $\overline{\mathbf{v}}$ vektorok másfajta számolása esetén másfajta felbontást kapunk.

3.29. A felbontás megadható pl. Householder-tükrözéssel, amit a második oszlop utolsó két eleméből álló vektorra alkalmazunk. A következő QR-felbontást nyerhetjük:

$$\begin{bmatrix} 4 & 2 & 1 \\ 0 & 3 & 0 \\ 0 & 4 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -3/5 & -4/5 \\ 0 & -4/5 & 3/5 \end{bmatrix} \begin{bmatrix} 4 & 2 & 1 \\ 0 & -5 & -12/5 \\ 0 & 0 & 9/5 \end{bmatrix}.$$

3.30.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2/3} & 1/\sqrt{3} \\ 0 & -1/\sqrt{3} & \sqrt{2/3} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 \\ 0 & \sqrt{2} \\ 1/\sqrt{2} & \sqrt{2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & \sqrt{3} \\ 0 & 0 \end{bmatrix}.$$

3.31. Leolvasható, hogy

$$\mathbf{R} = \begin{bmatrix} -1 & -1 \\ 0 & -1 \\ 0 & 0 \end{bmatrix}, \ \overline{\mathbf{v}}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \ \overline{\mathbf{v}}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

A $\overline{\mathbf{v}}_1$ vektorral megkonstruáljuk a $\mathbf{H}_1$ mátrixot és a $\overline{\mathbf{v}}_2$ vektorral a $\tilde{\mathbf{H}}_2$ mátrixot, amely a $\mathbf{H}_2$ mátrix $\mathbf{H}_2(2:3, 2:3)$ almátrixa lesz. $(\mathbf{H}_2)_{11} = 1$. (A többi elem nulla.) Ebből

$$\mathbf{A} = \mathbf{H}_1\mathbf{H}_2\mathbf{R} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}.$$

3.32. A feltételek mellett a szereplő $\mathbf{Q}$ és $\mathbf{R}$ mátrixok nemszingulárisak. A $\mathbf{Q}_1\mathbf{R}_1 = \mathbf{Q}_2\mathbf{R}_2$ egyenlőségből $\mathbf{R}_1\mathbf{R}_2^{-1} = \mathbf{Q}_1^T\mathbf{Q}_2$ következik, ahol a $\mathbf{Q}$ mátrixok ortogonalitását használtuk. Jelöljük az $\mathbf{R}_1\mathbf{R}_2^{-1}$ mátrixot $\mathbf{D}$-vel. Ez felső háromszögmátrix, másrészt az $\mathbf{R}_1\mathbf{R}_2^{-1} = \mathbf{Q}_1^T\mathbf{Q}_2$ egyenlőség miatt ortogonális, azaz inverze a transzponáltja. Mivel felső háromszögmátrixok inverze felső háromszögmátrix, így az inverze csak úgy lehet a transzponáltja (ami alsó háromszögmátrix), ha $\mathbf{D}$ diagonális. $\mathbf{D}$ ortogonalitása miatt $\mathbf{D}^{-1} = \mathbf{D}^T = \mathbf{D}$, azaz $\mathbf{D}^2 = \mathbf{E}$. Így a $\mathbf{D} = \mathbf{R}_1\mathbf{R}_2^{-1} = \mathbf{Q}_1^T\mathbf{Q}_2$ egyenlőségből következik az állítás. Az állítás pedig közvetlenül azt jelenti, hogy pozitív főátlójú $\mathbf{R}$ mátrixszal a QR-felbontás egyértelmű.

3.33. A k. oszlop eliminálásánál kell egy osztás (a főátló alatt csak egy nemnulla elem van) az $l_{k+1,k}$ kiszámításához, majd a k. sor $k+1:n$ elemeinek $l_{k+1,k}$-szorosát kivonjuk a $k+1$. sor $k+1:n$ elemeiből. Ez $1+2(n-k)$ flop, azaz összesen

$$\sum_{k=1}^{n-1}(1+2(n-k)) = n-1+2\sum_{k=1}^{n-1}(n-k) = n-1+2\frac{(n-1)n}{2} = n^2-1 \ flop.$$

Az $\mathbf{U}$ mátrix felső háromszögmátrix lesz, $\mathbf{L}$ pedig olyan mátrix, hogy a főátló felett és a szubdiagonális alatt nullák vannak és a főátlóban egyesek.

Ha kész az LU-felbontás, akkor két visszahelyettesítés kell. Egy az $\mathbf{U}$ mátrixszal, ez n^2 flop és egy az $\mathbf{L}$ mátrixszal, ami $2(n-1)$ flop. Tehát összesen n^2-2n-2 flopba kerül a megoldás.

Iterációs módszerek

Klasszikus iterációs módszerek

3.34. Az iterációs mátrix a $\mathbf{B}_{GS(\omega)}$ alsó háromszögmátrix lesz; minden főátlóbeli eleme $1-\omega$. Így $\varrho(\mathbf{B}_{GS(\omega)}) = |1-\omega| < 1$ feltétele $0<\omega<2$, és a spektrálsugár akkor a legkisebb (nulla), ha $\omega = 1$ (Gauss–Seidel-módszer). Ekkor lesz a leggyorsabb a konvergencia.

3.35. Az iteráció a következő

$$\overline{\mathbf{x}}^{(k+1)} = \underbrace{\begin{bmatrix} 0 & -1/2 \\ -1/2 & 0 \end{bmatrix}}_{=\mathbf{B}_J} \overline{\mathbf{x}}^{(k)} + \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}.$$

Ha a nullvektorról indítjuk az iterációt, akkor $\overline{\mathbf{x}}^{(1)} = [1/2, 1/2]^T$ és $\overline{\mathbf{x}}^{(2)} = [1/4, 1/4]^T$. Mivel $\|\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(0)}\|_\infty = 1/2$ és $\|\mathbf{B}_J\|_\infty = 1/2$, így a hibabecslés

$$\|\overline{\mathbf{x}}^{(j)} - \overline{\mathbf{x}}^{(0)}\|_\infty \le \frac{(1/2)^j}{1-1/2}\frac{1}{2} < 10^{-6}.$$

Innét kapjuk, hogy a 20. tag már teljesíti a feltételt.

3.36. A Jacobi-módszer iterációs mátrixa $\mathbf{B}_J = \mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})$, amely most a

$$\mathbf{B}_J = (2\mathbf{E})^{-1}(-\mathbf{A}+2\mathbf{E})$$

alakban írható, melyet átalakítva $\mathbf{B}_J = (1/2)(-\mathbf{A}+2\mathbf{E})$ adódik. Ennek sajátértékei $(1/2)(2-\lambda_k) = 1-\lambda_k/2 = \cos(k\pi/(n+1))$ alakúak. A spektrálsugár $k=1$-re adódik $\varrho(\mathbf{B}_J) = \cos(\pi/(n+1))$. Mivel ez 1-nél kisebb, így a módszer mindig konvergens lesz. Nagy n-ekre lassú a konvergencia.

3.37.

$$\mathbf{B}_J = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix}, \quad \mathbf{B}_{GS} = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/4 \end{bmatrix}.$$

Így $\varrho(\mathbf{B}_J) = 1/2$ és $\varrho(\mathbf{B}_J) = 1/4$. A Gauss–Seidel-módszer konvergál gyorsabban. Az iteráció

$$\overline{\mathbf{x}}^{(k+1)} = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/4 \end{bmatrix} \overline{\mathbf{x}}^{(k)} + \begin{bmatrix} 1 \\ 1/2 \end{bmatrix}.$$

Így $\overline{\mathbf{x}}^{(1)} = [1, 3/2]^T$ és $\|\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(0)}\| = \sqrt{13}/2$. Továbbá $\varrho(\mathbf{B}_{GS}^T \mathbf{B}_{GS}) = 5/16$, azaz $\|\mathbf{B}_{GS}\|_2 = \sqrt{5}/4$ és a hibabecslő formulából (1.4. tétel)

$$\|\overline{\mathbf{x}}^{(k)} - \overline{\mathbf{x}}^{\star}\| \leq \frac{(\sqrt{5}/4)^k}{1 - (\sqrt{5}/4)} \frac{\sqrt{13}}{2} \leq 10^{-6},$$

azaz $k \geq 26.17$ kell legyen.

3.38. A Jacobi-módszer iterációs mátrixa

$$\mathbf{B}_J = \begin{bmatrix} 0 & -2 & -2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{bmatrix}.$$

Ennek sajátértékei 2 és $-1 \pm \sqrt{5}$, azaz a módszer nem lesz konvergens (tetszőleges kezdővektorra).

A Gauss–Seidel-módszer iterációs mátrixa

$$\mathbf{B}_{GS} = \begin{bmatrix} 0 & -2 & -2 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}.$$

Ennek sajátértékei 0 és 2, azaz a módszer nem lesz konvergens (tetszőleges kezdővektorra).

Tehát egyik módszer sem lesz konvergens.

3.39. A Jacobi-módszer iterációs mátrixa

$$\mathbf{B}_J = \begin{bmatrix} 0 & -1/2 & -1 \\ -1/2 & 0 & -1 \\ 2 & -2 & 0 \end{bmatrix}.$$

Ennek spektrálsugara $1/2$, azaz a módszer konvergens.

A Gauss–Seidel-módszer iterációs mátrixa

$$\mathbf{B}_{GS} = \begin{bmatrix} 0 & -1/2 & -1 \\ 0 & 1/4 & -1/2 \\ 0 & -3/2 & -1 \end{bmatrix}.$$

Ennek spektrálsugara $(3+\sqrt{73})/8 \approx 1.443$, azaz a módszer nem lesz konvergens.

Tehát a Gauss–Seidel-módszer nem, míg a Jacobi-módszer konvergens lesz az adott együtthatómátrixú egyenletrendszerre.

3.40. A Gauss–Seidel-módszer esetén az iterációs mátrix a

$$\mathbf{B}_{GS} = \begin{bmatrix} 0 & 10/8 \\ 0 & -5/8 \end{bmatrix}$$

mátrix lesz. Ennek spektrálsugara 5/8, ami kisebb 1-nél, így a módszer valóban konvergens lesz. Az első lépés eredménye $[-1/4, 13/8]^T$.

3.41. Az, hogy az adott x_k értékek megoldások, egyszerű behelyettesítéssel igazolható:

$$\frac{3}{4}x_{k-1} + \frac{1}{4}x_{k+1} = \frac{3}{4}\left(1 - \frac{3^{k-1}-1}{3^{20}-1}\right) + \frac{1}{4}\left(1 - \frac{3^{k+1}-1}{3^{20}-1}\right)$$

$$= 1 - \frac{1}{4}\frac{3^k-3}{3^{20}-1} - \frac{1}{4}\frac{3^{k+1}-1}{3^{20}-1} = 1 - \frac{3^k - 3 + 3\cdot 3^k - 1}{4(3^{20}-1)} = 1 - \frac{1-3k-1}{3^20-1} = x_k,$$

a $k=0$ és $k=20$ eset egyszerű behelyettesítéssel adódik. Különböző ω relaxálási paraméterekre futtatva a SOR módszert, az alábbi táblázat mutatja, hogy hány iterációra van szükség a 10^{-10}-es hiba (2-es normában) eléréséhez. Ahogy látható, az alulrelaxálás nem javít a konvergencia sebességén, de a túlrelaxálás igen. Kb. 1.3 és 1.35 között van valahol az optimális ω érték.

omega	iterációszám
0.80	129
0.85	114
0.90	101
0.95	89
1.00	79
1.05	69
1.10	60
1.15	51
1.20	43
1.25	34
1.30	25
1.35	25
1.40	30
1.45	35
1.50	42

```
1.55    51
1.60    61
1.65    74
1.70    91
1.75    116
1.80    154
1.85    217
1.90    343
1.95    751
```

3.42. Az adott egyenletrendszerre alkalmazva a fenti képletet, azt kapjuk, hogy

$$\overline{\mathbf{x}}^{(k+1)} = \begin{bmatrix} 1-\omega & -\omega/4 \\ -2\omega/3 & 1-\omega \end{bmatrix} \overline{\mathbf{x}}^{(k)} + \begin{bmatrix} \omega/4 \\ 2\omega/3 \end{bmatrix}.$$

Az iterációs mátrix sajátértékei: $1-\omega+\omega/\sqrt{6}$, $1-\omega-\omega/\sqrt{6}$, így a spektrálsugár akkor a legkisebb, ha $\omega = 1$, azaz a Jacobi-módszerről van szó. Ezzel az iteráció

$$\overline{\mathbf{x}}^{(k+1)} = \begin{bmatrix} 0 & -1/4 \\ -2/3 & 0 \end{bmatrix} \overline{\mathbf{x}}^{(k)} + \begin{bmatrix} 1/4 \\ 2/3 \end{bmatrix},$$

tehát az iterációs mátrix maximumnormája $2/3$ és $\overline{\mathbf{x}}^{(1)} = [1/4, 2/3]^T$. A

$$\frac{(2/3)^k}{1-(2/3)}\frac{2}{3} \leq 10^{-6}$$

feltételt kell garantálni, ami $k \geq 35.78$ esetén teljesül, azaz a 36. iterációs lépés már 10^{-6}-nál jobban megközelíti a sorozat maximumnormában a megoldást.

3.43. Cseréljük ki az egyenletrendszer első két sorát, mert akkor diagonálisan szigorúan domináns együtthatómátrixot kapunk, amire pl. a Jacobi-módszer konvergálni fog. A Jacobi-módszert felírva az egyenletrendszerre $\overline{\mathbf{x}}^{(1)} = [1/2, 1/5, 3/7]^T$ adódik, és az iterációs mátrix maximumnormája 0.6 lesz. A 3.1. tétel szerint azt kapjuk, hogy legalább 28-at kell lépnünk az iterációval az adott hiba eléréséhez.

3.44. Az ω paraméter megválasztásával azt kell garantálnunk, hogy az iterációs mátrix spektrálsugara a lehető legkisebb 1-nél kisebb szám legyen. Az iterációs mátrix λ sajátértékei a feltételek szerint az $[1-\omega\beta, 1-\omega\alpha]$ intervallumba esnek. Mivel nem tudjuk, hogy mik lesznek a sajátértékek, válasszuk ω-t úgy, hogy

$$\min_{\lambda \in [1-\omega\beta, 1-\omega\alpha]} |\lambda|$$

minimális legyen. Az $\omega \mapsto |1-\omega\beta|$ és $\omega \mapsto |1-\omega\alpha|$ függvények grafikonjait ábrázolva láthatjuk, hogy a megfelelő kifejezés akkor lesz minimális, amikor a két függvény metszi egymást (az $\omega = 1$ pont kivételével). Ez akkor van, ha $\omega = 2/(\alpha+\beta)$. Ez lesz a feltételekből következő legjobb ω választás.

3.45. Az iterációs mátrix $\mathbf{B} = \mathbf{E} + \alpha\mathbf{A}$. Ennek a spektrálsugarát kell minimalizálni. Az $\mathbf{A}$ mátrix sajátértékei 1 és 4, így $\mathbf{B}$ sajátértékei $1+\alpha$ ill. $1+4\alpha$. Az α paraméter értéke akkor optimális, ha $|1+\alpha| = |1+4\alpha|$ és $\alpha \neq 0$. Ebből az optimális α-ra -0.4 adódik.

Variációs módszerek

3.46. Az együtthatómátrix transzponáltjával balról szorozva az egyenletet kapjuk a normálegyenletet:

$$\begin{bmatrix} 20 & 10 \\ 10 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 7 \end{bmatrix}.$$

Erre alkalmazva a gradiens módszert kapjuk, hogy

$$\overline{\mathbf{x}}^{(1)} = \begin{bmatrix} 452/1445 \\ 791/2890 \end{bmatrix}.$$

3.47. $\overline{\mathbf{x}}_0 = [0,0]^T$, $\overline{\mathbf{x}}_1 = [1/3,0]^T$, $\overline{\mathbf{x}}_2 = [1/3,-1/12]^T$.

3.48. Kiinduló adatok: $\overline{\mathbf{x}}_0 = \mathbf{0}$, $\overline{\mathbf{r}}_0 = \overline{\mathbf{b}} = [1,0,1]^T$, $\overline{\mathbf{p}}_1 = \overline{\mathbf{b}} = [1,0,1]^T$. Innét: $\alpha_1 = 1/2$, $\overline{\mathbf{x}}_1 = [1/2,0,1/2]^T$, $\overline{\mathbf{r}}_1 = [0,1,0]^T$, $\beta_1' = 1/2$, $\overline{\mathbf{p}}_2 = [1/2,1,1/2]$, $\overline{\mathbf{x}}_2 = [1,1,1]^T$, $\overline{\mathbf{r}}_2 = [0,0,0]^T$. Azaz két lépésben megkaptuk a megoldást.

3.49. Kiinduló adatok: $\overline{\mathbf{x}}_0 = \mathbf{0}$, $\overline{\mathbf{r}}_0 = \overline{\mathbf{b}} = [1,1]^T$, $\overline{\mathbf{p}}_1 = \overline{\mathbf{b}} = [1,1]^T$. Innét: $\alpha_1 = 1/3$, $\overline{\mathbf{x}}_1 = [1/3,1/3]^T$, $\overline{\mathbf{r}}_1 = [0,0]^T$. Ez azt mutatja, hogy az első lépésben megkaptuk már a pontos megoldást.

3.50. A mátrix szimmetrikus, pozitív definit, így alkalmazható rá a módszer. Az alábbi módon számolhatunk: $\overline{\mathbf{x}} = [0,0]^T$, $\overline{\mathbf{r}} = [1,0]^T$, $\overline{\mathbf{p}} = [1,0]^T$, $\alpha = 1/3$, $\overline{\mathbf{x}} = [1/3,0]^T$, $\overline{\mathbf{r}} = [0,-1/3]^T$, $\beta = -1/9$, $\overline{\mathbf{p}} = [1/9,-1/3]^T$, $\alpha = 3/11$, $\overline{\mathbf{x}} = [4/11,-1/11]^T$, ami már az egyenletrendszer megoldását adja.

3.51. Az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenlet megoldását megkaphatjuk úgy is, hogy megoldjuk az $\mathbf{A}(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{y}}$ egyenletet és a megoldáshoz hozzáadunk $\overline{\mathbf{y}}$-t. Így a korábbi programot a $(konj)grad(A, b - Ay, toll, nmax) + y$ módon kell alkalmazni.

3.52. 10 lépésből megkapjuk a megoldást (a nullvektorról indulva):

```
x =
  [10 19 27 34 40 45 49 52 54 55 55 54 52 49 45 40 34 27 19 10]'
iter =
    10
```

3.53. A konjugált gradiens módszerrel 4 lépés kell a pontos megoldáshoz, a gradiens módszerrel pedig 49 lépés után kapjuk meg a megoldást 10^{-10}-es pontossággal. A megoldás:

```
x =
   0.067814671156387
  -0.260943693789694
   0.545032181066451
   0.164869357657389
```

Túlhatározott lineáris egyenletrendszerek megoldása

3.54. Ebben a feladatban több helyes megoldás is lehetséges, hiszen a tükrözésekhez használt $\overline{\mathbf{v}}$ vektor nem egyértelmű, így a QR-felbontás sem lesz egyértelmű. Most a $\overline{\mathbf{v}} = \overline{\mathbf{x}} + \|\overline{\mathbf{x}}\|_2 \overline{\mathbf{e}}_1$ képlettel fogunk számolni. Először a

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

vektorral meghatározzuk a tükörsík normálvektorát

$$\overline{\mathbf{v}}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 1 \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix},$$

majd ezzel az első Householder-tükrözést:

$$\mathbf{H}_1 = \mathbf{E} - 2\frac{\overline{\mathbf{v}}_1 \overline{\mathbf{v}}_1^T}{\overline{\mathbf{v}}_1^T \overline{\mathbf{v}}_1} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

$$\mathbf{H}_1 \mathbf{A} = \begin{bmatrix} -1 & -3 \\ 0 & 0 \\ 0 & 2 \end{bmatrix}.$$

Ez a mátrix még nem felső háromszög, így még egy tükrözésre lesz szükség. Most a $[0, 2]^T$ vektorhoz kell egy tükrözést keresnünk.

$$\overline{\mathbf{v}}_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix} + 2 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix},$$

mellyel a tükrözés

$$\tilde{\mathbf{H}}_2 = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

és

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix},$$

valamint

$$\mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{bmatrix} -1 & -3 \\ 0 & -2 \\ 0 & 0 \end{bmatrix} = \mathbf{R}.$$

A $\mathbf{Q}$ mátrix a

$$\mathbf{Q} = \mathbf{H}_1\mathbf{H}_2 = \begin{bmatrix} 0 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}$$

képlettel adódik.

A túlhatározott egyenletrendszer megoldása az $\overline{\mathbf{x}}_{LS}$ megoldás megkeresését jelenti. Ez a QR-felbontással úgy határozható meg, hogy együtthatómátrixnak az $\mathbf{R}(1:2,1:2)$ mátrixot vesszük (az $\mathbf{R}$ mátrix felső, négyzetes almátrixa), jobb oldalnak pedig a $\mathbf{Q}^T[1,1,1]^T$ mátrix első két eleméből álló oszlopvektort. Így a

$$\begin{bmatrix} -1 & -3 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

egyenletrendszerhez jutunk, melynek megoldása $x_1 = -1/2$, $x_2 = 1/2$.

3.55. A normálegyenletet úgy kapjuk, hogy balról szorozzuk az egyenletrendszert az $\mathbf{A}$ mátrix transzponáltjával.

$$\begin{bmatrix} 1 & 3 \\ 3 & 13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}.$$

Ezen egyenletrendszer megoldása $\overline{\mathbf{x}}_{LS} = [-1/2, 1/2]$. (Vö. 3.54. feladat.)

3.56. A két lehetséges módszer közül a normálegyenlet alkalmazása a könnyebben végrehajtható megoldási mód. Ezt használva $\overline{\mathbf{x}}_{LS} = [-1/6, 1/3, 1/2]^T$ adódik.

3.57. Az egyenlet normálegyenlete

$$\begin{bmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 32 \\ 138 \end{bmatrix},$$

aminek megoldása $\overline{\mathbf{a}}_{LS} = [1/5, -2/35, 1/7]$. Ez a vektor annak a legfeljebb másodfokú polinomnak adja meg az együtthatóit ($x^2/7 - 2x/35 + 1/5$), amely az $(1,0), (2,2), (3,-1), (4,4)$ és $(5,3)$ pontokhoz legkisebb négyzetek értelemben a legközelebb halad.

3.58. A normálegyenlet felírásával és megoldásával megkaphatjuk, hogy

$$x = \frac{a+c+d}{3}, \quad y = \frac{b-c+d}{3}.$$

3.59. A pontos megoldás:

$$x = \frac{1+2\cdot 10^k+2\cdot 100^k}{2\cdot 100^k+1}, \quad y = \frac{-1+2\cdot 10^k-2\cdot 100^k}{2\cdot 100^k+1}, \quad k = 6,7,8.$$

MATLAB-ban számolva rendre az alábbi 2-es normában számolt hibákat nyerjük. Az eredmények azt mutatják, hogy a QR-felbontással nyert megoldás sokkal pontosabb. A $k = 8$ esetben a Cholesky-felbontásos megoldás nem ad használható értéket, mert az együtthatómátrix a MATLAB pontosságán belül már szinguláris.

```
k=6
hibaQR =
    1.049175818250489e-010
hibaCholesky
    1.257132582260048e-004

k=7
hibaQR =
    9.724246745738770e-010
hibaCholesky
    0.017034004439712

k=8
hibaQR =
    1.961820871056758e-009
hibaCholesky
    NaN
```

Sajátértékfeladatok numerikus megoldása

Sajátértékbecslések

4.1. Alkalmazzuk a Gersgorin-tételt! Eszerint a sajátértékek a $K_{0.3}(1)$, $K_{0.4}(3)$ és a $K_{0.2}(-2)$ körök uniójában vannak, ahol most $K_\varepsilon(x)$ az x közepű, ε-sugarú, zárt körlapot jelenti a komplex számsíkon. Mivel ezek a körök diszjunktak, így a második Gersgorin-tétel szerint mindegyik körben pontosan egy sajátérték van, ami azt jelenti, hogy mindhárom sajátérték valós. A korábbi becslésünket javíthatjuk úgy, hogy észrevesszük, hogy $\mathbf{A}^T$ sajátértékei megegyeznek $\mathbf{A}$ sajátértékeivel, így a transzponáltra alkalmazhatjuk a Gersgorin-tételt: a sajátértékek a $K_{0.3}(1)$, $K_{0.3}(3)$ és a $K_{0.2}(-2)$ körök uniójában vannak. A korábbi becslésekkel ezeket összevetve kapjuk az alábbi sajátértékbecsléseket: 1 ± 0.3, 3 ± 0.3, -2 ± 0.2. (A tényleges sajátértékek rendre: 0.967332067785579, 3.031395311434764, -1.998727379220341.)

4.2. A feladatban kapott becslés függ a becslés módszerétől, így különböző normákban, vagy más sajátvektorokat választva, más-más becslés nyerhető. Két lehetséges megoldást mutatunk.

Számítsuk ki a harmadik sajátértékét a mátrixnak és a hozzá tartozó egyik sajátvektort! A karakterisztikus egyenlet $(1-\lambda)\lambda(1+\lambda) = 0$. Így a harmadik sajátérték -1. Mivel minden sajátérték különböző, a mátrix diagonalizálható. A -1-hez tartozó sajátvektor pl. $[1, -1, 0]^T$. Így a diagonalizáló mátrix

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 0 \\ -1 & -2 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

lesz.

1. megoldás: Alkalmazzuk a Bauer–Fike-tételt (4.2. tétel)! Ehhez $\mathbf{S}$ kondíciószámára van szükségünk (mondjuk maximumnormában, mert ezt könnyű számolni). Ehhez $\mathbf{S}$

inverzét kell meghatároznunk:

$$\mathbf{S}^{-1} = \begin{bmatrix} -2 & 1 & 0 \\ -1 & 1 & 0 \\ 2 & -2 & -1 \end{bmatrix}.$$

Így maximumnormában a $\kappa_\infty(\mathbf{S}) = 25$ értéket kapjuk, hiszen $\mathbf{S}$-nek és $\mathbf{S}^{-1}$-nek is 5 a maximumnormája. Mivel $\mathbf{B}$ maximumnormája 3, így a

$$|\lambda_j(0) - \lambda_j(\varepsilon)| \leq 25 \cdot 3 \cdot \varepsilon = 75\varepsilon$$

becslést kapjuk.

2. megoldás: Az

$$\mathbf{S}^{-1}(\mathbf{A} + \varepsilon\mathbf{B})\mathbf{S} = \mathbf{D} + \varepsilon\mathbf{S}^{-1}\mathbf{B}\mathbf{S} = \mathbf{D} + \varepsilon \begin{bmatrix} 2 & 3 & -1 \\ -4 & -6 & 2 \\ -2 & -3 & 1 \end{bmatrix} = \begin{bmatrix} -1+2\varepsilon & 3\varepsilon & -\varepsilon \\ -4\varepsilon & -6\varepsilon & 2\varepsilon \\ -2\varepsilon & -3\varepsilon & 1+\varepsilon \end{bmatrix}$$

egyenlőségből, ahol $\mathbf{D} = \operatorname{diag}(-1, 0, 1)$ a sajátértékeket tartalmazó diagonális mátrix, a Gersgorin-tételeket felhasználva nyerjük, hogy

$$-1 - 2\varepsilon \leq \lambda_1(\varepsilon) \leq -1 + 6\varepsilon,$$

$$-12\varepsilon \leq \lambda_2(\varepsilon) \leq 0,$$

$$1 - 4\varepsilon \leq \lambda_3(\varepsilon) \leq 1 + 6\varepsilon.$$

Ez a Bauer–Fike-tételnél jobb becslést ad.

4.3. A Gersgorin-körök a következők: $K_1(2)$, $K_1(-2), K_1(3), K_4(5)$, ahol $K_\varepsilon(x)$ jelenti az x körüli ε sugarú zárt körlapot a komplex számsíkon. Mivel $K_1(-2)$ diszjunkt a többi körlaptól, Gersgorin második tétele miatt ebben a körben pontosan egy sajátérték van, ami nem lehet nemnulla képzetes részű, hiszen akkor a sajátérték konjugáltja is a körbe esne. A körön belül a $[-3, -1]$ negatív valós számok vannak. A többi kör a komplex sík pozitív valós részű oldalára esik. Így biztosan pontosan egy negatív valós sajátérték van.

4.4. Most a Gersgorin-tétel csak annyit mond, hogy a sajátértékek a $K_2(3)$ körben vannak, ebből még nem következik a bizonyítandó állítás.

Rendezzük át a sorokat a 2.,4.,1.,3. sorrendre és az oszlopokat is ugyanígy. Ez egy hasonlósági transzformáció egy permutációs mátrixszal, ami közben a sajátértékek nem változnak meg. Így kapjuk az alábbi mátrixot:

$$\begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & 1 & 2 \end{bmatrix}.$$

Ennek a mátrixnak a sajátértékeit a bal felső 2×2-es ill. a jobb alsó 2×2-es mátrixok sajátértékei adják. A bal felső mátrix szimmetrikus, így minden sajátértéke valós, a bal alsónál pedig a sajátértékek 1 és 4, ami könnyen látható onnét, hogy a sajátértékek szorzatának 4-nek (det), az összegének pedig 5-nek (trace) kell lennie. Így ezek a sajátértékek is valósak.

4.5. Az $\mathbf{A}$ mátrix szimmetrikus, így az őt diagonalizáló mátrix ortogonális, melynek 2-es normája 1. Emiatt érdemes a becslést 2-es normában elvégezni. Ekkor a 4.2. tétel szerint a sajátértékek nem változhatnak nagyobbat, mint a perturbáló mátrix 2-es normája, ami $\sqrt{3}/10 \approx 0.1732$. (Valóban így van, hiszen a mátrix sajátértékei -2.109772228646443, 2.000000000000000, 7.109772228646442, míg a perturbált mátrixé -2.108311239790695, 2.018761474726724, 7.189549765063973.)

A $\mathbf{B}$ mátrix nem szimmetrikus, így először meg kell határoznunk a diagonalizáló mátrixát. A mátrix sajátértékei 1, 2 és 3, a hozzájuk tartozó sajátvektorok rendre $[1, -4, 1]^T$, $[-1, 1, 0]^T$ és $[1, 0, 0]^T$, így a

$$\mathbf{V} = \begin{bmatrix} 1 & -1 & 1 \\ -4 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

mátrix diagonalizálni fogja a $\mathbf{B}$ mátrixot. Számoljunk most maximumnormában, mert azt egyszerű meghatározni. $\|\mathbf{V}\|_\infty = 5$,

$$\mathbf{V}^{-1} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 4 \\ 1 & 1 & 3 \end{bmatrix}$$

és $\|\mathbf{V}^{-1}\|_\infty = 5$, azaz $\kappa_\infty(\mathbf{V}) = 25$. Mivel a perturbáló mátrix maximumnormája 0.1, ezért a sajátértékek maximális változása $25 \cdot 0.1 = 2.5$. Természetesen más normában vagy más sajátvektorokat megadva ennél jobb felső becslés is adható a maximális változásra. (A perturbált mátrix sajátértékei 1.5752, 1.1462, 3.3786).

4.6. A keresett szorzó az $\mathbf{A}$ mátrix $\overline{\mathbf{x}}$ vektorhoz tartozó Rayleigh-hányadosa:

$$\alpha = \frac{\overline{\mathbf{x}}^T \mathbf{A} \overline{\mathbf{x}}}{\overline{\mathbf{x}}^T \overline{\mathbf{x}}} = \frac{20}{6}.$$

4.7. Sajátvektorból a Rayleigh-hányadossal tudunk jó sajátértékközelítést mondani:

$$\lambda \approx 6.7113.$$

Hatványmódszer és változatai

4.8. Először számítsuk ki az $\mathbf{A}^4\overline{\mathbf{x}}^{(0)}$ vektort:

$$\mathbf{A}^4\overline{\mathbf{x}}^{(0)} = \begin{bmatrix} 103 \\ 102 \end{bmatrix}.$$

Ez lesz a sajátvektor egy közelítése, vagy pl. 2-es normában normálva $[0.7105, 0.7036]^T$. A sajátérték közelítését a Rayleigh-hányadossal számítjuk: 3.995.

4.9. A sajátértékek valósak, Gersgorin tétele miatt van egy -1, 5 és 10 közelében (1, 1 ill. 2 sugarú környezetekben). A $\mathbf{C}-10\mathbf{E}$ mátrixszal a hatványmódszer az abszolút értékben domináns sajátértéket és a hozzá tartozó sajátvektort határozza meg. A $\mathbf{C}-10\mathbf{E}$ mátrix sajátértékei $\mathbf{C}$ sajátértékeinél 10-zel kisebbek, így lesz egy -11, egy -5 és egy 0 közelében. Így a -11 körüli sajátérték lesz domináns abszolút értékű, az ehhez tartozó sajátvektor a $\mathbf{C}$ mátrix λ_1 sajátértékéhez tartozó sajátvektor.

Egy lépést végrehajtva az

$$\overline{\mathbf{x}}^{(1)} = \begin{bmatrix} -\frac{20}{3} \\ -1 \\ 1 \end{bmatrix}$$

vektorhoz jutunk. Ezzel a Rayleigh-hányados -10.9641, azaz -10.9641+10=-0.9641 egy becslést ad a λ_1 sajátértékre.

4.10.

$$\overline{\mathbf{x}}^{(1)} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \ \overline{\mathbf{x}}^{(2)} = \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix}, \ \overline{\mathbf{x}}^{(3)} = \begin{bmatrix} 6 \\ -8 \\ 6 \end{bmatrix}, \ \overline{\mathbf{x}}^{(4)} = \begin{bmatrix} 20 \\ -28 \\ 20 \end{bmatrix}.$$

Az $\overline{\mathbf{x}}^{(4)}$ vektorral kiszámolva a Rayleigh-hányadost a sajátérték becslése 3.4141. (A pontos érték $2+\sqrt{2}$.)

4.11. A Gersgorin-tétel miatt az $\mathbf{A}$ mátrix sajátértékei a $[0, 4]$ intervallumból kerülhetnek csak ki. Az $\mathbf{A}-4\mathbf{E}$ sajátértékei $\mathbf{A}$ sajátértékeinél 4-gyel kisebbek, így ezek a $[-4, 0]$ intervallumban vannak. Tehát $\mathbf{A}-4\mathbf{E}$ legnagyobb abszolút értékű sajátértéke 4-gyel kisebb, mint $\lambda_{\min}$. Ezzel az állítást igazoltuk.

Az $(\overline{\mathbf{x}}^{(0)})^T = [1, 1, 1, 1]^T$ vektort háromszor kell a hatványmódszer szerint $\mathbf{A}$-val szoroznunk. ($\overline{\mathbf{x}}^{(3)}$-at az $\mathbf{A}(\mathbf{A}(\mathbf{A}\overline{\mathbf{x}}^{(0)}))$ módon és nem az $(\mathbf{A}^3)\overline{\mathbf{x}}^{(0)}$ módon érdemes számítani, továbbá az iterációs vektorok első és második eleme a szimmetria miatt ugyanaz lesz, mint az utolsó és az utolsó előtti elem.) Innét kapjuk, hogy $\overline{\mathbf{x}}^{(1)} = [-3, -4, -4, -3]^T$, $\overline{\mathbf{x}}^{(2)} = [10, 15, 15, 10]^T$, $\overline{\mathbf{x}}^{(3)} = [-35, -55, -55, -35]^T$. Ebből a Rayleigh-hányadossal kaphatunk becslést a domináns sajátértékre. Erre -3.6176 adódik, azaz $\lambda_{\min} \approx -3.6176+4 = 0.3824$.

4.12. Az α értéket úgy kell meghatározni, hogy a $20-\alpha$ körüli érték legyen az $\mathbf{A}-\alpha\mathbf{E}$ mátrix domináns sajátértéke. Az kell tehát, hogy $|20-\alpha|$ ne legyen kisebb $|10-\alpha|$, $|5-\alpha|$ és $|1-\alpha|$ egyikénél sem. Ezen függvények (α változójú) egyszerű ábrázolásából látszik, hogy ez $\alpha \geq 10.5$ esetén már teljesül. A konvergencia akkor a leggyorsabb, ha a második és első domináns sajátértékek aránya a legkisebb. Ez akkor teljesül, ha $\alpha = 12.5$.

4.13. Használjuk a powmeth.m programot a feladat megoldására! A legkisebb sajátértéket úgy kaphatjuk meg, ha a mátrix inverzére alkalmazzuk a hatványmódszert (a mátrix szimmetrikus, pozitív definit, így minden sajátértéke pozitív), majd a kapott sajátértéknek vesszük a reciprokát!

```
% A sajátvektora és a legnagyobb sajátérték.
>> [v,s,iter]=powmeth(toeplitz([2,-1,0]),100,10^-6)
v =
   0.499671783135477
  -0.707106705035206
   0.500328109176836
s =
   3.414213257777039
iter =
     16
% A sajátvektora és a legkisebb sajátérték.
>>[v,s]=powmeth(inv(toeplitz([2,-1,0])),100,10^-6)
   v =
      0.499817259171219
      0.707103662253327
      0.500187083262356
   s =
      1.707106698611546
   iter =
        6
>> s=1/s
   s =
      0.585786465962167
```

4.14. Az inverz iteráció megvalósítható pl. az alábbi módon. Az `est` bemenő paraméter a keresett sajátértékre vonatkozó becslés.

```
function [y,nu,iter]=invpowmeth(A,est,nmax,toll);
[n,n]=size(A);
y=rand(n,1);
```

```
y=y/norm(y);
[L,U]=lu(A-est*eye(n));
nuold=y'*A*y;
  y=L\y;
  y=U\y;
   y=y/norm(y);
   nu=y'*A*y;
   err=abs(nu-nuold);
iter=1;
while err>toll && iter<nmax
   iter=iter+1;
   y=L\y;
   y=U\y;
   y=y/norm(y);
   nuold=nu;
   nu=y'*A*y;
   err=abs(nu-nuold);
end;
```

4.15. A 4.14. feladatban megkonstruált programot használjuk.

```
% A keresett sajátvektor és sajátérték!
>> [v,s,iter]=invpowmeth(hilb(6),1/4,100,10^-6)
v =
   0.614533738941380
  -0.211052124913086
  -0.365884211463252
  -0.394690747896188
  -0.388215561286880
  -0.370731600452637
s =
   0.242360869819382
iter=
   3
```

4.16. A legnagyobb abszolút értékű sajátérték a hatványmódszerrel a másik kettő az inverz iterációval határozható meg úgy, hogy a sajátértékbecsléseket rendre 0-nak ill. 15-nek választjuk. Az eredményeknél csak a sajátértékeket adjuk meg az alábbiakban.

```
% Legnagyobb abszolút érték:
>> A=toeplitz([10:-1:1])-5*eye(10);
```

```
>> [v,s,iter]=powmeth(A,100,10^-6)
s =
  62.840398619704381
% Legkisebb abszolút érték:
>> [v,s,iter]=invpowmeth(A,0,100,10^-6)
s =
  -0.544007837288017
% A 15 közeli:
>> [v,s,iter]=invpowmeth(A,15,100,10^-6)
s =
  15.431729094507599
```

4.17. A mátrix pozitív definit, így minden sajátérték pozitív. Először meghatározzuk a legnagyobb sajátértéket (s) és a hozzá tartozó sajátvektort ($\overline{\mathbf{v}}$), majd megkonstruáljuk az $\mathbf{A}_1 = \mathbf{A} - s\overline{\mathbf{v}}\overline{\mathbf{v}}^T$ mátrixot, amire szintén alkalmazzuk a hatványmódszert. Ennek domináns sajátértéke már a kívánt sajátértéket adja.

```
>> A=hilb(5);
>> [v1,s1,iter]=powmeth(A,100,10^-6)
v1 = % sajátvektor
   0.767827161255247
   0.445803700165107
   0.321597758639363
   0.253459279120881
   0.209842290359855
s1 =   % a legnagyobb sajátérték
   1.567050688247044
>> A1=A-s1*v1*v1';
>> [v2,s2,iter]=powmeth(A1,100,10^-6)
v2 =
  -0.602118950814364
   0.275703037787104
   0.424756424821781
   0.443840086261477
   0.428985502581610
s2 =  % a második legnagyobb sajátérték
   0.208534238819530
```

4.18. Az alábbi módon számolhatunk a MATLAB-ban:

```
>> [v1,s1,iter]=powmeth(A,100,10^-6)
```

```
v1 = % sajátvektor
   0.767827161255247
   0.445803700165107
   0.321597758639363
   0.253459279120881
   0.209842290359855
s1 = % legnagyobb sajátérték
   1.567050688247044
>> v=v1+norm(v1)*[1;0;0;0;0]; % tükrözési vektor
>> H=eye(5)-2*(v*v')/(v'*v); % Householder-tükrözés
>> A1=H*A*H;
>> A2=A1(2:5,2:5);
>> [v2temp,s2,iter]=powmeth(A2,100,10^-6)
v2temp =
   0.427631921230542
   0.534349185556795
   0.530205824197741
   0.500483438113850
s =
   0.208534220726836
>> [v,s,iter]=invpowmeth(A,s,100,10^-6)
v = % ez a sajátvektor
  -0.601871478353973
   0.275913417432098
   0.424876622351521
   0.443903038699774
   0.429013353681693
s = % ez pedig a második legnagyobb sajátérték
   0.208534218611013
```

Jacobi- és QR-iterációk

4.19. Ha $a = d$, akkor a $\cos^2\theta = c^2 - s^2 = 0$ egyenlőségből és a Pitagorasz-tételből ($s^2 + c^2 = 1$) kapjuk, hogy az $s^2 = c^2 = 1/2$ választás megfelelő.

Ha $a \neq d$, akkor az $s = 0$ (ekkor $c = \pm 1$) vagy a $c = 0$ (ekkor $s = \pm 1$) értékek nyilvánvalóan nem adnak megfelelő forgatást, így feltehetjük, hogy sem s, sem c nem nulla. Ekkor

$$\operatorname{tg}^2\theta + 2\operatorname{ctg}(2\theta)\cdot\operatorname{tg}\theta - 1 = \frac{s^2}{c^2} + 2\frac{c^2-s^2}{2sc}\frac{s}{c} - 1 = \frac{s^2}{c^2} + \frac{c^2-s^2}{c^2} - 1 = \frac{c^2}{c^2} - 1 = 0.$$

Ezek alapján tehát s/c hányadosra kaptunk egy másodfokú egyenletet:

$$x^2 + \frac{d-a}{b}x - 1 = 0,$$

majd a Pitagorasz-tételből az kapjuk, hogy az egyenlet x megoldásaival az $s^2 = x^2/(x^2+1)$ és $c^2 = 1/(x^2+1)$ értékek megfelelők lesznek a transzformációhoz.

4.20. A 4.19. feladat eredményét használjuk fel. Mivel az $\mathbf{A}$ mátrix esetén $a = d$, így az $s = c = 1/\sqrt{2}$ választás megfelelő lesz. Valóban

$$\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \cdot \begin{bmatrix} 2 & 4 \\ 4 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & 6 \end{bmatrix}.$$

Tehát $\mathbf{A}$ sajátértékei -2 és 6. A $\mathbf{B}$ mátrix esetén $a = 1, b = -2, d = 4$, azaz $a \neq d$. Így először megoldjuk az $x^2 + (4-1)x/(-2) - 1 = x^2 - 1.5x - 1 = 0$ egyenletet: $x_{1,2} = 2$ és $-1/2$. Válasszuk mondjuk az $x = 2$ értéket, és ezzel határozzuk meg az s és c értékeket. Az

$$s = \frac{2}{\sqrt{5}}, \quad c = \frac{1}{\sqrt{5}}$$

választások megfelelőek lesznek. Valóban, hiszen ezekkel az értékekkel:

$$\begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \cdot \begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix} \cdot \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix}.$$

Ez mutatja, hogy a $\mathbf{B}$ mátrix sajátértékei 0 és 5.

4.21. A Jacobi-módszert a 4.5. tétel alapján hajtjuk végre. Az első sor harmadik eleméhez tartozó $\mathbf{S}_{13}$ Jacobi-transzformációs mátrix

$$\mathbf{S}_{13} = \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix},$$

amivel

$$\mathbf{A}^{(1)} := \mathbf{S}_{13}^T \mathbf{A} \mathbf{S}_{13} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 2\sqrt{2} \\ 0 & 2\sqrt{2} & 5 \end{bmatrix}.$$

Ezek után a második sor harmadik eleméhez készítjük el a transzformációs mátrixot. Az $\mathbf{S}_{23}$ mátrix

$$\mathbf{S}_{23} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{6}/3 & 1/\sqrt{3} \\ 0 & -1/\sqrt{3} & \sqrt{6}/3 \end{bmatrix},$$

amivel

$$\mathbf{A}^{(2)} := \mathbf{S}_{23}^T \mathbf{A}^{(1)} \mathbf{S}_{23} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 7 \end{bmatrix}.$$

Itt azt vesszük észre, hogy a létrejött iterációs mátrix már egy diagonális mátrix, aminek a hasonlósági transzformációk miatt ugyanazok a sajátértékei, mint az eredeti mátrixnak. Így – most kivételesen – a Jacobi-módszer az $\mathbf{A}$ mátrix pontos sajátértékeit adja: 1,1 és 7.

4.22. A Jacobi-módszert a 4.5. tétel alapján hajtjuk végre. Az első sor negyedik eleméhez tartozó $\mathbf{S}_{14}$ Jacobi-transzformációs mátrix

$$\mathbf{S}_{14} = \begin{bmatrix} 1/\sqrt{2} & 0 & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1/\sqrt{2} & 0 & 0 & 1/\sqrt{2} \end{bmatrix},$$

amivel

$$\mathbf{A}^{(1)} := \mathbf{S}_{14}^T \mathbf{A} \mathbf{S}_{14} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 2 & 2\sqrt{2} \\ 0 & 2 & 3 & 2\sqrt{2} \\ 0 & 2\sqrt{2} & 2\sqrt{2} & 5 \end{bmatrix}.$$

Ezek után a második sor negyedik eleméhez készítjük el a transzformációs mátrixot. Az $\mathbf{S}_{24}$ mátrix

$$\mathbf{S}_{24} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \sqrt{2}/\sqrt{3} & 0 & 1/\sqrt{3} \\ 0 & 0 & 1 & 0 \\ 0 & -1/\sqrt{3} & 0 & \sqrt{2}/\sqrt{3} \end{bmatrix},$$

amivel

$$\mathbf{A}^{(2)} := \mathbf{S}_{24}^T \mathbf{A}^{(1)} \mathbf{S}_{24} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 2\sqrt{3} \\ 0 & 0 & 2\sqrt{3} & 7 \end{bmatrix}.$$

Ezek után a Gersgorin-tételt alkalmazva kapjuk, hogy az 1 kétszeres sajátértéke a mátrixnak, továbbá van két sajátérték a $[3-2\sqrt{3}, 3+2\sqrt{3}]$ és $[7-2\sqrt{3}, 7+2\sqrt{3}]$ intervallumok uniójában, azaz a $[3-2\sqrt{3}, 7+2\sqrt{3}] \approx= [-0.4641, 10.4641]$ intervallumban. (Ez valóban így van, mert a mátrix pontos sajátértékei: 1,1,1,9.)

4.23. A feladat megoldásához használhatjuk pl. a [10] könyvbeli algoritmust, ahol bemenő paraméterként csak a kiinduló mátrixot ill. a toleranciaszintet kell megadnunk, ami a jelen esetben 1/1000.

A 33. lépés utáni iterációs mátrix már megfelel a feltételnek (négy tizedes jegyre kerekítve):

$$\mathbf{A}^{(33)} = \begin{bmatrix} 3.7321 & -0.0000 & -0.0001 & -0.0000 & 0.0000 \\ -0.0000 & 0.2679 & 0.0000 & -0.0000 & -0.0000 \\ -0.0001 & 0.0000 & 1.0000 & 0.0000 & -0.0000 \\ -0.0000 & -0.0000 & 0.0000 & 3.0000 & -0.0000 \\ 0.0000 & -0.0000 & -0.0000 & 0.0000 & 2.0000 \end{bmatrix}.$$

A Gersgorin-tétel szerint a sajátértékek kb. a főátlóbeli értékek, és az értékek hibája a sor többi elemének abszolút értékben vett összege, azaz a sajátértékek: $3.7321 \pm 1.6627e-004$, $0.2679 \pm 6.0414e-005$, $1 \pm 1.7468e-004$, $3 \pm 1.5299e-006$, $2 \pm 1.2959e-007$.

4.24. A mátrix sajátértékei és a hozzájuk tartozó hibaértékek a Gersgorin-tétel alapján.

```
sajatertekek =
   3.000000000000000
  12.999999999999996
   2.999999999999999
   2.999999999999998
   3.000000000000000
   3.000000000000000
   2.999999999999998
   2.999999999999998
   3.000000000000000
   3.000000000000001
hibaertekek =
   1.0e-014 *
                     0
     0.351331785544957
     0.073729316405455
     0.086224455293544
     0.198982285567787
     0.043644558500404
     0.086775314369678
     0.169353318490030
     0.164133436184475
     0.188356161912350
```

4.25. Egy Givens-forgatás az $\mathbf{A}$ mátrixot már felső háromszögmátrixba transzformálja, így maga a $\mathbf{G}$ mátrix lesz a QR-felbontás $\mathbf{Q}$ mátrixának transzponáltja. Így az első transzformáció alakja $\mathbf{A}^{(1)} = \mathbf{G}\mathbf{A}\mathbf{G}^T$ lesz, ahol $\mathbf{G}$ az első oszlopból számított Givens-forgatási

mátrix

$$\mathbf{G} = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}.$$

Azaz

$$\mathbf{A}^{(1)} = \frac{1}{5} \begin{bmatrix} 19 & -3 \\ -8 & -4 \end{bmatrix}.$$

A következő lépés ugyanilyen, csak most az $\mathbf{A}^{(1)}$ mátrixszal hajtjuk végre $\mathbf{A}$ helyett. Így kapjuk, hogy (tizedestörtekkel kiírva)

$$\mathbf{A}^{(2)} = \begin{bmatrix} 3.8941 & 1.3765 \\ 0.3765 & -0.8941 \end{bmatrix}.$$

A Gersgorin-tételt alkalmazva lehet becslést mondani a sajátértékekre: 3.8941 ± 0.3765 és -0.894 ± 0.3765.

4.26. Alkalmazzunk Givens-forgatást! Az első oszlop elemeiből meghatározhatók a forgatási szög szinusza és koszinusza: $c = 1/\sqrt{2}$, $c = 1/\sqrt{2}$. Így a forgatási mátrix és az azzal való szorzás:

$$\mathbf{GA} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ 0 & \sqrt{2} \end{bmatrix} = \mathbf{R}.$$

Ez a mátrix lesz az $\mathbf{R}$ mátrix, $\mathbf{G}$ transzponáltja pedig $\mathbf{Q}$.

A sajátértékek meghatározására való QR-iteráció első lépéséhez az

$$\mathbf{RQ} = \begin{bmatrix} 0 & 2 \\ -1 & 1 \end{bmatrix}$$

szorzatot kell kiszámolni.

4.27. Egy lehetséges megvalósítás a következő:

```
function [s,h]=qriter(A,nmax,toll)
Ak=A; Dnorm=norm(Ak,'fro'); epsi=toll*Dnorm; iter=1;
while Dnorm > epsi && iter<nmax
    [Q,R]=qr(Ak);
    Ak=R*Q;
    Dnorm=norm(Ak-diag(diag(Ak)),'fro');
    iter=iter+1;
end;
if iter<nmax
    s=diag(Ak)';
    h=sum((abs(Ak-diag(diag(Ak))))');
else
    error('Nem értük el az adott iterációszámmal a kívánt pontosságot.')
end;
```

4.28. Az $\overline{\mathbf{x}} = \mathbf{A}(2:3,1)$ vektorhoz keresünk Householder-tükrözést: $\overline{\mathbf{v}} = [9,3]^T$, ahonnét

$$\tilde{\mathbf{H}} = \frac{1}{5}\begin{bmatrix} -4 & -3 \\ -3 & 4 \end{bmatrix},$$

és a megfelelő Householder-tükrözési mátrix

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -4/5 & -3/5 \\ 0 & -3/5 & 4/5 \end{bmatrix}.$$

Ezzel a mátrixszal a

$$\mathbf{HAH} = \begin{bmatrix} 4 & -13/5 & 9/5 \\ -5 & 32/5 & -11/5 \\ 0 & 4/5 & 8/5 \end{bmatrix}$$

mátrix már felső Hessenberg-alakú lesz, és a hasonlósági transzformáció miatt a sajátértékei megegyeznek $\mathbf{A}$ sajátértékeivel.

4.29. A Householder-tükrözési mátrix ugyanaz lesz, mint a 4.28. feladatban:

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -4/5 & -3/5 \\ 0 & -3/5 & 4/5 \end{bmatrix}.$$

Ezzel a mátrixszal a

$$\mathbf{HAH} = \begin{bmatrix} 4 & -5 & 0 \\ -5 & 196/25 & -28/25 \\ 0 & -28/25 & 4/25 \end{bmatrix}$$

mátrix már felső Hessenberg-alakú lesz, és a hasonlósági transzformáció miatt a sajátértékei megegyeznek $\mathbf{A}$ sajátértékeivel. Mivel $\mathbf{A}$ szimmetrikus mátrix, így a transzformáltja is szimmetrikus, azaz tridiagonális mátrix lesz.

4.30. Az alábbi parancsokkal számolhatunk. Két Householder-transzformációra van szükség. Az A2 mátrix már a kívánt tulajdonságú mátrix lesz.

```
>> A=toeplitz([4,3,2,1])
A =
     4     3     2     1
     3     4     3     2
     2     3     4     3
     1     2     3     4
>> x=A(2:4,1)
x =
```

```
     3
     2
     1
>> v=x+norm(x)*[1,0,0]'
v =
    6.7417
    2.0000
    1.0000
>> H=eye(3)-2*(v*v')/(v'*v)
H =
   -0.8018   -0.5345   -0.2673
   -0.5345    0.8414   -0.0793
   -0.2673   -0.0793    0.9604
>> H1=blkdiag(1,H)
H1 =
    1.0000         0         0         0
         0   -0.8018   -0.5345   -0.2673
         0   -0.5345    0.8414   -0.0793
         0   -0.2673   -0.0793    0.9604
>> A1=H1*A*H1
A1 =
    4.0000   -3.7417   -0.0000   -0.0000
   -3.7417    8.2857   -1.3014   -2.2543
   -0.0000   -1.3014    1.0707    0.9113
   -0.0000   -2.2543    0.9113    2.6436
>> x=A1(3:4,2)
x =
   -1.3014
   -2.2543
>> v=x+norm(x)*[1,0]'
v =
    1.3016
   -2.2543
>> H=eye(2)-2*(v*v')/(v'*v)
H =
    0.5000    0.8660
    0.8660   -0.5000
>> H2=blkdiag(1,1,H)
H2 =
    1.0000         0         0         0
         0    1.0000         0         0
```

```
         0         0    0.5000    0.8660
         0         0    0.8660   -0.5000
>> A2=H2*A1*H2
A2 =
    4.0000   -3.7417   -0.0000   -0.0000
   -3.7417    8.2857   -2.6030   -0.0000
   -0.0000   -2.6030    3.0396   -0.2254
   -0.0000   -0.0000   -0.2254    0.6747
```

4.31. Csupán a program második sorának első két parancsát kell módosítanunk, hiszen a `hess` parancs elvégzi a Hessenberg-alakra hozást:

```
Dnorm=norm(A,'fro'); Ak=hess(A); epsi=toll*Dnorm; iter=1;
```

4.32. Az alábbi módon lehet pl. a programot futtatni:

```
>>[s,h]=qriter(toeplitz([4,3,2,1]),200,10^-6)
s =
   11.0990    3.4142    0.9010    0.5858
h =
  1.0e-005 *
    0.0000    0.0000    0.5698    0.5698
```

4.33. Az alábbi módon lehet pl. a programot futtatni:

```
>> [s,h]=qriter(toeplitz([2,-1,zeros(1,18)]),1000,10^-8)
s =
  Columns 1 through 7
    3.9777    3.9111    3.8019    3.6525    3.4661    3.2470    3.0000
  Columns 8 through 14
    2.7307    2.4450    2.1495    1.8505    1.5550    1.2693    1.0000
  Columns 15 through 20
    0.7530    0.5339    0.3475    0.1981    0.0889    0.0223
h =
  1.0e-007 *
  Columns 1 through 7
    0.7571    0.7572    0.0001    0.0000    0.0000    0.0000    0.0000
  Columns 8 through 14
    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
  Columns 15 through 20
    0.0000    0.0000    0.0000    0.0000    0.0000         0
```

Nemlineáris egyenletek és egyenletrendszerek

Sorozatok konvergenciája, hibabecslése

5.1. Mindkét sorozat a nullához tart és nemnegatív elemű. Ezért ahhoz, hogy belássuk, hogy pl. az e_k sorozat rendje (legalább) r (≥ 1), azt kell igazolni, hogy van egy olyan K k-tól független konstans, mellyel $e_{k+1} \leq K e_k^r$, azaz $e_{k+1}/e_k^r \leq K$. Természetesen a cél a lehető legnagyobb megfelelő r megkeresése.

Az első sorozatra tehát

$$\frac{a_{k+1}}{a_k^r} = \frac{1/(k+1)}{1/k^r} = \frac{k^r}{k+1},$$

amely csak $r \leq 1$ esetén marad korlátos. Így a sorozat konvergenciarendje 1.

A másik sorozatra

$$\frac{b_{k+1}}{b_k^r} = \frac{2^{-(k+1)}}{2^{-rk}} = 2^{-1-k(1-r)},$$

amely csak $r \leq 1$ esetén marad korlátos. Így a sorozat konvergenciarendje 1.

5.2. Mindkét sorozat a nullához tart és nemnegatív elemű. Ezért ahhoz, hogy belássuk, hogy pl. az e_k sorozat rendje (legalább) r (≥ 1), azt kell igazolni, hogy van egy olyan K k-tól független konstans, mellyel $e_{k+1} \leq K e_k^r$, azaz $e_{k+1}/e_k^r \leq K$. Természetesen a cél a lehető legnagyobb megfelelő r megkeresése.

Az első sorozatra tehát

$$\frac{e_{k+1}}{e_k^r} = \frac{10^{-2^{k+1}}}{10^{-r2^k}} = 10^{2^k(r-2)},$$

amely csak $r \leq 2$ esetén marad korlátos. Így a sorozat konvergenciarendje 2.

A másik sorozatra

$$\frac{f_{k+1}}{f_k^r} = \frac{10^{-(k+1)^2}}{10^{-rk^2}} = 10^{-k^2-2k-1+rk^2} = 10^{(r-1)k^2-2k-1},$$

amely csak $r \leq 1$ esetén marad korlátos. Így a sorozat konvergenciarendje 1.

5.3. Azt kell megnéznünk, hogy az $e_k = |x_k - 2|$ jelöléssel melyik az a legnagyobb r pozitív egész szám (feltételezzük, hogy egész lesz a legnagyobb ilyen szám), melyre teljesül, hogy e_{k+1}/e_k^r korlátos marad $k \to \infty$ esetén. MATLAB-ban az alábbi módon kísérletezhetünk. Látszik, hogy a sorozat negyedrendben konvergens.

```
x=[2.100000000000000 2.040000000000000 2.001024000000000 2.000000000439805]
x =
    2.1000    2.0400    2.0010    2.0000
>> e=abs(x-2)
e =
    0.1000    0.0400    0.0010    0.0000
>> e(2:4)./e(1:3).^1
ans =
    0.4000    0.0256    0.0000 % korlátos marad, nullához tart
>> e(2:4)./e(1:3).^2
ans =
    4.0000    0.6400    0.0004 % korlátos marad, nullához tart
>> e(2:4)./e(1:3).^3
ans =
   40.0000   16.0000    0.4096 % korlátos marad, nullához tart
>> e(2:4)./e(1:3).^4
ans =
  400.0000  400.0000  400.0004 % korlátos marad kb. 400-hoz tart
>> e(2:4)./e(1:3).^5
ans =
  1.0e+005 *
    0.0400    0.1000    3.9063 % nem marad korlátos
```

5.4. Azt kell megnéznünk, hogy az $e_k = |x_k - 5|$ jelöléssel igaz-e, hogy e_{k+1}/e_k^2 korlátos marad. MATLAB-ban számolva könnyen látszik, hogy ez tényleg így van.

```
>> x=[5.200000000000000
      5.080000000000000
      5.012800000000000
      5.000327680000000
      5.000000214748365
      5.000000000000092]'
x =
    5.2000    5.0800    5.0128    5.0003    5.0000    5.0000
```

```
>> e=abs(x-5)
e =
    0.2000    0.0800    0.0128    0.0003    0.0000    0.0000
>> e(2:6)./e(1:5).^2
ans =
    2.0000    2.0000    2.0000    2.0000    2.0030
% látható, hogy ez a sorozat korlátos
```

5.5. Induljunk ki a Lagrange-féle középértéktételből, melynek feltételei nyilván teljesülnek az f függvényre: létezik olyan c szám az x és $x^\star$ értékek között, mellyel

$$f'(c) = \frac{f(x) - f(x^\star)}{x - x^\star} = \frac{f(x)}{x - x^\star},$$

amiből

$$|x - x^\star| \leq \frac{|f(x)|}{m_1}$$

már következik.

Zérushelyek lokalizációja

5.6. Mivel $f(1) = -1$, $f(e) = e - 1 > 0$ és f folytonos függvény, így az 5.1. tétel miatt az adott intervallumban valóban van zérushely. Mivel $f'(x) = \ln x > 0$, ha $x > 1$, ezért a függvény szigorúan monoton növő az adott intervallumon, ami mutatja, hogy csak egy zérushely van az adott intervallumban.

5.7. Mivel $p'(x) = 3x^2 - 4 * x + 4$ és ennek a polinomnak a diszkriminánsa negatív, így a polinom szigorúan monoton növő. Az is nyilvánvaló, hogy $-\infty$-ben $-\infty$-hez tart, ∞-ben pedig ∞-hez. Ebből következik, hogy a polinomnak egyetlen zérushelye van csak. Mivel $p(0) = -4$, így olyan $x > 0$ értéket kell keresnünk, melyre $p(x) > 0$. Az $x = 2$ választás megfelelő, hiszen $p(2) = 4$. Tehát a [0,2] intervallum tartalmazza az egyetlen zérushelyet.

5.8. Az nyilvánvaló, hogy a polinom $-\infty$-ben $-\infty$-hez tart, ∞-ben pedig ∞-hez. Mivel $p'(x) = 3x^2 - 4x + 1$, aminek zérushelyei 1/3 és 1, ezért 1/3-ig növő, majd 1-ig csökkenő, majd 1-től újra növő a függvény. Továbbá mivel $p(1/3) = 13/270$, így 1/3-tól balra pontosan egy zérushely lehet csak. Mivel $p(0) = -1/10$, így ez a zérushely a $[0, 1/3]$ intervallumba kell hogy essen. Mivel $p(1) = -1/10$, így 1-től jobbra is van egy zérushely. Mivel $p(2) = 19/10$, így az $[1, 2]$ intervallum is tartalmaz egy zérushelyet. Továbbá a korábbiak alapján az $[1/3, 1]$ intervallumban is kell lennie zérushelynek.

5.9. Legyen $f(x) = x^2 e^x$ és $g(x) = \sin x$. Azt kell megmondanunk, hogy a két függvény grafikonja hányszor metszi egymást és hol. Az $f(x)$ függvény $-\infty$-ben 0+-hoz tart, ∞-ben ∞-hez. Továbbá -2-ig (itt az értéke 0.54136) növő, majd 0-ig (értéke 0) csökkenő, és ezután újra szigorúan monoton növő a függvény. Ha ezt a grafikont összevetjük a $g(x)$ függvény grafikonjával, akkor láthatjuk, hogy végtelen sok megoldása lesz az egyenletnek. Pozitív megoldás egyetlen egy lesz csak a $[0, \pi]$ intervallumon belül. A legnagyobb negatív megoldás pedig a $[-3\pi/2, -\pi]$ intervallumba esik.

5.10. Használjuk az 5.2. tételt! $A = a = 5$, így $1/(1 + 5/4) = 4/9 \leq |x| \leq 1 + 5/1 = 6$.

Intervallumfelezési módszer

5.11. Az adott intervallumban valóban van zérushely, hiszen a függvény folytonos és az intervallum végpontjaiban különböző az előjele: $a = 0$-ban $f(a) = -4$ és $b = 4$-ben $f(b) = 64$. Mivel $b - a = 4$, így $|x_0 - x^\star| \leq 2$ és az $|x_k - x^\star| \leq 2^{1-k} \leq 10^-2$ becslésből következik, hogy $k = 8$ már megfelelő lesz az adott hiba eléréséhez (5.3.). Az alábbi módon számolhatunk:

k	a	x_k	b	$f(x_k)$
0	0	2	4	6
1	0	1	2	-2
2	1	1.5	2	0.875
3	1	1.25	1.5	-0.7989
4	1.25	1.375	1.5	-0.0254
5	1.375	1.4375	1.5	0.4080
6	1.375	1.40625	1.4375	0.1872
7	1.375	1.390625	1.40625	0.0799
8	1.375	1.3828125	1.390625	0.0270

Azaz $x_8 = 1.3828125$ egy megfelelő közelítése a zérushelynek.

5.12. A keresett érték megoldása pl. az $x^3 - 25 = 0$ egyenletnek. Ez a megoldás a [2,3] intervallumban van. Így az 5.3. tétel miatt az elsőre adott hibához az $1/2^{k+1} \leq 1/10$ feltételnek kell teljesülni, azaz elegendő három iterációs lépést elvégezni.

k	a	x_k	b	$f(x_k)$
0	2	2.5	3	-9.375
1	2.5	2.75	3	-4.2031
2	2.75	2.875	3	-1.2363
3	2.875	2.9375	3	0.3474

Így 2.9375 megfelelő közelítése $\sqrt[3]{25}$-nek.

5.13. Az alábbi program az intervallumfelezési eljárást hajtja végre. A működéséhez meg kell adni rendre azt a függvényt, melynek a zérushelyét keressük, azt az [a,b] intervallumot, amelyben a zérushelyet kell keresni, ill. az elérni kívánt hibát. Ha ez utóbbit nem adjuk meg, akkor azt a program 10^{-6}-ra állítja be. A program az elérni kívánt hibaértékből kiszámítja a szükséges iterációszámot és annyi iterációs lépést hajt végre. Eredményül a kapott közelítést, a tényleges iterációszámot, és a számított szükséges iterációszámot adja vissza. (Lehet, hogy hamarabb leáll a program, mint a szükséges iterációszám, ha az egyik felezőpont már a zérushelyet adja.)

```
function [megold,iter,maxiter]=interfel(fv,a,b,hiba)
f=inline(fv,'x'); megold=a+(b-a)/2;
if f(a)*f(b)>0 error('Nem garantált, hogy van zérushely..
                                    az intervallumban.');
else
if nargin==3, hiba=10^-6; end;
maxiter=ceil(log((b-a)/hiba)/log(2)-1);
iter=0;
while iter<maxiter && abs(f(megold))>10^-60
if f(megold)*f(a)>0 a=megold; else b=megold; end
megold=a+(b-a)/2; iter=iter+1;
end
end
format long
```

5.14. Alkalmazzuk az 5.13. feladatban szereplő programot 10^{-6}-os pontossággal! Az $f(x)$ függvény zérushelyére 1.341851234436035, míg a $g(x)$-ére -2.191308021545410 adódik. Minkét esetben 19 iterációs lépés elegendő az adott pontosság eléréséhez.

Newton-módszer

5.15. A [4] jegyzet (5.1.4) egyenlősége szerint

$$|e_{k+1}| = \frac{f''(\xi_k)}{2f'(x_k)}|e_k|^2$$

(ennek gyengített változatát becslésként tartalmazza az 5.6. tétel), ahol ξ_k x_k és $x^\star$ közé esik. Itt a szokásos $e_k = x_k - x^\star$ jelölést használjuk az iterációs lépés hibájára. Ha feltesszük, hogy x_1 közel van a zérushelyhez, akkor használhatjuk az

$$e_{k+1} \approx \frac{f''(x_1)}{2f'(x_1)}|e_k|^2 \approx 0.57e_k^2$$

közelítést. Ezzel

$$e_{k+1} \approx 0.57e_k^2 \approx 0.57(0.57e_{k-1}^2)^2 = 0.57^3 e_{k-1}^4 \approx \ldots \approx 0.57^M e_0^{M+1},$$

ahol $M = 1+2+\ldots+2^k = 2^{k+1}-1$. Ez a becslés csak akkor mutat konvergenciát, ha e_0 elegendően kicsi. Mivel $x_1 - x_0 = x_1 - x^\star - (x_0 - x^\star) = e_1 - e_0 = -0.1$ és $e_1 \approx 0.57e_0^2$, így $e_0 - 0.1 \approx 0.57e_0^2$, ahonnét azt kapjuk, hogy $e_0 \approx 0.11$. (A másik gyök nem jöhet szóba, mert akkor π-nél nagyobb zérushelyet kapnánk.)

Így az

$$e_{k+1} \approx 0.57^M 0.11^{M+1} = 0.11 \cdot 0.0627^M \leq 5 \cdot 10^{-6}$$

feltételt kell garantálnunk az 5 tizedesjegyes pontossághoz. Innét M értékére $M = 3.61$ adódik, ami azt jelenti, hogy kb. két lépést kell elvégezni az adott pontosság eléréséhez ($k = 2$). (Valójában 3 iterációs lépés kell az adott pontosság eléréséhez.)

5.16. Az $e^{-x} = 10 - x^2$ egyenlőség két oldalán álló függvényeket ábrázolva könnyen látható, hogy két megoldás lesz. A pozitív zérushely valahol $\sqrt{10}$ közelébe esik, könnyen látható az is, hogy innét indítható is az iteráció. (A függvény és második deriváltja is pozitív a zérushelyig terjedő intervallumban.) Első lépésben 3.155539727, a másodikban 3.155532331 és a harmadikban ugyanaz adódik, így 3.155532331 már megfelelő közelítést ad.

5.17. Az $e^{-x} - \sin x$ függvény legkisebb pozitív zérushelye 0 és $\pi/2$ között van. Ezen az intervallumon a függvény második deriváltja pozitív és pl. az $x = 0$ pontban a függvényérték is. Az $x^{(0)} = 0$ pontból tehát indíthatjuk az iterációt!

$$x^{(1)} = 0.5,\ x^{(2)} = 0.585644,\ x^{(3)} = 0.588529,\ x^{(4)} = 0.588533.$$

Ez az eredmény már elfogadható.

5.18. Mivel páratlan fokszámú a polinom, ezért legalább egy zérushelye van. A derivált zérushelyei $\pm 1/\sqrt{3}$, és ezekben a pontokban negatív értéket vesz fel a polinom. Így egyetlen zérushely van az $(1/\sqrt{3}, \infty)$ intervallumban. Könnyen látható, hogy 1-ben negatív, 2-ben meg pozitív a polinom értéke, így a zérushely 1 és 2 között van valahol. Mivel a második derivált $6x > 0$, ha $x > 0$, így a Newton-módszer pl. az $x_0 = 2$ ponttól indítható. Valóban: 4 tizedesjegyre számolva a 3. lépésben már megfelelő eredményt kapunk: 1.7963.

5.19. Az x^4 és $x + 10$ függvényeket ábrázolva látható, hogy az [1,2] intervallumban van a keresett megoldás. Dolgozzunk a Newton-módszerrel! (Ez a módszer másodrendű, így talán nem kell sokat számolni az adott pontosság eléréséhez.) Az $f(x) = x^4 - x - 10$ jelöléssel, $f''(x) = 12x^2$, így pozitív függvényértéket adó helyről kell indítani az iterációt. Legyen ez $x_0 = 2$. Az $x_{k+1} = x_k - f(x_k)/f'(x_k)$ iterációval számolva a harmadik és a negyedik lépések között már nincs változás a negyedik tizedesjegyben, azaz az $x_4 = 1.855585$ érték már biztosan pontos lesz három tizedesjegyre. (Kisebb lesz a hiba, mint $5 \cdot 10^{-4}$.)

5.20. Indítsuk az iterációt az $x_0 = 2$ pontból! Ekkor a sorozat monoton csökkenő módon fog konvergálni az egyenlet megoldásához. Így $|f(x)| \leq x_k^2 - 2$ és $|f'(x)| \geq 2$ az $x^\star$ és x_k közötti intervallumon. Tehát érvényes a

$$|x_k - x^\star| \leq \frac{|x_k^2 - 2|}{2}$$

becslés. Leállási feltételt úgy kaphatunk, hogy a fenti egyenlőtlenség jobb oldalát minden iterációs lépésben kiszámítjuk, és ha az egy adott toleranciaszint alá kerül, akkor biztosak lehetünk benne, hogy x_k a toleranciaszintnél közelebb van a megoldáshoz. Pl. ha a toleranciaszint 10^{-10}, akkor a negyedik lépés után már leállíthatjuk az iterációt:

```
k =
     1
x_k =
   1.500000000000000
hibabecsles =
   0.125000000000000
k =
     2
x_k =
   1.416666666666667
hibabecsles =
   0.003472222222222
k =
     3
x_k =
   1.414215686274510
hibabecsles =
    3.003652441435634e-006
k =
     4
x_k =
   1.414213562374690
hibabecsles =
    2.255307052223543e-012
```

5.21. A zérushely a $[0, \pi/2]$ intervallumba esik. Mivel $f'(x) = -\sin x - 1$ és $f''(x) = -\cos x$, ezért pl. $x_0 = 1.5$-ről indítható az iteráció. Mivel a második derivált negatív, így monoton csökkenő lesz az iterációs sorozat. Az x_k és $x^\star$ pontok között tehát érvényes, hogy

$$|f(x)| = |\cos x - x| \leq |\cos x_k - x_k|, \;\; |f'(x)| = |-\sin x - 1| \geq 1,$$

így tehát

$$|x_k - x^\star| \leq \frac{|\cos x_k - x_k|}{1} = |\cos x_k - x_k|.$$

10^{-10}-es toleranciaszinttel számolva azt kapjuk a becslésből, hogy a negyedik lépés után már leállíthatjuk az iterációt.

```
k =
     1
x_k =
   0.785314737732760
hibabecsles =
   0.078148968153730
k =
     2
x_k =
   0.739534550025702
hibabecsles =
    7.522240085812149e-004
k =
     3
x_k =
   0.739085177791409
hibabecsles =
    7.460334550124514e-008
k =
     4
x_k =
   0.739085133215161
hibabecsles =
    7.771561172376096e-016
```

5.22. A módszer azért nem lesz másodrendű, mert a zérushelynél a derivált értéke nulla ($3 \cdot 1^2 - 3 = 0$), így az 5.6. tétel feltételei nem teljesülnek. Ha a Newton-módszert úgy módosítjuk, hogy

$$x_{k+1} = x_k - 2\frac{f(x_k)}{f'(x_k)},$$

akkor az iteráció már másodrendben fog konvergálni. Ezt az alábbi módon igazolhatjuk.

Mivel $x^\star$ kétszeres zérushely, így $f(x^\star) = f'(x^\star) = 0$. Vonjunk ki az iterációs formula minkét oldalából $x^\star$-ot, majd szorozzunk $f'(x_k)$-val.

$$(x_{k+1} - x^\star)f'(x_k) = (x_k - x^\star)f'(x_k) - 2f(x_k). \tag{10.2}$$

Taylor-sorfejtést használva az $x^\star$ pont körül, azt kapjuk, hogy

$$f'(x_k) = f'(x^\star) + f''(\xi)(x_k - x^\star) = f''(\xi)(x_k - x^\star),$$

valamint a sorfejtést a jobb oldalon álló $G(x) = (x - x^\star)f'(x) - 2f(x)$ függvényre használva, mivel

$$G(x^\star) = 0,\ G'(x^\star) = f'(x_k) + (x^\star - x^\star)f''(x^\star) - 2f'(x^\star) = 0,$$

$$G''(x^\star) = (x^\star - x^\star)f'''(x^\star) = 0,$$

így

$$G(x_k) = \frac{G'''(\eta)}{6}(x_k - x^\star)^3.$$

A fenti kifejezéseket visszahelyettesítve a 10.2 egyenlőségbe azt kapjuk, hogy

$$(x_{k+1} - x^\star)f''(\xi)(x_k - x^\star) = \frac{G'''(\eta)}{6}(x_k - x^\star)^3.$$

Ezt átrendezve, majd abszolút értéket véve kapjuk az alábbi becslést, ami már mutatja, hogy a konvergencia másodrendű.

$$|x_{k+1} - x^\star| \leq \frac{G'''_{\max}}{6f''_{\min}}|x_k - x^\star|^2,$$

ahol $G'''_{\max}$ egy felső becslés G harmadik deriváltjának abszolút értékére és $f''_{\min}$ egy alsó becslés f első deriváltjának abszolút értékére az $\mathbf{x}^\star$ zérushely egy környezetében.

A másodrendű konvergencia az 5.8. tételre támaszkodva is igazolható. Ehhez a módszert olyan fixpont iterációnak tekintjük, amelynek iterációs függvénye

$$F(x) = x - 2\frac{f(x)}{f'(x)}.$$

Az f függvényre érvényes, hogy $f(x^\star) = f'(x^\star) = 0$ és $f''(x^\star) \neq 0$, továbbá

$$\lim_{x\to x^\star} F'(x) = \lim_{x\to x^\star}\left(1 - 2\frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2}\right)$$

$$= -1 + 2f''(x^\star)\lim_{x\to x^\star}\frac{f'(x)}{2f'(x)f''(x)} = -1 + 2f''(x^\star)\frac{1}{2f''(x^\star)} = 0.$$

Ez mutatja, hogy a konvergencia legalább másodrendű lesz. Az, hogy nem lesz magasabbrendű, az onnét látszik, hogy $F''(x)$-nek már nem nulla a határértéke $x^\star$-ban.

Általánosan beláttuk tehát azt a tételt, hogy ha f kétszer folytonosan deriválható és $x^\star$-ban kétszeres zérushelye van, akkor a Newton-módszer fenti módosítása már másodrendben konvergens zérushelyhez tartó sorozatot ad.

5.23. Mivel $x^\star$ m-szeres zérushely, így $f(x^\star) = \ldots = f^{(m-1)}(x^\star) = 0$. Induljunk ki az

$$x_{k+1} = x_k - m\frac{f(x_k)}{f'(x_k)}$$

iterációból és vonjunk ki mindegyik oldalából $x^\star$-ot:

$$x_{k+1} - x^\star = x_k - x^\star - m\frac{f(x_k)}{f'(x_k)},$$

majd szorozzunk $f'(x_k)$-val:

$$(x_{k+1} - x^\star)f'(x_k) = (x_k - x^\star)f'(x_k) - mf(x_k). \tag{10.3}$$

Most a bal oldalon álló $f'(x_k)$ értéket és a jobb oldalon álló $(x_k - x^\star)f'(x_k) - mf(x_k)$ értéket is a függvények $x^\star$ pont körüli sorfejtéséből számoljuk ki. Egyrészt

$$f'(x) = f'(x^\star) + \ldots + \frac{f^{(m-1)}(x^\star)}{(m-2)!}(x-x^\star)^{m-2} + \frac{f^{(m)}(\eta)}{(m-1)!}(x-x^\star)^{m-1} = \frac{f^{(m)}(\eta)}{(m-1)!}(x-x^\star)^{m-1},$$

másrészt a

$$G(x) := (x - x^\star)f'(x) - mf(x)$$

függvényre

$$G^{(j)}(x) := jf^{(j)}(x) + (x - x^\star)f^{(j+1)}(x) - mf^{(j)}(x),$$

ami azt mutatja, hogy

$$G^{(j)}(x^\star) = 0,$$

ha $j = 0, 1, \ldots, m$. Így tehát G $x^\star$ körüli sorfejtése

$$G(x) = \frac{(x - x^\star)^{m+1}}{(m+1)!}G^{(m+1)}(\xi).$$

A fenti $f(x)$ és $G(x)$ függvényeket az $x^\star$ helyen kiszámolva és a (10.3) képletbe helyettesítve kapjuk, hogy

$$(x_{k+1} - x^\star)\frac{f^{(m)}(\eta)}{(m-1)!}(x_k - x^\star)^{m-1} = \frac{(x_k - x^\star)^{m+1}}{(m+1)!}G^{(m+1)}(\xi),$$

azaz

$$|x_{k+1} - x^\star| \le \frac{G^{(m+1)}_{\max}}{m(m+1)f^{(m)}_{\min}}|x_k - x^\star|^2,$$

ahol $G^{(m+1)}_{\max}$ a G függvény $m+1$-edik derivált abszolút értékének maximumát, míg $f^{(m)}_{\min}$ az f függvény abszolút értékének minimumát adja meg az $x^\star$ pont egy környezetében. Ilyen értékek nyilvánvalóan léteznek és az utóbbi nem nulla. Ez a képlet nyilvánvalóan azt mutatja, hogy a konvergencia másodrendű.

Természetesen sokszor nem tudható előre, hogy az f függvénynek a keresett zérushelye hányszoros zérushely. Ezzel kapcsolatban lásd az 5.24. feladatot.

5.24. Ha f-nek $x^\star$ m-szeres zérushelye, akkor felírható $f(x) = (x-x^\star)^m h(x)$ alakban, ahol már $h(x^\star) \neq 0$, hasonlóan $f'(x)$-nek $x^\star$ $m-1$-szeres zérushelye, így hasonlóan $f(x)$-hez $f'(x)$ az $f'(x) = (x-x^\star)^{m-1} j(x)$ alakban írható. Így

$$g(x) = \frac{f(x)}{f'(x)} = \frac{(x-x^\star)h(x)}{j(x)},$$

így $x^\star$ valóban egyszeres zérushely, mert h és j olyan függvények, melyek $x^\star$-ban nem nullát vesznek fel.

Ezek alapján azt mondhatjuk, hogy a klasszikus Newton-módszer mindig másodrendben fog konvergálni, ha a $g(x) = f(x)/f'(x)$ függvényre alkalmazzuk, azaz ha az $f(x) = 0$ egyenlet megoldását az

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$$

iterációval keressük.

5.25. Az iteráció sajnos ciklikusan ismétlődő lépéseket állít elő: $x_0 = 0$, $x_1 = -1$, $x_2 = 0$, $x_3 = -1$, ..., így az nem fog konvergálni a megoldáshoz.

Az x_0 kezdőérték nincs elegendően közel a megoldáshoz ahhoz, hogy az 5.6. tétel biztosítsa a konvergenciát. Mivel a megoldás -0.4 környékén van, $m_1 = 1$ és $M_2 = 10$ megfelelő választások, így az $|x_0 - x^\star| \leq 0.2$ feltétel már biztosítaná a konvergenciát. Valóban, a [-0.6,-0.2] intervallumból indítva az iterációt az konvergens lesz és az $x^\star = -0.3977508105$ értéket adja eredményül.

Húr- és szelőmódszerek

5.26. Mivel $f(-1) < 0$ és $f(1) > 0$, így valóban van megoldás az adott intervallum belsejében. A számításokat az alábbi táblázatban foglaltuk össze. A számítások leellenőrizhetők pl. a chord.m program segítségével.

k	a	b	x_k	$f(x_k)$
1	-1	1	0.2939	0.9162
2	-1	0.2939	-0.3280	-0.0852
3	-0.3280	0.2939	-0.2751	0.0205
4	-0.3280	-0.2751	-0.2854	-1.8886e-005

5.27. A szelőmódszerrel az alábbi módon számolhatunk a [-1,1] intervallumról indítva az eljárást. A számítások leellenőrizhetők pl. a secant.m program segítségével.

k	x_k
0	1
1	-1
2	0.293858536281304
3	-0.328029789700643
4	-0.275142163493936
5	-0.285407606179304
6	-0.285398162735863
7	-0.285398163397448

5.28. A Newton-módszer esetén minden lépésben egy új függvényértéket és egy új derivált értéket kell kiszámolnunk, míg a szelőmódszer esetén elegendő minden lépésben egy újabb függvényérték számolása. Így fordulhat elő, hogy egy alacsonyabb rendű módszer időben gyorsabban megtalálja a zérushelyet egy adott pontossággal, mint a másodrendű Newton-módszer. Ha pl. addig végezzük az iterációt, amíg két egymás utáni közelítés már 10^{-15}-nél kisebb lesz, akkor a Newton-módszernek 6 iterációra, míg a szelőmódszernek 7 iterációra van szüksége. Viszont a Newton-módszer 6 iterációja 0.005944 másodpercet, míg a szelőmódszer 7 iterációja 0.003552 másodpercet vesz igénybe. (Természetesen a futási idő függ a használt számítógéptől, de a futási idők aránya hasonló lesz.) A pontos megoldás 0.685504401504941.

Fixpont iterációk

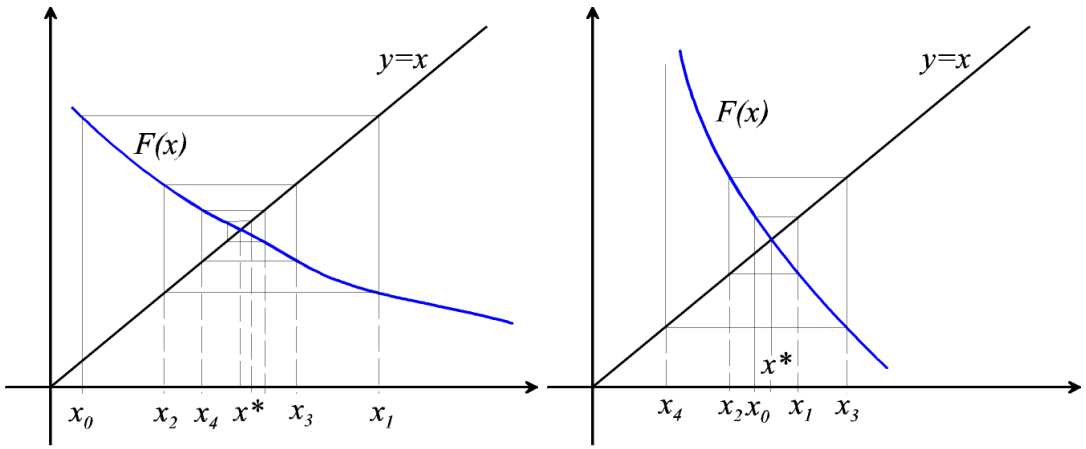
5.29. A fixpont iteráció grafikusan úgy szemléltethető, hogy az $(x_0, 0)$ ponton keresztül húzunk egy olyan függőleges szakaszt, ami metszi az $F(x)$ függvény grafikonját. A metszéspontot vízszintes szakasszal összekötjük az $y = x$ egyenes grafikonjával. Ahol ez a szakasz metszi az egyenest, azt a pontot összekötjük ismét egy függőleges szakasszal az $F(x)$ függvény grafikonjával, majd azt ismét egy vízszintes szakasszal az $y = x$ egyenessel, stb. A függőlegesen húzott szakaszok meghosszabításainak x-tengellyel alkotott metszéspontjai adják rendre az $x_1, x_2, \ldots$ iterációs lépéseket. A 10.5 ábrán egy konvergens és egy divergens esetet szemléltettünk. Az első esetben $|F(x)| \leq q < 1$, a másodikban pedig $|F(x)| > 1$.

5.30. Az iterációs függvény

$$F(x) = \ln(1 + x) - x + x^2/2,$$

melynek deriváltja

$$F'(x) = \frac{x^2}{1 + x},$$



10.5. ábra. Egy konvergens és egy nem konvergens fixpont iterációs eljárás szemléltetése.

ami $x = 0$ esetén nulla és folytonos, így az origó egy megfelelő környezetében biztosan kisebb abszolút értékű lesz, mint $q < 1$. Tekintsük pl. a $[-0.5, 0.5]$ intervallumot. Ebben

$$\left|\frac{x^2}{1+x}\right| \leq \frac{1}{4/2} = \frac{1}{2} = q,$$

azaz ebből az intervallumból indítva az iterációt az a fixponthoz fog konvergálni.

Mivel $F'(0) = F''(0) = 0$, de $F'''(0) \neq 0$, így a konvergencia harmadrendű lesz.

5.31. Legyen az iterációs függvény

$$F(x) = x + A\left(\frac{x^2-2}{x}\right) + B\left(\frac{x^2-2}{x^3}\right).$$

Ennek nyilván fixpontja az $x^\star = \pm\sqrt{2}$ pont. A konvergenciarend annál nagyobb, minél magasabb rendű deriváltja tűnik el F-nek a fixpontban. Az első és a második derivált is eltűnik, ha teljesülnek az $1 + 2A + B = 0$ és $-A - 5B/2 = 0$ feltételek. Ezen egyenletrendszer megoldása $A = 1/4$, $B = -5/8$. Az iteráció ezek alapján harmadrendben lesz konvergens.

5.32. Most a Newton-módszer, vagy az $x_{k+1} = 2/x_k$ iteráció nem jöhet szóba, mert az $x_0 = 0$ pontban az iterációs függvények nincsenek értelmezve. A megfelelő iterációs függvény megkonstruálására egy lehetőség pl. az alábbi.

Próbálkozzunk iterációt konstruálni az

$$x = x + g(2 - x^2)$$

ekvivalens egyenlettel, ahol g megfelelő pozitív konstans. A konvergenciához biztosítani kell, hogy pl. az [1,2] intervallumon (ebben van a zérushely) az $F(x) = x + g(2 - x^2)$ függvény kontrakció legyen. $F'(x) = 1 - 2gx$. Garantáljuk pl., hogy $|1 - 2gx| \leq 1/2$. Ehhez elég pl. g értékét 1/4-nek választani.

Ezzel az iterációnk lehet az $x_{k+1} = x_k + (2 - (x_k)^2)/4$ alakú, és eddig azt tudjuk, hogy az $[1, 2]$ intervallumból indítva az iteráció az egyenlet megoldásához konvergál.

Már csak azt kellene megmutatni, hogy $x_0 = 0$-ról is konvergálni fog, amihez elegendő megmutatni, hogy az iteráció valamelyik lépésben belekerül az [1,2] intervallumba. Az $x_0 = 0$ pontból indítva az iterációt $x_1 = 1/2$, $x_2 = 15/16$ és $x_3 = 1247/1024$, ami már beleesik az $[1, 2]$ intervallumba. (Ez az $x + (2 - x^2)/4$ iterációs függvény grafikonjából is látszik.) Ezt akartuk megmutatni.

A hibabecslést csak az $[1, 2]$ intervallumban lévő sorozatrészre lehet a tanult képlettel csinálni ($k \in \mathbb{N}$):

$$|x_{3+k} - x^\star| \leq \frac{(1/2)^k}{1/2}|x_4 - x_3| \leq \frac{1}{2^{k-1}} < 10^{-6}$$

(ahol egyszerűen 1-gyel becsüljük felülről az $|x_4 - x_3|$ értéket), ahonnét $k = 21$ adódik. Azaz az eredeti sorozattal legalább 24 lépés szükséges az adott pontossághoz.

5.33. Ha a fixpont iteráció a k-adik lépésben az x_k pontból az x_{k+1} pontba lép, akkor a Banach-féle fixponttételnél tanult becslés alapján

$$|x_{k+s} - x^\star| \leq \frac{q^s}{1 - q}|x_{k+1} - x_k|,$$

ahol most $x^\star$ az $F(x) = 0.5 + \sin x$ iterációs függvény fixpontja és q a kontrakciós tényező. Az iterációs függvény deriváltja $\cos x$, így az $[1, 1.5]$ intervallumon (1-ről indulunk ($k = 0$ a fenti becslésben) és 1.5 körüli eredményt várunk fixpontnak) $F(x)$ kontrakciós tényezője $\cos 1 \approx 0.55$, amivel csak a

$$|x_{10} - x^\star| \leq \frac{q^{10}}{1 - q}|x_1 - x_0| = 0.001922$$

becslést nyerjük, ami még nem megfelelő számunkra. Láthatjuk, hogy $x_1 = 1.34147$, így most vizsgáljuk meg, hogy az $[x_1, 1.5]$ intervallumon mekkora a kontrakciós tényező: $\cos x_1 \approx 0.227$. Ezzel

$$|x_{1+9} - x^\star| \leq \frac{q^9}{1 - q}|x_2 - x_1| = 2.7391 \cdot 10^{-7},$$

ami mutatja, hogy x_{10} már legalább 6 tizedesjegyre pontos lesz.

5.34. A Banach-féle fixponttétel biztosítja a konvergenciát, ha igazoljuk, hogy az

$$F(x) = \frac{x}{3} + \frac{1}{x}$$

függvény kontrakció az $[1,2]$ intervallumon, és hogy az intervallumot önmagába képezi.

A kontrakcióhoz elég belátni, hogy $|F'(x)| < 1$ az [1,2] intervallumon, hiszen $F'(x)$ folytonos.

$$F'(x) = \frac{1}{3} + \frac{-1}{x^2},$$

így

$$\frac{-2}{3} \leq F'(x) \leq \frac{1}{3}.$$

Így a leképezés valóban kontrakció, a kontrakciós tényező 2/3.

$F(x)$ $\sqrt{3}$-tól balra csökkenő, jobbra növő az [1,2] intervallumon. Ez látszik a deriváltból. Így maximuma $\max\{4/3, 7/6\} = 4/3 = 1.3333$, minimuma $2/\sqrt{3} \approx 1.1547$. Azaz az intervallumot önmagába képezi.

A Banach-féle fixponttétel miatt tehát az iteráció az [1,2] intervallumbeli egyetlen fixponthoz tart, ami $\sqrt{3/2}$.

Hibabecslés szintén a Banach-féle fixponttétellel adható. Ha $x_0 = 2$, akkor $x_1 = 7/6$, azaz

$$|x_k - \sqrt{3/2}| \leq \frac{(2/3)^k}{1-(2/3)} \left|2 - \frac{7}{6}\right| \leq 10^{-3},$$

ahonnét kapjuk, hogy a 20. tagtól már beleesik a kívánt környezetbe a sorozat.

5.35. Egyszerű számítással látható, ahogy $\sqrt[3]{21}$ mindegyik iterációnak fixpontja. Mivel $\sqrt[3]{21}$ értéke a [2,3] intervallumban van, így a konvergenciához elég megvizsgálni pl., hogy a Banach-féle fixponttétel feltételei teljesülnek-e.

Az első iteráció a $[2,3]$ intervallumot önmagába képezi, és az iterációs függvény deriváltjának abszolút értékének maximuma 166/189. Így teljesülnek a Banach-féle fixponttétel feltételei. Mivel az iterációs függvény deriváltja a fixpontban 6/7, azaz nem nulla, így az iteráció elsőrendben tart a fixponthoz.

A második iterációs függvény a [2.1,3] intervallumot önmagába képezi, és az iterációs függvény deriváltjának abszolút értékének maximuma 1118/1323. Így teljesülnek a Banach-féle fixponttétel feltételei. Mivel az iterációs függvény deriváltja a fixpontban 0, a második deriváltja nem 0 ($(2/21)21^{2/3}$), így az iteráció másodrendben konvergál a fixponthoz.

A harmadik iterációs függvény deriváltja a fixpontban 5.7057, azaz az iteráció nem konvergál a fixponthoz.

Összefoglalva tehát a harmadik iteráció nem konvergens, az első elsőrendben, a második pedig másodrendben konvergens.

5.36. Az iteráció

$$x_{k+1} = x_k - \frac{2f'(x_k)f(x_k)}{2(f'(x_k))^2 - f(x_k)f''(x_k)}$$

alakját fogjuk használni a bizonyításban.

Fejtsük sorba az $f(x)$ függvényt a harmadik ill. második tagig is az x_k pont körül, és alkalmazzuk a sorfejtést az $x^\star$ pontban!

$$0 = f(x^\star) = f(x_k) + f'(x_k)(x^\star - x_k) + \frac{f''(x_k)}{2!}(x^\star - x_k)^2 + \frac{f'''(\xi)}{3!}(x^\star - x_k)^3,$$

$$0 = f(x^\star) = f(x_k) + f'(x_k)(x^\star - x_k) + \frac{f''(\eta)}{2!}(x^\star - x_k)^2,$$

ahol ξ és η megfelelő $x^\star$ és x_k közé eső számok. Szorozzuk be az első egyenletet $2f'(x_k)$-val, a másodikat $f''(x_k)(x^\star - x_k)$-val, majd vonjuk ki az elsőből a másodikat, és rendezzük.

$$0 = 2f(x_k)f'(x_k) + (2(f'(x_k))^2 - f''(x_k)f(x_k))(x^\star - x_k) +$$
$$+ \left(\frac{f'(x_k)f'''(\xi)}{3} - \frac{f''(x_k)f''(\eta)}{2}\right)(x^\star - x_k)^3.$$

Osszuk el mindkét oldalt a $2(f'(x_k))^2 - f''(x_k)f(x_k)$ kifejezéssel.

$$0 = \frac{2f(x_k)f'(x_k)}{2(f'(x_k))^2 - f''(x_k)f(x_k)} + x^\star - x_k + \frac{2f'(x_k)f'''(\xi) - 3f''(x_k)f''(\eta)}{6(2(f'(x_k))^2 - f''(x_k)f(x_k))}(x^\star - x_k)^3.$$

A jobb oldalon álló első tag az iteráció képlete miatt éppen $x_k - x_{k+1}$, így x_k kiesik, és a képletet átrendezve az alábbi egyenlőséget kapjuk.

$$x_{k+1} - x^\star = -\frac{2f'(x_k)f'''(\xi) - 3f''(x_k)f''(\eta)}{6(2(f'(x_k))^2 - f''(x_k)f(x_k))}(x^\star - x_k)^3,$$

ami már mutatja a módszer harmadrendű konvergenciáját.

Nemlineáris egyenletrendszerek megoldása

5.37. Fejezzük ki az x_1, x_2, x_3 ismeretleneket rendre az egyes egyenletekből:

$$x_1 = \frac{1}{3}\cos(x_2x_3) + 1/6 =: F_1(x),$$
$$x_2 = \frac{\pm 1}{9}\sqrt{x_1^2 + \sin x_3 + 1.06} - 0.1 =: F_2(x),.$$
$$x_3 = \frac{-1}{20}e^{-x_1x_2} - (10\pi - 3)/60 =: F_3(x).$$

Először válasszuk $F_2(x)$-ben a gyökjel előtti pozitív előjelet. Megmutatjuk, hogy a fenti iteráció kielégíti az 5.9. tétel feltételeit a $-1 \le x_1, x_2, x_3 \le 1$ kockán. Először azt mutatjuk

meg, hogy a leképezés a kockába képez:

$$|F_1(x)| = \left|\frac{1}{3}\cos(x_2x_3) + 1/6\right| \le 1/2,$$
$$|F_2(x)| = \left|\frac{1}{9}\sqrt{x_1^2 + \sin x_3 + 1.06} - 0.1\right| \le \frac{1}{9}\sqrt{1^2 + \sin 1 + 1.06} - 0.1 \le 0.09,$$
$$|F_3(x)| = \left|\frac{-1}{20}e^{-x_1x_2} - (10\pi - 3)/60\right| \le \frac{e}{20} + (10\pi - 3)/60 \le 0.61.$$

Most megmutatjuk, hogy $|\partial F_i(x)/\partial x_k| \le 0.8430/3 = 0.281$ (azaz az 5.9. tételben $q = 0.8430$) tetszőleges $i, k = 1, 2, 3$ esetén.

$$\left|\frac{\partial F_1}{\partial x_1}\right| = 0, \quad \left|\frac{\partial F_2}{\partial x_2}\right| = 0, \quad \left|\frac{\partial F_3}{\partial x_3}\right| = 0,$$

$$\left|\frac{\partial F_1}{\partial x_2}\right| = \frac{1}{3}|x_3||\sin(x_2x_3)| \le \frac{1}{3}\sin 1 \le 0.281,$$
$$\left|\frac{\partial F_1}{\partial x_3}\right| = \frac{1}{3}|x_2||\sin(x_2x_3)| \le \frac{1}{3}\sin 1 \le 0.281,$$
$$\left|\frac{\partial F_2}{\partial x_1}\right| = \frac{|x_1|}{9\sqrt{x_1^2 + \sin x_3 + 1.06}} < \frac{1}{9\sqrt{0.218}} < 0.238,$$
$$\left|\frac{\partial F_2}{\partial x_3}\right| = \frac{|\cos(x_3)|}{18\sqrt{x_1^2 + \sin x_3 + 1.06}} < \frac{1}{18\sqrt{0.218}} < 0.119,$$
$$\left|\frac{\partial F_3}{\partial x_1}\right| = \frac{|x_2|}{20}e^{-x_1x_2} \le \frac{1}{20}e < 0.14,$$
$$\left|\frac{\partial F_3}{\partial x_2}\right| = \frac{|x_1|}{20}e^{-x_1x_2} \le \frac{1}{20}e < 0.14.$$

A tétel szerint tehát az iterációnak egyetlen fixpontja van az adott kockán belül. Ha $F_2(x)$-ben a negatív előjelet választjuk a gyökjel előtt, akkor a fentiekhez hasonlóan megmutatható, hogy annak az iterációnak is egyetlen fixpontja van. Így igazoltuk, hogy két fixpont van, mivel a kapott két fixpont nem esik egybe (5.38. feladat).

5.38. Az 5.37. feladat eredményét és az 5.9. tételt ($q = 0.8430$) felhasználva dolgozunk.

Indítsuk az első iterációt az $\overline{\mathbf{x}}_0 = [0.6, 0, -0.6]^T$ pontból! Ekkor

$$\overline{\mathbf{x}}_1 = [0.5000, -0.0041, -0.5237]^T,$$

és az alábbi hibabecslést kapjuk

$$\|\overline{\mathbf{x}}_k - \overline{\mathbf{x}}^\star\|_\infty \le \frac{q^k}{1-q}\|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0\|_\infty = \frac{0.8430^k}{1 - 0.8430}0.1 \le 10^{-6}.$$

Azaz legfeljebb 73 iterációs lépés elegendő az adott pontosság eléréséhez.

$$\overline{\mathbf{x}}_{73} = [0.500000000000000, -1.387778780781446e-017, -0.523598775598299]^T$$

($\overline{\mathbf{x}}^\star = [1/2, 0, -\pi/6]^T$). A másik iterációval hasonlóan számolhatunk.

$$\overline{\mathbf{x}}_1 = [0.5000, -0.1959, -0.5287]^T,$$

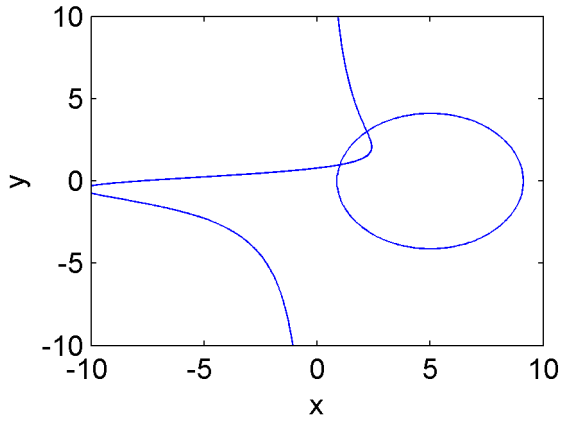
és

$$\|\overline{\mathbf{x}}_k - \overline{\mathbf{x}}^\star\|_\infty \leq \frac{q^k}{1-q}\|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0\|_\infty < \frac{0.8430^k}{1-0.8430}0.2 \leq 10^{-6},$$

azaz legfeljebb 83 iterációs lépésre van szükség.

$$\overline{\mathbf{x}}_{73} = [0.498144684589491, -0.199605895543780, -0.528825977573387]^T.$$

5.39. Ábrázoljuk MATLAB-ban az egyenletrendszer egyenleteit implicit függvényként az adott tartományon. Innét leolvasható, hogy az egyenletrendszernek összesen két megol-



10.6. ábra. Az 5.39. feladatban szereplő implicit függvények grafikonja.

dása van összesen az adott tartományon. (Lásd még az 5.40. és 5.41. feladatokat.)

5.40. Átírjuk az egyenletrendszert alkalmas módon fixpont iterációs alakba

$$x_1 = \frac{x_1^2 + x_2^2 + 8}{10} =: F_1(x_1, x_2),$$
$$x_2 = \frac{x_1 x_2^2 + x_1 + 8}{10} =: F_2(x_1, x_2),$$

majd megmutatjuk, hogy a $D = [0, 1.5] \times [0, 1.5]$ halmazon teljesülnek az 5.9. tétel feltételei.

Az könnyen ellenőrizhető, hogy az $\overline{\mathbf{F}}(x_1, x_2) = (F_1(x_1, x_2), F_2(x_1, x_2))$ leképezés a D halmazból a D halmazba képez. Továbbá

$$\begin{aligned}
\left|\frac{\partial F_1}{\partial x_1}\right| &= \left|\frac{x_1}{5}\right| \leq \frac{1}{5}, \\
\left|\frac{\partial F_1}{\partial x_2}\right| &= \left|\frac{x_2}{5}\right| \leq \frac{1}{5}, \\
\left|\frac{\partial F_2}{\partial x_1}\right| &= \left|\frac{x_2^2}{10} + \frac{1}{10}\right| \leq 0.325, \\
\left|\frac{\partial F_2}{\partial x_2}\right| &= \left|\frac{x_1 x_2}{5}\right| \leq 0.45,
\end{aligned}$$

ami miatt az 5.9. tétel feltételei érvényben vannak a $q = 0.9$ választással, azaz valóban egy fixpont van csak az adott tartományon belül. A fixpont megkereséséhet indítsuk az iterációt az $\overline{\mathbf{x}} = [1/2, 1/2]^T$ pontból! Ekkor $\overline{\mathbf{x}}_1 = [0.85, 0.90625]^T$, és a hibabecslő formulából azt nyerjük, hogy legfeljebb 145 iterációra van szükségünk az adott pontosság eléréséhez $\overline{\mathbf{x}}_{145} = [1, 1]^T$, ami a tényleges pontos megoldást adja az adott D tartományból.

5.41. Írjuk fel a Newton-iterációt az adott egyenletrendszerre! Az egyszerűség kedvéért az x és y változókat használjuk, és nem írjuk ki az iterációs lépések számát.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} 2x - 10 & 2y \\ y^2 + 1 & 2xy - 10 \end{bmatrix}^{-1} \begin{bmatrix} x^2 - 10x + y^2 + 8 \\ xy^2 + x - 10y + 8 \end{bmatrix}.$$

Most a tétel nem ad útmutatást a kezdőpont megválasztására. Azt tudjuk tenni, hogy a 10.6 ábra alapján elindítjuk az iterációt valahonnét a megoldás közeléből. Pl. az $\overline{\mathbf{x}}_0 = [0.5, 0.5]^T$ pontból indítva az az $\overline{\mathbf{x}}^\star = [1, 1]^T$ ponthoz tart, míg az $\overline{\mathbf{x}}_0 = [3, 3]^T$ pontból kezdve az iterációt az

$$\overline{\mathbf{x}}^\star = [2.193439415415308, 3.020466468123034]^T$$

megoldást kapjuk.

5.42. $\overline{\mathbf{x}}^\star = [0.121241911480502, 0.271105155792415]^T$.

Interpoláció és approximáció

Polinominterpoláció

Interpoláció Lagrange és Newton módszerével általános alappontokon

6.1. Legyen $x_0 = -1$, $x_1 = 2$, $x_2 = 3$, $x_3 = 4$ és $f_0 = 2$, $f_1 = 4$, $f_2 = 0$, $f_3 = 2$. Az x_0-hoz tartozó Lagrange-féle alappolinom

$$l_0(x) = \frac{(x-2)(x-3)(x-4)}{(-1-2)(-1-3)(-1-4)} = -\frac{1}{60}x^3 + \frac{3}{20}x^2 - \frac{13}{30}x + \frac{2}{5},$$

az x_1-hez tartozó

$$l_1(x) = \frac{(x-(-1))(x-3)(x-4)}{(2-(-1))(2-3)(2-4)} = \frac{1}{6}x^3 - x^2 + \frac{5}{6}x + 2,$$

az x_2-höz tartozó

$$l_2(x) = \frac{(x-(-1))(x-2)(x-4)}{(3-(-1))(3-2)(3-4)} = -\frac{1}{4}x^3 + \frac{5}{4}x^2 - \frac{1}{2}x - 2$$

és az x_3-hoz tartozó

$$l_3(x) = \frac{(x-(-1))(x-2)(x-3)}{(4-(-1))(4-2)(4-3)} = \frac{1}{10}x^3 - \frac{2}{5}x^2 + \frac{1}{10}x + 3/5.$$

Ezek alapján a Lagrange-féle interpolációs polinom az

$$L_3(x) = 2l_0(x) + 4l_1(x) + 0l_2(x) + 2l_3(x) = \frac{5}{6}x^3 - \frac{9}{2}x^2 + \frac{8}{3}x + 10$$

polinom lesz. Vegyük észre, hogy az interpolációs polinom meghatározásához az $l_2(x)$ polinomra nincs is szükség, hiszen $f_2 = 0$.

6.2. A 6.3. tételt használjuk. Először elkészítjük az osztott differencia táblázatot.

x_i	$f_i = [x_i]f$	$[.,.]f$	$[.,.,.]f$	$[.,.,.,.]f$
-1	$2 = c_0$			
		$\frac{4-2}{2-(-1)} = 2/3 = c_1$		
2	4		$\frac{-4-2/3}{3-(-1)} = -7/6 = c_2$	
		$\frac{0-4}{3-2} - 4$		$\frac{3-(-7/6)}{5} = 5/6 = c_3$
3	0		$\frac{2-(-4)}{2} = 3$	
		$\frac{2-0}{4-3} = 2$		
4	2			

A táblázatban bejelöltük a Newton-alakban felírt polinom együtthatóit. Így a keresett polinom

$$\begin{aligned} L_3(x) &= c_0 + c_1(x-(-1)) + c_2(x-(-1))(x-2) + c_3(x-(-1))(x-2)(x-3) \\ &= 2 + (2/3)(x-(-1)) + (-7/6)(x-(-1))(x-2) + (5/6)(x-(-1))(x-2)(x-3) \\ &= \frac{5}{6}x^3 - \frac{9}{2}x^2 + \frac{8}{3}x + 10. \end{aligned} \tag{10.4}$$

Az interpolációs polinom Horner-alakja a Newton-alak alábbi átrendezésével nyerhető:

$$L_3(x) = 2 + (x+1)\,(\ (2/3) + (x-2)\,(\ (-7/6) + (5/6)(x-3)\ )\ )\,.$$

A helyettesítési értékek számolásához ezt az alakot érdemes használni.

6.3. Természetesen mindkét módszerrel az

$$L_2(x) = \frac{5}{6}x^2 - \frac{29}{6}x + 9$$

polinomot kapjuk. A Lagrange-alakban megadott előállítás

$$L_2(x) = 5\frac{(x-3)(x-4)}{(1-3)(1-4)} + 2\frac{(x-1)(x-4)}{(2-1)(2-4)} + 3\frac{(x-1)(x-3)}{(4-1)(4-3)},$$

míg a Newton-féle előállítás Horner-alakban

$$L_2(x) = 5 + (x-1)\left(\frac{-3}{2} + \frac{5}{5}(x-3)\right).$$

6.4. A Lagrange-féle előállítás esetén minden egyes Lagrange-féle alappolinom helyettesítési értékének kiszámítása $4n-1$ flopba kerül. Ezeket kell szorozni az alappontokbeli

függvényértékekkel, majd össze kell adni őket. Ez összesen $(n+1)(4n-1+1)+n = 4n^2+5n$ flop.

A Newton-féle előállításnál minden osztott differencia 3 flop, valamint $1+2+\ldots+n = n(n+1)/2$ osztott differenciát kell kiszámolnunk. Tehát az osztott differenciák kiszámítása összesen $3n^2/2$ flopba kerül. Ezek után a polinom helyettesítési értékét a Horner-sémával érdemes számolni. Ennek költsége $3n$.

Látható tehát, hogy a Newton-féle előállítás kevésbé költséges. Már egy helyettesítési érték kiszámítása is kevesebb művelet, de ha eltároljuk az osztott differenciákat, akkor egy-egy újabb helyen a polinom helyettesítési értékének kiszámítása már csak egyenként $3n$ flop lesz.

6.5. Vezessük be a

$$q_k = \frac{1}{(x_k - x_0)\ldots(x_k - x_{k-1})(x_k - x_{k+1})\ldots(x_k - x_n)}$$

jelölést. Ezek az értékek csak az alappontoktól függnek, és függetlenek az alappontokbeli függvényértékektől is és x-től is. Ezek kiszámítása egyenként $2n$ flop, azaz összesen $2n(n+1) = 2n^2 + 2n$ flop. Ezek után az interpolációs polinom az

$$L_n(x) = w_{n+1}(x) \sum_{k=0}^{n} \frac{q_k f_k}{x - x_k}$$

alakban írható ($w_{n+1}(x)$ az alappontpolinom szokásos jelölése). Az alappontpolinomtól megszabadulhatunk, ha észrevesszük, hogy a konstans 1 függvény felírható

$$1 = w_{n+1}(x) \sum_{k=0}^{n} \frac{q_k}{x - x_k}$$

alakban, így

$$L_n(x) = \frac{L_n(x)}{1} = \frac{w_{n+1}(x) \sum_{k=0}^{n} \frac{q_k f_k}{x-x_k}}{w_{n+1}(x) \sum_{k=0}^{n} \frac{q_k}{x-x_k}} = \frac{\sum_{k=0}^{n} \frac{q_k f_k}{x-x_k}}{\sum_{k=0}^{n} \frac{q_k}{x-x_k}}.$$

Ezt az alakot baricentrikus interpolációs formulának hívjuk. Ha a q_k súlyokat már meghatároztuk, akkor egy helyettesítési érték számítása a fenti formulával $5n + 4$ flopba kerül.

6.6. Számítsuk ki először a baricentrikus súlyokat!

$$\begin{aligned} q_0(x) &= \frac{1}{(-1-2)(-1-3)(-1-4)} = -\frac{1}{60}, \\ q_1(x) &= \frac{1}{(2-(-1))(2-3)(2-4)} = \frac{1}{6}, \\ q_2(x) &= \frac{1}{(3-(-1))(3-2)(3-4)} = -\frac{1}{4}, \\ q_3(x) &= \frac{1}{(4-(-1))(4-2)(4-3)} = \frac{1}{10}. \end{aligned} \tag{10.5}$$

Ezek segítségével az interpolációs polinom az

$$L_3(x) = \frac{\dfrac{2\cdot(-1/60)}{x-(-1)} + \dfrac{4\cdot(1/6)}{x-2} + \dfrac{0\cdot(-1/4)}{x-3} + \dfrac{2\cdot(1/10)}{x-4}}{\dfrac{-1/60}{x-(-1)} + \dfrac{1/6}{x-2} + \dfrac{-1/4}{x-3} + \dfrac{1/10}{x-4}}$$

alakban adható meg, ami természetesen egyszerűsítés után a

$$L_3(x) = \frac{5}{6}x^3 - \frac{9}{2}x^2 + \frac{8}{3}x + 10$$

alakot ölti.

6.7. Természetesen ugyanazt a polinomot kapjuk mindegyik módszerrel. Mivel az egyes részfeladatok között csak egy-egy pont eltérés van, ezért célszerű a Newton-módszert használni.

$$a)\ x+3,\ \ b)\ \frac{1}{3}x^2 - \frac{1}{3}x + 4,\ \ c)\ -\frac{4}{3}x^3 - 11x^2 - \frac{77}{3}x + 20.$$

6.8. Ha $s \le n$, akkor az (x_k, x_k^s) $(k = 0, \dots, n)$ pontokra illesztett polinom éppen az $L_n(x) = x^s$ polinom lesz. Ez pont a keresett

$$x^s = \sum_{k=0}^{n} x_k^s l_k(x)$$

egyenlőséget jelenti. Speciális esetként $(s = 0)$ azt kapjuk, hogy az alappolinomok összege a konstans 1 függvényt adja.

6.9. Előállítjuk az interpolációs polinomot

$$L_2(x) = -\frac{1}{24}x^2 + \frac{3}{4}x - \frac{1}{3},$$

amibe 3-at helyettesítünk. Erre $37/24 = 1.5416$ adódik. Az $x = 3$ pontbeli hiba

$$|\log_2 3 - L_2(3)| \leq \frac{M_3}{6} w_3(3),$$

ahol $w_3(3)$ az alappontpolinom értéke 3-nál, azaz 5, M_3 pedig egy felső becslés $\log_2 x$ harmadik deriváltjára a $[2, 8]$ intervallumon, azaz pl. 0.37. Ebből a 0.3083-as felső becslés adódik a hibára.

6.10. Az interpolációs polinom előállítható pl. Newton- vagy Lagrange módszerével:

$$L_3(x) = \frac{1}{60}x^3 - \frac{1}{4}x^2 + \frac{37}{30}x.$$

Ennek a polinomnak az $x = 5$ pontbeli helyettesítési értéke 2. Ez lesz tehát a keresett közelítés.

6.11. Az osztóintervallumok hossza $h = 1/20$, $n = 10$ és az interpolálandó függvény 11. deriváltjára $(-11!x^{-12})$ a $11!2^{12}$ becslést adhatjuk. Innét a hibára 7.258×10^{-5} adódik (6.4. tétel).

6.12. Legyen az alappontok közti távolság h. A Newton-féle osztott differencia táblázat néhány elemét meghatározva a

$$c_k = [x_0, x_1, \ldots, x_k]f = \frac{\binom{k}{0}f_k - \binom{k}{1}f_{k-1} + \binom{k}{2}f_{k-2} + \ldots \mp \binom{k}{k}f_0}{h^k k!} = \frac{\sum_{i=0}^{k}\binom{k}{i}(-1)^i f_{k-i}}{h^k k!}$$

sejtésünk lehet.

Nyilvánvalóan $k = 0$ és $k = 1$ esetében igaz az állítás. Lássuk be, hogy ha l-re igaz, akkor $l + 1$ esetén is igaz!

$$\begin{aligned}
c_{l+1} &= [x_0, x_1, \ldots, x_{l+1}]f \\
&= \frac{[x_1, \ldots, x_{l+1}]f - [x_0, x_1, \ldots, x_l]f}{x_{l+1} - x_0} \\
&= \frac{[x_1, \ldots, x_{l+1}]f - [x_0, x_1, \ldots, x_l]f}{h(l+1)} \\
&= \frac{\sum_{i=-1}^{l-1}\binom{l}{i+1}(-1)^{i+1}f_{l-i} - \sum_{i=0}^{l}\binom{l}{i}(-1)^i f_{l-i}}{h^{l+1}(l+1)!} \\
&= \frac{\binom{l}{0}f_{l+1} + \sum_{i=0}^{l-1}\left(\binom{l}{i+1} + \binom{l}{i}\right)(-1)^{i+1}f_{l-i} - \binom{l}{l}(-1)^l f_0}{h^{l+1}(l+1)!} \\
&= \frac{\binom{l}{0}f_{l+1} + \sum_{i=0}^{l-1}\binom{l+1}{i+1}(-1)^{i+1}f_{l-i} - \binom{l}{l}(-1)^l f_0}{h^{l+1}(l+1)!} \\
&= \frac{\sum_{i=0}^{l+1}\binom{l+1}{i}(-1)^i f_{l+1-i}}{h^{l+1}(l+1)!}.
\end{aligned}$$

Ezzel igazoltuk a sejtésünket.

6.13. Mivel az alappontok egyforma távol vannak egymástól, így felhasználhatjuk a 6.12. feladat eredményét. Így tehát

$$\begin{aligned} f &= f_0 = 1, \\ [x_0, x_1]f &= \frac{f_1 - f_0}{h} = \frac{3-1}{2} = 1, \\ [x_0, x_1, x_2]f &= \frac{f_2 - 2f_1 + f_0}{2h^2} = \frac{8 - 2\cdot 3 + 1}{8} = \frac{3}{8}, \\ [x_0, x_1, x_2, x_3]f &= \frac{f_3 - 3f_2 + 3f_1 - f_0}{6h^3} = \frac{20 - 3\cdot 8 + 3\cdot 3 - 1}{48} = \frac{1}{12}, \end{aligned}$$

és a keresett interpolációs polinom

$$L_3(x) = 1 + (x-4) + \frac{3}{8}(x-4)(x-6) + \frac{1}{12}(x-4)(x-6)(x-8).$$

6.14. Az osztóintervallumok hossza $h = 1/20$, $n = 20$ és a függvény 21. deriváltjára $(20!x^{-20})$ a 20! becslést adhatjuk. Innét a hibára 1.3811×10^{-11} adódik. (6.4. tétel).

6.15. A feladatot a 6.4. tétel segítségével oldjuk meg. Az interpolációs hiba

$$|E_n(x)| \le \frac{M_{n+1}}{4(n+1)} h^{n+1} = \frac{M_{n+1}}{4(n+1)} \left(\frac{1}{n}\right)^{n+1}$$

becslését használjuk, ahol M_{n+1} egy becslés f $n+1$. deriváltjának abszolút értékére. Mivel

$$f^{(n+1)}(x) = \frac{(-1)^n n!}{x^{n+1}},$$

ezért $M_{n+1} = n!$ megfelelő választás. Így tehát

$$|E_n(x)| \le \frac{n!}{4(n+1)} \frac{1}{n^{n+1}},$$

ami x-től függetlenül nullához tart, ha n tart a végtelenbe. Azaz az interpolációs polinomok sorozata egyenletesen tart az $\ln x$ függvényhez.

6.16. Nyilvánvalóan $[x_0]f = 1/x_0$ és

$$[x_0, x_1]f = \frac{\frac{1}{x_1} - \frac{1}{x_0}}{x_1 - x_0} = \frac{-1}{x_0 x_1}.$$

Így az lehet a sejtésünk, hogy

$$[x_0, x_1, \ldots, x_n]f = \frac{(-1)^n}{x_0 x_1 \ldots x_n}.$$

Ezt teljes indukcióval igazolhatjuk. Az $n = 0$ és $n = 1$ választás esetén igaz az állítás. Tegyük fel, hogy $n-1$-re is igaz, azaz

$$[x_0, x_1, \ldots, x_{n-1}]f = \frac{(-1)^{n-1}}{x_0 x_1 \ldots x_{n-1}}.$$

Így tehát

$$[x_0, x_1, \ldots, x_n]f = \frac{[x_1, \ldots, x_n]f - [x_0, \ldots, x_{n-1}]f}{x_n - x_0} = \frac{\frac{(-1)^{n-1}}{x_1 \ldots x_n} - \frac{(-1)^{n-1}}{x_0 \ldots x_{n-1}}}{x_n - x_0} = \frac{(-1)^n}{x_0 x_1 \ldots x_n}.$$

Ezt akartuk megmutatni.

6.17. A 6.4. tételben szereplő hibabecslő formulát alkalmazzuk az $f(x) = x^{n+1}$ függvényre. Mivel $f^{(n+1)}(x) = (n+1)!$, ezért a hiba

$$x^{n+1} - L_n(x) = w_{n+1}(x).$$

A keresett $[x_0, \ldots, x_n]f$ osztott differencia az $L_n(x)$ interpolációs polinom főegyütthatója. Ez a fenti egyenlőségből meghatározható:

$$[x_0, \ldots, x_n]f = x_0 + x_1 + \ldots + x_n.$$

6.18. A program egy megvalósítása az alábbi linken található.

6.19. A MATLAB programmal számolva a 0.310268301038230 értéket kapjuk az integrál közelítésére. A „pontos” érték kb. 0.310268301723381.

6.20. A MATLAB programmal számolva 92.5 Hgmm-t kapunk $50°C$-nál a gőznyomásra.

6.21. A MATLAB programmal számolva rendre az alábbi közelítéseket kapjuk: 1.61, 1.72, 1.85, 2.08, 2.51.

Interpoláció Csebisev-alappontokon

6.22. Az alappontok a $\pm 1/\sqrt{2}$ pontok. Így az interpolációs polinom a lineáris

$$\sqrt{2}x \sin(\pi/(2\sqrt{2}))$$

polinom lesz. A hibabecsléshez a 6.6. tételt használhatjuk. Ezzel a hibára a

$$|E_1(x)| \leq \frac{M_2}{2!2^1} = \frac{\pi^2/4}{2!2^1} = 0.6169$$

becslést nyerhetjük.

6.23. A 6.6. tételt használjuk. Mivel $M_{n+1} = 1$ megfelelő választás, így az

$$|E_n(x)| \leq \frac{1}{(n+1)!2^n} \leq 10^{-6}$$

egyenlőtlenséget kell megoldanunk, pontosabban olyan n-t kellene mondani, amire teljesül az egyenlőtlenség. Ehhez pl. az

$$|E_n(x)| \leq \frac{1}{(n+1)!2^n} \leq \frac{1}{2^{2n}} \leq 10^{-6}$$

becslést használhatjuk. Logaritmus segítségével innét azt kapjuk, hogy a feltétel $n = 10$-től már teljesülni fog, azaz 10 Csebisev-alappont esetén már megfelelően kicsi lesz a hiba.

6.24. Az alappontok a $[0, 1]$ intervallumra transzformálva: $1/2$, $1/2 \pm \sqrt{3}/4$. MATLAB-bal számolva az integrál közelítésére 0.308368138117735 adódik.

6.25. A 6.6. tételt használhatjuk. A feladat állítása következik abból, ha megmutatjuk, hogy f Lipschitz-folytonos [-5,5]-ön, amihez elegendő megmutatni, hogy a deriváltja korlátos.

$$|f'(x)| = \left| \frac{-2x}{(1+x^2)^2} \right| \leq \frac{10}{(1+0^2)^2} = 10.$$

Ezt akartuk megmutatni.

Hermite-interpoláció

6.26. A megfelelő polinom az Hermite–Fejér-féle interpolációs polinom lesz. Ezt meghatározhatjuk pl. az osztott differenciák módszerével. A keresett polinom

$$q(x) = (x-1) + 2(x-1)^2 - 4(x-1)^2(x-2) + \frac{7}{4}(x-1)^2(x-2)^2 + \frac{3}{4}(x-1)^2(x-2)^2(x-3),$$

melyre $p(4) = 39$.

6.27. $f(0) = 0$, $f'(0) = 1$, $f(\pi/2) = 1$, $f'(\pi/2) = 0$. Ebből felírva az osztott differenciákat (6.8. tétel) a

$$H_3(x) = x - 2\,\frac{(-2+\pi)\,x^2}{\pi^2} + 4\,\frac{(-4+\pi)\,x^2\,(x - 1/2\,\pi)}{\pi^3}$$

polinomot nyerjük. Ebbe $\pi/4$-et helyettesítve 0.6963 adódik. A 6.9. tétel alapján ezen érték hibája kisebb, mint

$$|E_1(\pi/4)| \leq \frac{1}{24}\left(\frac{\pi}{4}\right)^4 \approx 0.01585.$$

Szakaszonkénti polinomiális interpoláció

6.28. Szakaszonként lineáris interpolációról van szó, így a 6.10. tételt használhatjuk. Mivel $f''(x) = 2\cos(2x)$, így ennek egy felső becslése 2. A hibára tehát a feladat feltétele szerint érvényes a

$$|p(x) - f(x)| \leq \frac{2}{8}\left(\frac{\pi}{n+1}\right)^2 = \frac{\pi^2}{4(n+1)^2} < 10^{-6}$$

becslés, amely $n > 1569.79$ esetén teljesül. Azaz n legalább 1570 legyen.

6.29. Az $(x_{k-1}, f(x_{k-1}))$, $(x_{k-1/2}, f(x_{k-1/2}))$, $(x_k, f(x_k))$ pontokra illesztett polinom interpolációs hibáját kell először kiszámolnunk ($x_{k-1/2}$ jelöli a k-adik intervallum felezőpontját). A 6.4. tétel miatt egy tetszőleges $\bar{x} \in [x_{k-1}, x_k]$ pontban az interpolációs hiba becslése

$$\left|\frac{f^{(3)}(\xi)}{3!}(\bar{x} - x_{k-1})(\bar{x} - x_{k-1/2})(\bar{x} - x_k)\right| \leq \frac{3}{8\cdot 6}\frac{2}{4}\left(\frac{1}{2n}\right)^2 = \frac{1}{128\cdot n^3} \leq 10^{-8},$$

ahol felhasználtuk, hogy $f'''(x) = 3x^{-5/2}/8$, melynek maximuma az $[1, 2]$ intervallumon $3/8$. A becslésből azt kapjuk, hogy ha n legalább 74, akkor teljesül a becslés.

6.30. Nyilvánvalóan elegendő a $-h, 0, h$ alappontokra vizsgálni az állítást. A 6.4. tétel miatt az interpolációs hiba alakja

$$E_2(x) = -\frac{f^{(3)}(\xi_x)}{3!}(x+h)x(x-h),$$

ahol ξ_x megfelelő szám a $(-h, h)$ intervallumból. Látható, hogy tulajdonképpen az $(x+h)x(x-h) = x(x^2-h^2)$ alappontpolinom értékére kell felső becslést adnunk. Egyszerű függvényvizsgálatot végrehajtva azt kapjuk, hogy az alappontpolinom az $x = \pm h/\sqrt{3}$ pontban veszi fel a legnagyobb abszolút értékű értékét, nevezetesen $\pm 2h^3/(3\sqrt{3})$-at. Így tehát az

$$|E_2(x)| = \frac{|f^{(3)}(\xi_x)|}{3!}|(x+h)x(x-h)| \leq \frac{\max_x\{|f'''(x)|\}2h^3}{6\cdot 3\sqrt{3}} = \frac{h^3}{9\sqrt{3}}\max_x\{|f'''(x)|\}$$

becslés adható. Ezt kellett igazolni.

6.31. A splinefüggvény deriváltjait a

$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{bmatrix}\begin{bmatrix} d_{-1} \\ d_0 \\ d_1 \end{bmatrix} = \frac{3}{h}\begin{bmatrix} f_0 - f_{-1} \\ f_1 - f_{-1} \\ f_1 - f_0 \end{bmatrix}$$

egyenletrendszerből kaphatjuk meg, ahonnét most csak a d_0 értéket kell meghatároznunk, hiszen ez adja meg az x_0-beli deriváltat. Az első sor és az utolsó sor (1/2)-szeresét kivonva a második sorból, azt kapjuk, hogy

$$3d_0 = \frac{3}{h}\left(f_1 - f_{-1} - \frac{1}{2}(f_0 - f_{-1} + f_1 - f_0)\right),$$

melyből egyszerűsítés után kapjuk, hogy

$$d_0 = \frac{f_1 - f_{-1}}{2h}.$$

Ezt kellett megmutatni.

6.32. Az alappontokbeli deriváltak értékére felírjuk az

$$\frac{1}{3}\begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{bmatrix}\begin{bmatrix} d_0 \\ d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix}$$

lineáris egyenletrendszert. Ennek megoldása $d_0 = -11/12$, $d_1 = -1/6$ és $d_2 = 7/12$. Ebből Hermite–Fejér-interpolációval kapjuk a [-1,0] intervallum polinomját:

$$s_1(x) = \frac{35}{12}x^3 + \frac{19}{4}x^2 - \frac{1}{6}x,$$

és a $[0, 1]$ intervallum polinomját

$$s_1(x) = \frac{19}{12}x^3 + \frac{11}{4}x^2 - \frac{1}{6}x.$$

Trigonometrikus interpoláció

6.33. Egy elsőfokú trigonometrikus polinom lesz megfelelő. Az együtthatókra tanult képletek alapján

$$t(x) = 1 + \frac{2}{\sqrt{3}}\sin x.$$

6.34. Mivel $n+1 = 4$ pontunk van, így $m = 2$ fokszámú kiegyensúlyozott trigonometrikus polinomot keresünk. Az együtthatók képleteit felhasználva kapjuk, hogy

$$T_2(x) = \frac{3}{4} - \frac{1}{2}\cos x + \sin x - \frac{5}{4}\cos(2x).$$

6.35. Legyenek az alappontok $x_k = 2k\pi/(n+1)$, $(k = 0, 1, \ldots, n)$, ahol $n+1 = 2m$ páros pozitív egész. Ekkor a komplex diszkrét Fourier-transzformációhoz a

$$c_j = \frac{1}{n+1} \sum_{k=0}^{n} f_k w^{-kj}, \ (j = -(m-1), \ldots, m)$$

együtthatókat kell kiszámolni, ahol w $(n+1)$-edik komplex egységgyök és f_k az interpolálandó f függvény x_k pontbeli értéke. Ez a feladat lényegében a $p(z) = \sum_{k=0}^{n} f_k z^k$ polinom helyettesítési értékeinek kiszámítását követeli meg a w^{-j} $(j = -(m-1), \ldots, m)$ számokra. Ez a számolás, ha már előre kiszámoltuk w hatványait $(n+1)^2$ komplex szorzást igényel.

Vezessük be a $p_{ps}(z) = f_0 + f_2 z + \ldots + f_{n-1} z^{m-1}$ és $p_{ptl}(z) = f_1 + f_3 z + \ldots + f_n z^{m-1}$ polinomokat. Ezekkel $p(z)$ a $p(z) = p_{ps}(z^2) + z p_{ptl}(z^2)$. Ha ennek a polinomnak kiszámítottuk a helyettesítési értékét egy w^{-j} helyen $(j = -(m-1), \ldots, 0)$

$$p(w^{-j}) = p_{ps}(w^{-2j}) + w^{-j} p_{ptl}(w^{-2j}),$$

akkor a $w^{-(j+m)}$-nél vett helyettesítési érték már szorzás nélkül számolható, ugyanis

$$p(w^{-(j+m)}) = p_{ps}(w^{-2j} w^{-2m}) + w^{-j} w^{-m} p_{ptl}(w^{-2j} w^{-2m}) = p_{ps}(w^{-2j}) - w^{-j} p_{ptl}(w^{-2j}),$$

ami miatt a két szereplő tagot összeadás helyett csak ki kell vonni egymásból. (Kihasználtuk, hogy $w^m = -1$ és $w^{n+1} = 1$.) Így összesen a két $(m-1)$-ed fokú polinom helyettesítési értékét kell kiszámolni, ami $2(n+1)^2/4$ szorzás, ill. az egyik polinomértéket szorozni kell még a megfelelő egységgyök hatványával. Ez további $(n+1)/2$ szorzás. Azaz a fent ismertetett technikával $(n+1)^2$ szorzás helyett csak $(n+1)^2/2 + (n+1)/2$ szorzást jelent.

6.36. A 6.35. feladat eredményét felhasználva a diszkrét Fourier-transzformáció a $p(z) = \sum_{k=0}^{n} f_k z^k$ polinom helyettesítési értékeinek kiszámítását követeli meg a w^{-j} $(j = -(m-1), \ldots, m)$ számokra.

Vezessük be a

$$\begin{aligned} p_0(z) &= f_0 + f_{t_1} z + \ldots + f_{(t_2-1)t_1} z^{t_2-1}, \\ p_1(z) &= f_1 + f_{t_1+1} z + \ldots + f_{(t_2-1)t_1+1} z^{t_2-1}, \\ &\vdots \\ p_{t_1-1}(z) &= f_{t_1-1} + f_{2t_1-1} z + \ldots + f_{t_1 t_2-1} z^{t_2-1} \end{aligned}$$

polinomokat, melyekkel $p(z)$ az alábbi alakban írható

$$p(z) = p_0(z^{t_1}) + z p_1(z^{t_1}) + z^2 p_2(z^{t_1}) + \ldots + z^{t_1-1} p_{t_1-1}(z^{t_1}).$$

Ha kiszámoltuk ennek a polinomnak az értékeit a w^{-j} ($j = -(m-1), \ldots, -(m-1)+t_2-1$) értékekre (ez $t_2(t_1t_2+t_1)$ szorzás), akkor a többi polinomérték ($t_1t_2-t_2$ darab) már polinom helyettesítési érték számolás nélkül számolható t_1 szorzás segítségével darabonként, ugyanis

$$p(w^{-(j+st_2)}) = p_0(w^{-jt_1}w^{-t_1t_2s}) + w^{-j}w^{-st_2}p_1(w^{-jt_1}w^{-t_1t_2s})$$
$$+w^{-2j}w^{-2st_2}p_2(w^{-jt_1}w^{-t_1t_2s}) + \ldots + w^{-jt_1}w^{-t_1st_2}p_{t_1-1}(w^{-jt_1}w^{-t_1t_2s})$$
$$= p_0(w^{-jt_1}) + w^{-j}w^{-st_2}p_1(w^{-jt_1}) + w^{-2j}w^{-2st_2}p_2(w^{-jt_1}) + \ldots + w^{-jt_1}w^{-t_1st_2}p_{t_1-1}(w^{-jt_1}),$$

ha $s = 1, \ldots, t_1 - 1$ valamilyen egész szám. Így a komplex szorzások száma ezzel a módszerrel $(t_1t_2)^2$ helyett csak $t_2(t_1t_2 + t_1) + t_1(t_1t_2 - t_2) = t_1t_2(t_1 + t_2)$ lesz. A két szorzásszám aránya $(t_1 + t_2)/(t_1t_2)$.

Approximáció polinomokkal és trigonometrikus polinomokkal

6.37. A

$$\begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 1 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 0 \end{bmatrix}$$

túlhatározott lineáris egyenletrendszer legkisebb négyzetek értelemben legjobban közelítő megoldását kell meghatározni. Ez a

$$\begin{bmatrix} 10 & 4 \\ 4 & 4 \end{bmatrix} \begin{bmatrix} a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$$

normálegyenlet megoldását követeli meg. Ennek megoldása $a_1 = -0.5$, $a_0 = 1.75$. Így tehát a legjobban közelítő egyenes az $y = -0.5x + 1.75$ lesz.

6.38. A megoldandó normálegyenlet

$$\begin{bmatrix} 82 & 28 & 10 \\ 28 & 10 & 4 \\ 10 & 4 & 4 \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix},$$

aminek megoldása $a_2 = -1/2$, $a_1 = 1$, $a_0 = 3/2$, azaz a legjobban közelítő legfeljebb másodfokú polinom $-x^2/2 + x + 3/2$.

6.39. Először meghatározunk olyan polinomokat, melyek a $-1, 0, 1, 3$ alappontrendszeren ortogonálisak. Ehhez ortonormálnunk kell az $\overline{\mathbf{x}}_0 = [1, 1, 1, 1]^T$ és $\overline{\mathbf{x}}_1 = [-1, 0, 1, 3]^T$ vektorokat:

$$\begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}, \quad \frac{\sqrt{35}}{70} \begin{bmatrix} -7 \\ -3 \\ 1 \\ 9 \end{bmatrix}.$$

Ezek alappontokon vett interpolációjával kapjuk az ortogonális polinomokat: $q_0 = 1/2$, $q_1 = (2/35)\sqrt{35}x - (3/70)\sqrt{35}$. Így a keresett legjobban közelítő elsőfokú polinom

$$p(x) = \overline{\mathbf{f}}^T q_0(\overline{\mathbf{x}}) q_0(x) + \overline{\mathbf{f}}^T q_1(\overline{\mathbf{x}}) q_1(x) = -0.4x + 1.8,$$

ahol $\overline{\mathbf{f}} = [2, 1, 3, 0]$ az adott pontok második koordinátáinak vektora.

6.40. A 6.39. feladat eredményét használhatjuk az eggyel magasabb fokú közelítő polinom meghatározásához. Ehhez a $q_2(x)$ ortonormált polinomot kell már csak meghatározni. Most az $\overline{\mathbf{x}}_0 = [1, 1, 1, 1]T$, $\overline{\mathbf{x}}_1 = [-1, 0, 1, 3]^T$ és $\overline{\mathbf{x}}_2 = [(-1)^2, 0^2, 1^2, 3^2]^T$ vektorokat kell ortonormálni (az első kettő már a hivatkozott feladatban ortonormálva lett) vektorokat:

$$\begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}, \quad \frac{\sqrt{35}}{70} \begin{bmatrix} -7 \\ -3 \\ 1 \\ 9 \end{bmatrix}, \quad \begin{bmatrix} (1/22)\sqrt{154} \\ (-2/77)\sqrt{154} \\ (-4/77)\sqrt{154} \\ (5/154)\sqrt{154} \end{bmatrix}.$$

Így $q_3(x) = (1/44)\sqrt{154}x^2 - (15/308)\sqrt{154}x - (2/77)\sqrt{154}$, és a legjobban közelítő polinom

$$\begin{aligned} p(x) &= \overline{\mathbf{f}}^T q_0(\overline{\mathbf{x}}) q_0(x) + \overline{\mathbf{f}}^T q_1(\overline{\mathbf{x}}) q_1(x) + \overline{\mathbf{f}}^T q_2(\overline{\mathbf{x}}) q_2(x) = \\ &= -0.4x + 1.8 + \overline{\mathbf{f}}^T q_2(\overline{\mathbf{x}}) q_2(x) = \frac{-7}{22}x^2 + \frac{31}{110}x + \frac{119}{55}. \end{aligned}$$

6.41. Az interpolációs polinom legfeljebb elsőfokú részletösszege lesz a legjobban közelítő polinom. Erre $T_1(x) = 5/2 - \cos x - \sqrt{3}\sin x$ adódik.

Numerikus deriválás és numerikus integrálás

Numerikus deriválás

7.1. Az approximáló kifejezést jelöljük $\Delta f(h)$-val! Alkalmazzunk Taylor-sofejtést az egyes tagokra!

$$f(x_0+h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(x_0) + \frac{h^3}{3!}f'''(x_0) + O(h^4)$$

$$f(x_0+2h) = f(x_0) + 2hf'(x_0) + \frac{(2h)^2}{2!}f''(x_0) + \frac{(2h)^3}{3!}f'''(x_0) + O(h^4)$$

Ekkor

$$\Delta f(h) = \frac{2hf'(x_0) - \frac{2}{3}h^3f'''(x_0) + O(h^4)}{2h} = f'(x_0) - \frac{f'''(x_0)}{3}h^2 + O(h^3).$$

Azaz $\Delta f(h)$ az első deriváltat másodrendben közelíti és hibája:

$$-\frac{f'''(x_0)}{3}h^2 + O(h^3).$$

7.4. Az approximáló kifejezést jelöljük $\Delta f(h)$-val! Továbbá vezessük be az alábbi jelöléseket:

$a = f(x_0 - 2h)$,

$b = -4f(x_0 - h)$,

$c = 6f(x_0)$,

$d = -4f(x_0 + h)$,

$e = f(x_0 + 2h)$!

Alkalmazzunk Taylor-sorfejtést az $a-e$ tagokra! Ekkor eredményeinket az alábbi táblázatban foglalhatjuk össze:

	$f(x_0)$	$f'(x_0)$	$f''(x_0)$	$f'''(x_0)$	$f^{(4)}(x_0)$	$f^{(5)}(x_0)$	$f^{(6)}(x_0)$	$f^{(7)}(x_0)$
a	1	$-2h$	$2h^2$	$-\frac{4}{3}h^3$	$2/3h^4$	$-4/15h^5$	$4/45h^6$	$-8/315h^7$
b	-4	$4h$	$-2h^2$	$2/3h^3$	$-1/6h^4$	$1/30h^5$	$-1/180h^6$	$1/1260h^7$
c	6	0	0	0	0	0	0	0
d	-4	$-4h$	$-2h^2$	$-2/3h^3$	$-1/6h^4$	$-1/30h^5$	$-1/180h^6$	$-1/1260h^7$
e	1	$2h$	$2h^2$	$4/3h^3$	$2/3h^4$	$4/15h^5$	$4/45h^6$	$8/315h^7$

10.8. táblázat. A Taylor-sorfejtés során számolt egyes együtthatók.

A fenti táblázat megfelelő oszlopainak összegzésével az approximáló kifejezés az alábbi alakot ölti:

$$\Delta f(h) = \frac{h^4 f^{(4)}(x_0) + \frac{1}{6}h^6 f^{(6)}(x_0) + O(h^8)}{h^4} = f^{(4)}(x_0) + \frac{1}{6}h^2 f^{(6)}(x_0) + O(h^4).$$

Azaz $\Delta f(h)$ a negyedik deriváltat másodrendben közelíti és hibája:

$$\frac{1}{6}h^2 f^{(6)}(x_0) + O(h^4).$$

7.5. Vezessük be az alábbi jelöléseket:

$$f(x_0) = f_0,\ f(x_0+h) = f_1,\ f(x_0-h) = f_{-1}!$$

Tegyük fel, hogy az $f_{-1},\ f_0,\ f_1$ értékeket megváltoztatjuk ϵ-nál kisebb értékekkel. Azaz legyen

$$\tilde{f}_{-1} = f_{-1} + \epsilon_{-1},\ \tilde{f}_0 = f_0 + \epsilon_0,\ \tilde{f}_1 = f_1 + \epsilon_1,$$

ahol $|\epsilon_{-1}|,\ |\epsilon_0|,\ |\epsilon_{-1}| \leq \epsilon$. Ekkor

$$\frac{\tilde{f}_1 - \tilde{f}_{-1}}{2h} = \frac{f_1 - f_{-1}}{2h} + \frac{\epsilon_1 - \epsilon_{-1}}{2h} = f'(x_0) + \frac{h^2}{6}f'''(\xi) + \frac{\epsilon_1 - \epsilon_{-1}}{2h}.$$

Azaz

$$\left|\frac{\tilde{f}_1 - \tilde{f}_{-1}}{2h} - f'(x_0)\right| \leq \frac{h^2}{6}M_3 + \frac{2\epsilon}{2h},$$

ahol $M_3 = \sup|f'''(x)|$. A felső becslést adó függvény nem más, mint

$$g(h) = \frac{h^2}{6}M_3 + \frac{\epsilon}{h}.$$

Ez azt mutatja, hogy akkor lesz kicsi a hiba, ha h se nem túl nagy, se nem túl kicsi. Egy közelítő optimális értéket úgy nyerhetünk, hogy a felső becslést minimalizáljuk h-ban. Azaz

$$g'(h) = 0 \Leftrightarrow \frac{h}{3}M_3 - \frac{\epsilon}{h^2} = 0 \Leftrightarrow h^3 = \frac{3\epsilon}{M_3}.$$

Ekkor az ϵ-hibával terhelt középponti szabály optimális lépéshossza:

$$h_{\text{opt}} = \sqrt[3]{\frac{3\epsilon}{M_3}}.$$

7.8. Fejtsük Taylor-sorba az egyes tagokat!

$$Af(x_0 - h) = Af(x_0) - Ahf'(x_0) + \frac{h^2}{2}f''(x_0) + O(h^3).$$

$$Cf(x_0 + h) = Af(x_0) + Ahf'(x_0) + \frac{h^2}{2}f''(x_0) + O(h^3).$$

Ekkor

$$(A + B + C)f(x_0) + (C - A)f'(x_0) + (A\frac{h^2}{2} + C\frac{h^2}{2})f''(x_0) + O(h^3).$$

Azaz ezen tagokkal a második derivált approximálásának feltételei az egyes együtthatókra nézve az alábbi:

$$\begin{cases} A + B + C = 0, \\ -A + C = 0, \\ A + C = \dfrac{2}{h^2}. \end{cases} \Rightarrow A = \frac{1}{h^2},\ B = -\frac{2}{h^2},\ C = \frac{1}{h^2}.$$

7.9. A feladat által megadott lépésközök mellett a megírt program (lásd forráskódját a feladat végén) eredményeit az alábbi táblázat szemlélteti:

h	hiba
10^{-2}	$3.9951996795 \cdot 10^{-6}$
10^{-3}	$3.9970202259 \cdot 10^{-8}$
10^{-4}	$4.4626258244 \cdot 10^{-10}$
10^{-5}	$4.0478514135 \cdot 10^{-7}$
10^{-6}	$3.1236571060 \cdot 10^{-5}$

10.9. táblázat. Adott h lépésköz melletti hibaérték.

A táblázat jól mutatja, hogy a hiba értéke különböző h mellett ingadozik. Mi lehet ennek az oka?

Ennek a kérdésnek a megválaszolásához a 7.7. feladat eredményét kell felhasználni. Nevezetesen azt, hogy a második deriváltat másodrendben közelítő centrális differencia séma optimális lépéshossza ϵ pontosságú adatok esetén:

$$h_{\text{opt}} = \sqrt[4]{\frac{48\epsilon}{M_4}},$$

ahol $M_4 = \sup |f^{(4)}(x)|$.

A MATLAB program $\epsilon = 10^{-16}$ pontossággal közelít tetszőleges értéket. Továbbá a feladatunkban $M_4 = \sup |\sin^{(4)}(x)| = 1$. Ekkor a fenti eredményben ezeket az értékeket helyettesítve kapjuk az optimális paraméter értéket, mely:

$$h_{\text{opt}} = 2.6321480259 \cdot 10^{-4}$$

Azaz $h = 2.63215 \cdot 10^{-4}$-es érték mellett lesz a két érték különbsége a lehető legkisebb. Így nem meglepő módon a kisebb lépésközű értéktől haladva a hiba h_{opt} értékig csökken, míg onnantól kezdve folyamatosan nő a hiba.

A szükséges közelítő másodrendű centrális differencia programja, mely adott lépésköz mellett a pontos és közelítő érték abszolút hibáját mutatja.

```
function [hiba]=szinusznumder(h)
f_pontos=-sin(0.5);
f_centralis_differencia=(sin(0.5+h)-2*sin(0.5)+sin(0.5-h))/(h^2);
hiba=abs(f_pontos-f_centralis_differencia);
```

Numerikus integrálás

7.11. A feladatban szereplő formulák a Newton–Cotes-típusú kvadratúrák speciális esetei:

$$\int_a^b f(x)\,\mathrm{d}x = \underbrace{(b-a)f\Big(\frac{a+b}{2}\Big)}_{I_E(f)} + \frac{(b-a)^3}{24} f''(\eta),$$

ahol $\eta \in [a,b]$ és $I_E(f)$ az érintőformula.

$$\int_a^b f(x)\,\mathrm{d}x = \underbrace{(b-a)\Big(\frac{f(a)+f(b)}{2}\Big)}_{I_{\text{Tr}}(f)} - \frac{(b-a)^3}{12} f''(\eta),$$

ahol $\eta \in [a,b]$ és $I_{\mathrm{Tr}}(f)$ a trapézformula.

$$\int_a^b f(x)\,\mathrm{d}x = \underbrace{\frac{(b-a)}{6}\Big(f(a)+4f\Big(\frac{a+b}{2}\Big)+f(b)\Big)}_{I_{\mathrm{Simp}}(f)} + \frac{(b-a)^5}{180}f^{(4)}(\eta),$$

ahol $\eta \in [a,b]$ és $I_{\mathrm{Simp}}(f)$ a Simpson-formula. Ezeket a formulákat alkalmazva az alábbi közelítéseket nyerjük:

$$I_E(f) = (1-0)f(1/2) = 1/4.$$

$$I_{\mathrm{Tr}}(f) = (1-0)\cdot\frac{f(0)+f(1)}{2} = 1/2.$$

$$I_{\mathrm{Simp}}(f) = \frac{(1-0)}{6}\cdot\Big(f(0)+4f\Big(\frac{0+1}{2}\Big)+f(1)\Big) = 1/3.$$

Jelöljük a feladatban szereplő integrált $I(f)$-ként! Ekkor az adott formulára vonatkozó hiba nem lesz más, mint

$$|I(f)-I_E(f)| \le \frac{(1-0)^3}{24}M_2 = \frac{1}{24}\cdot 2 = \frac{1}{12},$$

$$|I(f)-I_{\mathrm{Tr}}(f)| \le \frac{(1-0)^3}{12}M_2 = \frac{1}{12}\cdot 2 = \frac{1}{6},$$

$$|I(f)-I_{\mathrm{Simp}}(f)| \le \frac{(1-0)^5}{180}M_4 = 0,$$

ahol $M_2 = \sup\limits_{x\in[0,1]}|f''(x)| = \sup\limits_{x\in[0,1]}|2| = 2$ és $M_4 = \sup\limits_{x\in[0,1]}|f^{(4)}(x)| = \sup\limits_{0\in[0,1]}|2| = 0.$

Azaz az érintő-, trapéz- és Simpson-formula az integrál pontos értékét rendre $1/12$, $1/6$ és 0 hibával közelíti.

7.12. A feladat megoldásához szükséges képlet, ahol az $[a,b]$ intervallumot n egyenlő részre osztottuk:

$$\int_a^b f(x)\,\mathrm{d}x = \underbrace{\frac{(b-a)}{2n}\Big(f(a)+f(b)+2\sum_{i=1}^{n-1}f(x_i)\Big)}_{I_{n,\mathrm{Tr}}(f)} - \frac{(b-a)^3}{12n^2}f''(\eta),$$

ahol $\eta \in [a,b]$ és $I_{n,\mathrm{Tr}}(f)$ az összetett trapézformula. A feladat kiírása alapján a három részre osztott [0,1] intervallum a $0,\ 1/3,\ 2/3,\ 1$ osztópontokat jelöli ki. Ekkor az összetett trapézformula az alábbi közelítő integrálértéket adja:

$$I_{3,\mathrm{Tr}} = \frac{1-0}{2\cdot 3}\Big[f(0)+f(1)+2\Big(f\Big(\frac{1}{3}\Big)+f\Big(\frac{2}{3}\Big)\Big)\Big] = \frac{1}{6}\Big(4+\frac{89}{130}\Big) = \frac{203}{260}.$$

A hiba meghatározáshoz meg kell adnunk a második derivált szuprémumát a $[0,1]$ intervallumon.

$$M_2 = \sup_{x\in[0,1]} |f''(x)| = \sup_{x\in[0,1]} |-2(1+x^2)^{-2}+8x^2(1+x^2)^{-1}| = 5.$$

Ekkor

$$|I(f)-I_{n,\mathrm{Tr}}| \le \frac{(1-0)^2}{12\cdot 3^2}M_2 = \frac{1}{108}\cdot 5 = \frac{5}{108}.$$

Azaz a három részre történő osztás esetén az összetett trapézformula az integrál pontos értékét 1/108 hibával közelíti.

7.13. A 7.12. feladat megoldása során meghatároztuk a hibaképletben szereplő $M_2 = 5$ értéket. Ekkor a feladatunk az intervallumszámot megadó n paraméter kiszámítása lesz, feltéve ha az összetett trapézformula eredménye 10^{-5} pontossággal közelíti az integrál pontos értékét.

$$|I(f)-I_{n,\mathrm{Tr}}(f)| \le \frac{1}{12n^2}\cdot 5 < 10^{-5}.$$

Ebből az $n \approx 204.124$ értéket nyerjük, azaz a szükséges intervallumok száma legalább 125.

7.16. A feladatra megírt ossztrapez.m fájl forráskódja az alábbi:

```
function ossztrapez(a,b,n,fv)

format long
h=(b-a)/n;

  fprintf('\n');
  disp('A feladat megoldása összetett trapézformulával.')

  x=[a:h:b];
  y=eval(fv);
  ((b-a)/(2*n))*(y(1)+2*sum(y(2:1:n))+y(n+1))
```

Tesztelve a programot a 7.15. feladatra a MATLAB által beépített `trapz` függvény által kiszámított 4 értéket adja vissza. Ennek beírása a MATLAB-ban és eredménye:

```
>> ossztrapez(-2,2,23,'x.^5-3*x.^3+2*x+1')

A feladat megoldása összetett trapézformulával.

ans =

     4
```

7.18. A feladat eredményét realizáló osszformulak.m MATLAB fájl forráskódja az alábbi:

```
function osszformulak(a,b,n,fv,method)

format long
h=(b-a)/n;

switch method
  case {'Erinto'}
    fprintf('\n');
    disp('A feladat megoldása összetett érint■formulával.')
       x=[a:h/2:b];
       y=eval(fv);
       ((b-a)/n)*sum(y(2:2:2*n))

  case {'Trapez'}
    fprintf('\n');
    disp('A feladat megoldása összetett trapézformulával.')
       x=[a:h:b];
       y=eval(fv);
       ((b-a)/(2*n))*(y(1)+2*sum(y(2:1:n))+y(n+1))

  case {'Simpson'}
    fprintf('\n');
    disp('A feladat megoldása összetett Simpson-formulával.')
       x=[a:h:b];
       y=eval(fv);
       ((b-a)/(3*n))*(y(1)+4*sum(y(2:2:n))+2*sum(y(3:2:n-1))+y(n+1))

otherwise
        fprintf('\n');
        disp('Nem megfelel■ módszer.')
end
```

A program futtatása például az

$$\int_0^1 e^x \,\mathrm{d}x = e^1 - 1 \approx 1.718281828459046$$

integrál esetén az alábbi értékeket adja vissza 100 intervallumra történő osztás esetén:

```
>> osszformulak(0,1,100,'exp(x)','Erinto')

A feladat megoldása összetett érint■formulával.

ans =

   1.718274668972308
>> osszformulak(0,1,100,'exp(x)','Trapez')

A feladat megoldása összetett trapézformulával.

ans =

   1.718296147450418
>> osszformulak(0,1,100,'exp(x)','Simpson')

A feladat megoldása összetett Simpson-formulával.

ans =

   1.718281828554504
>> osszformulak(0,1,100,'exp(x)','Butaság')

Nem megfelel■ módszer.
```

7.20. A 7.19. feladat megoldása alapján az ún. Boole-formulát alkalmazzuk a konkrét feladatra:

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{(b-a)}{90}\Big(7f(a) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(b)\Big),$$

ahol $x_i = a + i(b-a)/n,\ i = 1, \ldots, 3$. Azaz

$$\int_0^2 2 + \cos(2\sqrt{x})\,\mathrm{d}x \approx \frac{(2-0)}{90}\Big(7f(0) + 32f(0.5) + 12f(1) + 32f(1.5) + 7f(2)\Big) =$$

$$\frac{1}{45}\Big(21{+}32(2{+}\cos(\sqrt{2})){+}12(2{+}\cos(\sqrt{2})){+}32(2{+}\cos(\sqrt{6})){+}7(2{+}\cos(2\sqrt{2}))\Big) \approx 3.459998.$$

Azaz a zárt $N^{4,k}$ Newton–Cotes-formula segítségével meghatározott integrálközelítő értéke 3.459998.

7.21. A harmadfokú Legendre-polinom zérushelyei:

$$-\sqrt{\frac{3}{5}},\ 0,\ \sqrt{\frac{3}{5}},$$

melyek az alappontok lesznek. A Legendre-polinom súlyfüggvénye a konstans 1 függvény. A kvadratúra előállításához szükséges együtthatókat a megfelelő Lagrange-polinomok integráljaként nyerjük. Tekintsük az első alappontot:

$$a_0 = \int_{-1}^{1} \frac{\Big(x-0\Big)\Big(x-\sqrt{\frac{3}{5}}\Big)}{\Big(-\sqrt{\frac{3}{5}}-0\Big)\Big(-\sqrt{\frac{3}{5}}-\sqrt{\frac{3}{5}}\Big)}\,\mathrm{d}x = \int_{-1}^{1} \frac{x^2 - x\sqrt{\frac{3}{5}}}{\frac{6}{5}}\,\mathrm{d}x = \frac{5}{9}.$$

$$a_1 = \int_{-1}^{1} \frac{\Big(x+\sqrt{\frac{3}{5}}\Big)\Big(x-\sqrt{\frac{3}{5}}\Big)}{\Big(0+\sqrt{\frac{3}{5}}\Big)\Big(0-\sqrt{\frac{3}{5}}\Big)}\,\mathrm{d}x = \int_{-1}^{1} \frac{x^2 - \frac{3}{5}}{-\frac{3}{5}}\,\mathrm{d}x = \frac{8}{9}.$$

$$a_2 = \int_{-1}^{1} \frac{\Big(x+\sqrt{\frac{3}{5}}\Big)\Big(x-0\Big)}{\Big(\sqrt{\frac{3}{5}}+\sqrt{\frac{3}{5}}\Big)\Big(\sqrt{\frac{3}{5}}-0\Big)}\,\mathrm{d}x = \int_{-1}^{1} \frac{x^2 + x\sqrt{\frac{3}{5}}}{\frac{6}{5}}\,\mathrm{d}x = \frac{5}{9}.$$

Ekkor a Gauss–Legendre-kvadratúra képlete nem más, mint

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx \frac{5}{9} f\Big(-\sqrt{3/5}\Big) + \frac{8}{9} f\Big(0\Big) + \frac{5}{9} f\Big(\sqrt{3/5}\Big),$$

amely pontos lesz minden legalább ötödfokú polinomra.

7.23. Ismeretes, hogy $n+1$ alappont esetén a kvadratúra $2n+1$-edfokú polinomokra pontos, ezért elegendő a feladat megoldásához $n = 2$ értéket megválasztani. Ekkor a $T_3(x) = 4x^3 - 3x$ Csebisev-polinom gyökei:

$$-\frac{\sqrt{3}}{2},\ 0,\ \frac{\sqrt{3}}{2}.$$

A formula ötödfokú polinomokra pontos, így

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,\mathrm{d}x \approx \frac{\pi}{3}\Big(f(-\sqrt{3}/2) + f(0) + f(\sqrt{3}/2)\Big),$$

azaz

$$\int_{-1}^{1} \frac{x^4}{\sqrt{1-x^2}}\,\mathrm{d}x = \frac{\pi}{3}\Big((-\sqrt{3}/2)^4 + (\sqrt{3}/2)^4\Big) = \frac{3}{8}\pi.$$

Azaz az integrál pontos értéke $\dfrac{3\pi}{8}$.

7.24. Ahhoz, hogy a formula legfeljebb másodfokú polinomokra legyen pontos teljesülnie kell, hogy $f(x) = 1$, x és x^2 polinomokra pontos. Ezeket a feltételeket behelyettesítve nyerjük, hogy:

$$\int_0^4 1\,\mathrm{d}x = c_1 + c_2 + c_3 = 4,$$

$$\int_0^4 x\,\mathrm{d}x = c_1 + 2c_2 + 4c_3 = 8,$$

$$\int_0^4 x^2\,\mathrm{d}x = c_1 + 4c_2 + 16c_3 = 64/3.$$

A lineáris algebrai egyenletrendszert megoldva megkapjuk a kérdéses együtthatókat. Nevezetesen,

$$c_1 = \frac{16}{9},\ c_2 = \frac{4}{3},\ c_3 = \frac{8}{9}.$$

Ekkor a

$$\frac{16}{9}f(1) + \frac{4}{3}f(2) + \frac{8}{9}f(4)$$

közelítő integrálás minden legfeljebb másodfokú polinomra pontos.

7.25. A hibaszámításhoz szükséges a pontos integrál értéke: 1.640533. A feladatban megadott intervallumszám esetén MATLAB-ban a Crank–Nicolson módszer értékeit a `trapz` parancs segítségével számolhatjuk ki. Ezek eredményeit az alábbi táblázatban foglalhatjuk össze:

intervallum száma	h	`trapz` értéke	hiba (%-ban)
1	0.8	0.1728	89.5
2	0.4	1.0688	34.9
4	0.2	1.4848	9.5

10.10. táblázat. A Crank–Nicolson módszer értékei, hibája adott lépésköz mellett.

A Richardson-extrapoláció két kevésbé pontos megoldásból egy pontosabbat állít össze. Nevezetesen, ha egy módszer másodrendű (például a Crank–Nicolson módszer), akkor h_1 és h_2 lépésközű rácshálón való $R(h_1)$ és $R(h_2)$ numerikus értékeket az

$$R = R(h_2) + \frac{1}{\left(\frac{h_1}{h_2}\right)^2 - 1}\Big[R(h_2) - R(h_1)\Big]$$

módon kombinálva $O(h^4)$ nagyságrendű módszert kapunk. A formula intervallumfelezés esetén ($h_2 = 0.5h_1$) az alábbi módon egyszerűsödik:

$$R = \frac{4}{3}R(h_2) - \frac{1}{3}R(h_1).$$

Ekkor a Richardson-extrapoláció értékeit az alábbi módon számíthatjuk:

$$R = \frac{4}{3}1.0688 - \frac{1}{3}0.1728 = 1.36747.$$

Ekkor a hiba 16.6%-os.

$$R = \frac{4}{3}1.4848 - \frac{1}{3}1.0688 = 1.62347.$$

Ekkor a hiba 1%-os.

A hibaeredményekből jól látható, hogy a Richardson-extrapoláció a másodrendű Crank–Nicolson-módszer megfelelő súlyozásával negyedrendű módszert állít elő.

7.26. A Romberg-módszer nem jelent mást, mint Richardson-extrapolációk végrehajtását az alsóbb szinteken. Azaz a 7.25. feladat gondolatmenetéhez hasonlóan számíthatjuk ki a negyedrendben közelítő numerikus értékeket.

Ha az adott módszer negyedrendű és felezzük az intervallumot ($h_2 = 0.5h_1$), akkor a Romberg-módszer az alábbi alakot ölti:

$$R = \frac{16}{15}R(h_2) - \frac{1}{15}R(h_1).$$

Hatodrendű módszer esetén:

$$R = \frac{64}{63}R(h_2) - \frac{1}{63}R(h_1).$$

Ekkor a Romberg-módszerrel számított közelítő integrál értékei az alábbiak:

$O(h^2)$	$O(h^4)$	$O(h^6)$	$O(h^8)$
0.172800			
	1.367467		
1.068800		1.640533	
	1.623467		1.640533
1.484800		1.640533	
	1.639467		
1.600800			

A közönséges differenciálegyenletek kezdetiérték-feladatainak numerikus módszerei

Egylépéses módszerek

8.1. A konzisztencia definícióját és Taylor-sorfejtést alkalmazva az alábbi eredményeket kapjuk (a többi eredmény az Útmutatások, végeredmények fejezetben megtalálható):

(a) $\psi_n = (y(t_{n-1}) - y(t_n))/h - f(t_n, y(t_n)) = (y(t_n) + hy'(t_n) + h^2/2y''(t_n) + O(h^3) - y(t_n))/h - y'(t_n) = 1/2hy''(t_n) + O(h^2)$, azaz az explicit Euler módszer első rendben konzisztens.

(b) $\psi_n = (y(t_{n-1}) - y(t_n))/h - f(t_{n+1}, y(t_{n+1})) = (y(t_{n+1}) - y(t_{n+1}) + hy'(t_{n+1}) - h^2/2y''(t_{n+1}) + O(h^3))/h - y'(t_{n+1}) = -1/2hy''(t_{n+1}) + O(h^2)$, azaz az implicit Euler módszer első rendben konzisztens.

8.2. Alkalmazzuk a feladatban szereplő módszereket $h = 1/2$-es lépésköz esetén! Az $[1, 2]$ intervallumon így minden egyes módszer esetén 2 rácspont értékét kell kiszámítanunk. (A kezdeti feltétel miatt a kiinduló x_0 érték adott.)

(a) Az $y_{n+1} = y_n + hf(t_n, y_n)$ képletet alkalmazva az alábbi értékeket kapjuk:
$y_0 = 1$,
$y_1 = 1 + 1/2 \cdot f(1, 1) = 1 + 1/2 \cdot 2 = 2$,
$y_2 = 2 + 1/2 \cdot f(3/2, 2) = 2 + 1/2 \cdot 4 \cdot 2/3 = 10/3 \approx 3.33333333$.

Azaz a feladat közelítő megoldásának értéke a $t = 2$ pontban 3.33333333.

(b) Az $y_{n+1} = y_n + hf(t_{n+1}, y_{n+1})$ képletet alkalmazva az alábbi értékeket kapjuk:
$y_0 = 1$,
$y_1 = 1 + 1/2 \cdot f(3/2, x_1)$, melyből y_1-et kifejezve kapjuk:

$$y_1 = \frac{1}{1 - \frac{2 \cdot 1/2}{3/2}} = 3,$$

$y_2 = 3 + 1/2 \cdot f(2, y_2)$, melyből y_2-t kifejezve kapjuk:

$$y_2 = \frac{3}{1 - \frac{2 \cdot 1/2}{2}} = 6.$$

Azaz a feladat közelítő megoldásának értéke a $t = 2$ pontban 6.

(c) Az $y_{n+1} = y_n + h/2(f(t_n, y_n) + f(t_{n+1}, y_{n+1}))$ képletet alkalmazva az alábbi értékeket kapjuk:
$y_0 = 1$
$y_1 = 1 + 1/4 \cdot (f(1, 1) + f(3/2, y_1))$, melyből y_1-et kifejezve kapjuk:

$$y_1 = \frac{1 + \frac{1/2 \cdot 1}{1}}{1 - \frac{1/2}{3/2}} = \frac{9}{4},$$

$y_2 = 9/4 + 1/4 \cdot (f(3/2, 9/4) + f(2, y_2))$, melyből y_2-t kifejezve kapjuk:

$$y_2 = \frac{9/4 + \frac{1/2 \cdot 9/4}{3/2}}{1 - \frac{1/2}{2}} = 4.$$

Azaz a feladat közelítő megoldásának értéke a $t = 2$ pontban 4.

(d) Az $y_{n+1} = y_n + hf(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n))$ képletet alkalmazva az alábbi értékeket kapjuk:
$y_0 = 1$,
$y_1 = 1 + 1/2 \cdot f(1 + 1/4, 1 + 1/4 \cdot f(1, 1)) = 11/5$,
$y_2 = 11/5 + 1/2 \cdot f(3/2 + 1/4, 11/5 + 1/4 \cdot f(3/2, 11/5)) = 407/105 \approx 3.87619047$.
Azaz a feladat közelítő megoldásának értéke a $t = 2$ pontban $3,87619047$.

(e) Az $y_{n+1} = y_n + h(1/2 f(t_n, y_n) + 1/2 f(t_n + h, y_n + hf(t_n, y_n)))$ képletet alkalmazva az alábbi értékeket kapjuk:
$y_0 = 1$,
$y_1 = 1 + 1/2(1/2 f(1, 1) + 1/2 f(3/2, 1 + 1/2 f(1, 1))) = 13/6$,
$y_2 = 13/6 + 1/2(1/2 f(3/2, 13/6) + 1/2 f(2, 13/6 + 1/2 f(3/2, 13/6))) = 91/24$.
Azaz a feladat közelítő megoldásának értéke a $t = 2$ pontban $3,79166666$.

A többi lépésköz esetén a módszerek hasonlóan alkalmazhatóak. Ezek eredményei az Útmutatások, végeredmények fejezetben megtalálhatóak.

8.8. A hiba méréséhez a numerikus és pontos megoldás értékeit kell meghatároznunk. Az előbbi meghatározására futtassuk a 8.6. feladatban megírt programjainkat a 8.3. feladatra a $h = 1/2, 1/4, 1/8$ és $h = 1/16$ lépésközökkel. A kezdetiérték-feladat pontos megoldásának értéke a $t = 2$ pontban 4. Ekkor a hiba a két érték különbségének abszolút értékeként számítható ki. Az alábbi táblázatban ezek eredményét láthatjuk:

hiba	EE	IE	CN	JE	EH
h=1/2	0.66666666	2.00000000	0.00000000	0.12380952	0.20833333
h=1/4	0.39999999	0.66666666	0.00000000	0.03820081	0.06957695
h=1/8	0.22222222	0.28571428	0.00000000	0.01059999	0.02020452
h=1/16	0.11764705	0.13333333	0.00000000	0.00278829	0.00544302

10.11. táblázat. Hibaértékek adott lépésköz és módszer mellett.

A táblázatból jól látható, hogy az explicit és implicit Euler módszerek esetében a hiba a lépésköz felezésével feleződik, míg a javított Euler és Euler–Heun-módszerek esetében negyedelődik. Ennek magyarázata abban áll, hogy előbbi kettő elsőrendben, míg utóbbi kettő módszer másodrendben konvergens. A Crank–Nicolson-séma (másodrendű módszer) a legelső lépésben a feladat pontos értékét adja.

8.10. A 8.2. feladat módszerei az implicit Euler- és Crank–Nicolson-módszerek kivételével explicitek. Azaz az explicit Euler, javított Euler és Euler–Heun-módszerek Butcher-táblázatát kell meghatároznunk. Az explicit Runge–Kutta általános s-lépcsős módszerek általános alakjának birtokában meghatározzuk a szükséges paramétereket.

(a) Tekintsük az explicit Euler-módszer képletét. A módszer egylépcsős, így a c_1 paraméter értékét kell meghatároznunk.

$$y_{n+1} = y_n + h \cdot \underbrace{1}_{c_1} \cdot \underbrace{f(t_n, y_n)}_{k_1}$$

Azaz a módszert felírhatjuk $y_{n+1} = y_n + h \cdot 1 \cdot k_1$ alakban. Ekkor a Butcher-táblázat:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

(b) Tekintsük a javított Euler-módszer képletét. Meghatározzuk a szükséges paramétereket.

$$y_{n+1} = y_n + \underbrace{1}_{c_2} \cdot h \underbrace{f(t_n + \underbrace{\frac{1}{2}}_{a_2} h, y_n + \underbrace{\frac{1}{2}}_{b_{21}} h \underbrace{f(t_n, y_n)}_{k_1})}_{k_2}$$

Azaz a kétlépcsős módszert felírhatjuk $y_{n+1} = y_n + h \cdot 1 \cdot k_2$ alakban. Ekkor a Butcher-táblázat:

0	0	0
1/2	1/2	0
	0	1

(c) Tekintsük az Euler–Heun-módszer képletét. Meghatározzuk a szükséges paramétereket.

$$y_{n+1} = y_n + h(\underbrace{\frac{1}{2}}_{c_1}\underbrace{f(t_n, y_n)}_{k_1} + \underbrace{\frac{1}{2}}_{c_2}\underbrace{f(t_n + \underbrace{1}_{a_2}h, y_n + \underbrace{1}_{b_{21}}h\underbrace{f(t_n, y_n)}_{k_1})}_{k_2})$$

Azaz a kétlépcsős módszert felírhatjuk $y_{n+1} = y_n + h(1/2k_1 + 1/2k_2)$ alakban. Ekkor a Butcher-táblázat:

0	0	0
1	1	0
	1/2	1/2

8.11. Tekintsük a klasszikus negyedrendű Runge–Kuta-módszer Butcher-táblázatát.

0	0	0	0	0
1/2	1/2	0	0	0
1/2	0	1/2	0	0
1	0	0	1	0
	1/6	1/3	1/3	1/6

A Butcher-táblázatból könnyen leolvasható, hogy a módszer négylépcsős, továbbá a szükséges együtthatókat is könnyen felírhatjuk. Nevezetesen:

$k_1 = f(t_n, y_n)$
$k_2 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1)$
$k_3 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_2)$
$k_4 = f(t_n + h, y_n + k_3)$.

Azaz a megadott módszer alakja: $y_{n+1} = y_n + h(1/6k_1 + 1/3k_2 + 1/3k_3 + 1/6k_4)$.

8.14. Az explicit Runge-Kutta-típusú módszerek Butcher-táblázatainak konzisztencia feltételeit fogjuk ellenőrizni.

(a) Az explicit Euler-módszer Butcher-táblázata megtalálható a 8.10. feladat megoldásának (a) részénél. A konzisztenciarend számításához ellenőrizzük a feltételeket!

$$\mathbf{c}^T \cdot \mathbf{e} = 1 \cdot 1 = 1$$

$$\mathbf{c}^T \cdot \mathbf{a} = 1 \cdot 0 \neq 1/2$$

Azaz az explicit Euler-módszer elsőrendben konzisztens.

(c) Az Euler–Heun-módszer Butcher-táblázata megtalálható a 8.10. feladat megoldásának (c) részénél. A konzisztenciarend számításához ellenőrizzük a feltételeket!

$$\mathbf{c}^T \cdot \mathbf{e} = 1 \cdot 1 = 1$$

$$\mathbf{c}^T \cdot \mathbf{a} = 1/2 \cdot 0 + 1/2 \cdot 1 = 1/2$$

$$\mathbf{c}^T \cdot \mathbf{a}^2 = 1/2 \cdot 0 + 1/2 \cdot 1 \neq 1/3$$

Azaz az Euler–Heun-módszer másodrendben konzisztens.

A további eredmények az Útmutatások, végeredmények fejezetben megtalálhatóak.

8.15. Két feladatrésszel foglalkozunk részletesen: (további eredmények az Útmutatások, végeredmények fejezetben)

(b) Tekintsük a b feladatrész képletét. Meghatározzuk a szükséges paramétereket.

$$y_{n+1} = y_n + h[\underbrace{(1 - \frac{1}{2\alpha})}_{c_1} \underbrace{f(t_n, y_n)}_{k_1} + \underbrace{\frac{1}{2\alpha}}_{c_2} \underbrace{f(t_n + \underbrace{\alpha}_{a_2} h, y_n + \underbrace{\alpha}_{b_{21}} h \underbrace{f(t_n, y_n)}_{k_1})}_{k_2}]$$

Azaz a módszert felírhatjuk $y_{n+1} = y_n + h((1 - 1/(2\alpha))k_1 + (1/2\alpha)k_2)$ alakban. Ekkor a Butcher-táblázat:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1 - \frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}$$

(d) Tekintsük az d feladatrész képletét. Meghatározzuk a szükséges paramétereket.

$$y_{n+1} = y_n + h \left[\underbrace{\frac{1}{4}}_{c_1} \underbrace{f(t_n, y_n)}_{k_1} + \underbrace{\frac{3}{4}}_{k_2} f \left(t_n + \underbrace{\frac{2}{3}}_{a_2} h, y_n + \underbrace{\frac{2}{3}}_{b_{21}} \underbrace{f(t_n, y_n))}_{k_2} \right) \right]$$

Azaz a kétlépcsős módszert felírhatjuk $y_{n+1} = y_n + h(1/4k_1 + 3/4k_2)$ alakban. Ekkor a Butcher-táblázat:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 2/3 & 2/3 & 0 \\ \hline & 1/4 & 3/4 \end{array}$$

8.16. Az egylépéses módszerek tesztfeladatra történő alkalmazása adott y_0 kezdeti érték mellett az $y_{n+1} = R(z)y_n$ iterációt jelenti.

(a) Az explicit Euler-módszert a tesztfeladatra alkalmazva nyerjük, hogy

$$y_{n+1} = y_n + hf(t_n, y_n) = y_n + h\lambda y_n = (1+z)y_n,$$

azaz $R(z) = 1 + z$.

(b) Az implicit Euler-módszert a tesztfeladatra alkalmazva nyerjük, hogy

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}) = y_n + \lambda h y_{n+1} = (1+z)y_n.$$

Ezt rendezve kapjuk, hogy $y_{n+1} = \dfrac{1}{1-\lambda h}y_n = \dfrac{1}{1-z}y_n$, azaz $R(z) = \dfrac{1}{1-z}$.

(c) A Crank–Nicolson-módszert a tesztfeladatra alkalmazva nyerjük, hogy

$$y_{n+1} = y_n + \frac{h}{2}\Big(f(t_n, y_n) + f(t_{n+1}, y_{n+1})\Big)$$

$$= y_n + \frac{h}{2}\Big(\lambda y_{n+1} + \lambda y_n\Big) = y_n + \frac{z}{2}\Big(y_{n+1} + y_n\Big).$$

Ezt rendezve kapjuk, hogy

$$y_{n+1} = \frac{1 + z/2}{1 - z/2}y_n,$$

azaz $R(z) = \dfrac{2+z}{2-z}$.

(e) A javított Euler-módszert a tesztfeladatra alkalmazva nyerjük, hogy

$$y_{n+1} = y_n + hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}f(t_n, y_n)\right) = y_n + h\lambda\left(y_n + \frac{h}{2}\lambda y_n\right) = \left(1 + z + \frac{z^2}{2}\right)y_n.$$

Azaz $R(z) = 1 + z + \dfrac{z^2}{2}$.

A további feladatrészek eredményei az Útmutatások, végeredmények fejezetben megtalálhatóak.

8.17. Az adott numerikus módszer pontosan akkor teljesíti az abszolút stabil tulajdonságot, ha $|R(z)| \leq 1$. Továbbá a módszert A-stabilnak nevezzük, ha az abszolút stabilitási tartománya tartalmazza a $\mathbb{C}^- \subset \mathbb{C}$ félsíkot, azaz a $\mathrm{Re}(z) \leq 0$ komplex számokat.

(a) Az explicit Euler-módszer stabilitási függvénye: $R(z) = 1 + z$. Ekkor a stabilitási tartományt az alábbi módon határozhatjuk meg:

$$|R(z)| = |1+z| \leq 1,\ \forall z \in \mathbb{C} \Leftrightarrow |1+x+iy| \leq 1, \forall x, y \in \mathbb{R} \Leftrightarrow (1+x)^2+y^2 \leq 1\ \forall x, y \in \mathbb{R}.$$

Azaz az explicit Euler-módszer abszolút stabilitási tartománya a komplex síkon a $(-1, 0)$ középpontú 1 sugarú körlapot jelöli ki, mely nem tartalmazza a $\mathbb{C}^- \subset \mathbb{C}$ félsíkot, így a módszer nem A-stabil.

(b) Az implicit Euler-módszer stabilitási függvénye: $R(z) = \dfrac{1}{1-z}$. Ekkor a stabilitási tartományt az alábbi módon határozhatjuk meg:

$$|R(z)| = \left|\frac{1}{1-z}\right| \leq 1,\ \forall z \in \mathbb{C} \Leftrightarrow |1-z| \geq 1,\ \forall z \in \mathbb{C} \Leftrightarrow (1-x)^2-y^2 \geq 1\ \forall x, y \in \mathbb{R}.$$

Azaz az implicit Euler-módszer abszolút stabilitási tartománya a komplex síkon az $(1, 0)$ középpontú 1 sugarú körlap komplementerét jelöli ki, mely tartalmazza a $\mathbb{C}^- \subset \mathbb{C}$ félsíkot, így a módszer A-stabil.

(c) A Crank–Nicolson-módszer stabilitási függvénye: $R(z) = \dfrac{2+z}{2-z}$. Ekkor a stabilitási tartományt az alábbi módon határozhatjuk meg:

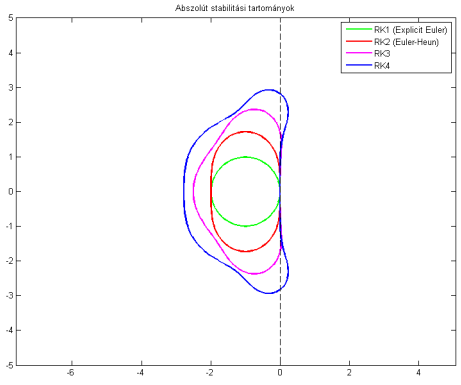
$$|R(z)| = \left|\frac{2+z}{2-z}\right| \leq 1,\ \forall z \in \mathbb{C} \Leftrightarrow |2+z|^2 \leq |2+z|^2,\ \forall z \in \mathbb{C} \Leftrightarrow |2+x| \leq |2-x|\ \forall x \in \mathbb{R}.$$

Azaz a Crank–Nicolson-módszer abszolút stabilitási tartománya a komplex síkon a $\mathrm{Re}(z) \leq 0$ komplex számokat jelöli, mely tartalmazza a $\mathbb{C}^- \subset \mathbb{C}$ félsíkot, így a módszer A-stabil.

(e) A javított Euler-módszer stabilitási függvénye: $R(z) = 1 + z + \dfrac{z^2}{2}$. Ekkor a stabilitási tartomány meghatározásához az $|R(z)| \leq 1$ összefüggésnek kell teljesülnie a komplex síkon. Azaz a javított Euler-módszer nem tartalmazza a $\mathbb{C}^- \subset \mathbb{C}$ félsíkot, így a módszer nem A-stabil.

8.19. A feladatot az Astabilitas2.m fájl oldja meg. A futtatás eredményeképpen kapjuk az alábbi ábrát, mely az egyes módszerek abszolút stabilitási tartományainak körvonalait jeleníti meg. A futtatás eredménye és az Astabilitas2.m fájl forráskódja:

```
[X,Y]=meshgrid(-5:0.1:5,-5:0.1:5);
Z=X +Y*i;
%Explicit Euler
M=abs(1+Z);
[c,h]=contour(X,Y,M,[1,1]);
set(h,'linewidth',2,'edgecolor','g')
hold on
%Euler-Heun-módszer
M=abs(1+Z/2.*(2+Z));
[c,h]=contour(X,Y,M,[1,1]);
set(h,'linewidth',2,'edgecolor','r')
%RK3
M=abs(1+Z/6.*(6+Z/3.*(9+3*Z)));
[c,h]=contour(X,Y,M,[1,1]);
set(h,'linewidth',2,'edgecolor','m')
%RK4
M=abs(1+Z/24.*(24+Z/12.*(144+Z/48.*(2304+576*Z))));
[c,h]=contour(X,Y,M,[1,1]);
set(h,'linewidth',2,'edgecolor','b')
axis equal
y=-5:0.1:5;
x=0*y;
plot(x,y,'k--')
title('Abszolút stabilitási tartományok')
legend('RK1 (Explicit Euler)', 'RK2 (Euler-Heun)', 'RK3', 'RK4')
```



10.7. ábra. A megadott módszerek abszolút stabilitási tartományainak körvonalai.

8.20. A stabilitási tartományt a stabilitási függvények segítségével határozhatjuk meg. Mivel mindkét módszer stabilitási függvényét már a 8.16. feladat (b) és (c) része során meghatároztuk, így ezeket kell a MATLAB-ban beprogramoznunk. Ezt az Astabilitas.m fájl realizálja. A forráskód idevágó részlete az alábbi:

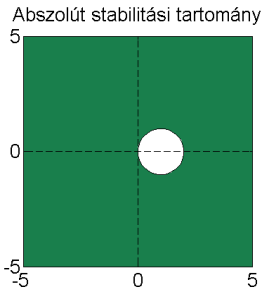
```
%Implicit Euler
clf;
[X,Y] = meshgrid(linspace(-5,5), linspace(-5,5));
Z = X+Y*1i;
phi = 1./(1-Z);
contourf(X,Y,1-abs(phi), [0 0], 'LineWidth', 1);
set(gca,'FontSize', 20, 'CLim', [0 1]);
colormap([.1 .5 .3; 0 0 0; 1 1 1]);
hold on;
plot([-5 5], [0 0], '--k', 'LineWidth', 1);
plot([0 0], [-5 5], '--k', 'LineWidth', 1);
title('Abszolút stabilitási tartomány')

%Crank-Nicolson
clf;
[X,Y] = meshgrid(linspace(-5,5), linspace(-5,5));
Z = X+Y*1i;
phi = (2+Z)./(2-Z);
contourf(X,Y,1-abs(phi), [0 0], 'LineWidth', 1);
set(gca,'FontSize', 20, 'CLim', [0 1]);
colormap([.1 .5 .3; 0 0 0; 1 1 1]);
hold on;
plot([-5 5], [0 0], '--k', 'LineWidth', 1);
plot([0 0], [-5 5], '--k', 'LineWidth', 1);
title('Abszolút stabilitási tartomány')
```

A kód az implicit Euler és Crank–Nicolson-módszerek abszolút stabilitási tartományait ábrázolja a $[-5, 5] \times [-5, 5]$ négyzeten.

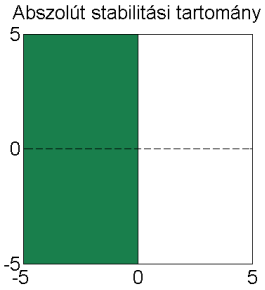
A futtatás eredményeként visszaköszön az elméletből ismert tény, miszerint mindkét módszer A-stabil. A MATLAB-ban való futtatáshoz a megfelelő részek kommentelése után alábbi parancsot írjuk be:

```
>> Astabilitas.m
```

Ekkor a következő két ábrát adja vissza az Astabilitas.m fájl:



10.8. ábra. Az implicit Euler-módszer stabilitási tartománya.



10.9. ábra. A Crank–Nicolson-módszer stabilitási tartománya.

8.21. A tesztfeladatra tetszőleges $\lambda < 0$ esetén a megoldás korlátos és monoton csökkenő ($y(t) = e^{\lambda t}$). Így csak azok a numerikus megoldások tudják a pontos megoldást jól approximálni, amelyekre az előállított numerikus megoldás is rendelkezik ezekkel a tulajdonságokkal, nevezetesen amikor teljesül az

$$|R(h\lambda)| \leq 1, \quad h > 0, \ \lambda \in \mathbb{R}$$

feltétel. Az explicit Euler-módszer esetében ez ekvivalens a $h \leq 2/(-\lambda)$ feltétellel, míg implicit Euler-módszer esetén tudjuk, hogy a módszer feltétel nélkül stabil. Azaz utóbbi esetében a táblázatban szereplő λ értékek tetszőleges h lépésköz mellett jól viselkednek. Ezzel szemben az explicit módszer numerikus értékei akkor approximálják jól a feladat pontos megoldásának értékeit, ha teljesül a már fent említett feltétel, azaz $h \leq h_0$, ha $h_0 = 2/(-\lambda)$.

8.22. A 8.16. feladat eredménye szerint

$$R(h\lambda) = \frac{2 + h\lambda}{2 - h\lambda}.$$

Könnyen láthatóan tetszőleges $h > 0$ és $\lambda < 0$ esetén $|R(h\lambda)| \leq 1$. Ugyanakkor a tesztfeladaton tetszőleges $h > 0$ mellett mégsem viselkedik jól a módszer. Ugyanis $h >$

$2(-\lambda)$ esetén $R(h\lambda) \in (-1, 0)$, ezért az ilyen rácshálókon a numerikus értékek lépésenként előjelet váltanak, azaz bár abszolút értékben csökkenek, de emellett oszcillálnak is, ami ellentmond a pontos megoldás szigorú monoton csökkenésének.

8.23. Alkalmazzuk a feladatra az explicit Euler-módszer. Ekkor kapjuk, hogy

$$y_{n+1} = y_n + hf(t_n, y_n) = y_n + h(1 - 10y_n) = (1 - 10h)^2 y_{n-1} - 10h^2 + 2h = \\ = \ldots = (1 - 10h)^{n+1} y_0 + K(h),$$

ahol $K(h)$ a h-tól függő maradéktagot jelöli. A feladat pontos megoldásának követéséhez a $|1 - 10h| < 1$ feltételnek kell teljesülnie. Azaz könnyen látható módon $h > 1/5$ esetén a módszer numerikus értékei oszcillálnak és a végtelenhez tartanak.

Többlépéses módszerek

8.25. Alkalmazzuk a Taylor-sorfejtést!

(a) Szorozzuk végig a többlépéses módszert hárommal! Ekkor nyerjük, hogy

$$3y(t_n) - 4y(t_{n-1}) + y(t_{n-2}) = 2hf(t_n, y(t_n)).$$

Fejtsünk sorba a t_{n-1} és t_{n-2} tagokat a t_n pont körül!

$$y(t_{n-1}) = y(t_n) - hy'(t_n) + \frac{h^2}{2} y''(t_n) + \frac{h^3}{3!} y'''(t_n) + O(h^4)$$

$$y(t_{n-2}) = y(t_n) - 2hy'(t_n) + 2h^2 y''(t_n) - \frac{4h^3}{3} y'''(t_n) + O(h^4)$$

Ezeket az eredeti egyenletbe helyettesítve, valamint $y'(t_n) = f(t_n, y(t_n))$ összefüggést használva kapjuk, hogy a módszer hibatagja $-\frac{2}{3}h^3 y'''(t_n)$. Azaz a módszer másodrendű.

A további feladatrészek eredményei az Útmutatások, végeredmények fejezetben megtalálhatóak.

8.26. A lineáris többlépéses módszer p-ed rendű, ha a módszert definiáló paraméterekre teljesülnek az alábbi feltételek:

$$\sum_{k=0}^{m} a_k = 0, \quad \frac{1}{j} \sum_{k=0}^{m} k^j a_k + \sum_{k=0}^{m} k^{j-1} b_k = 0 \quad j = 1, \ldots, p.$$

(a) Határozzuk meg a többlépéses módszer együtthatóit!

$$a_0 = 1,\ a_1 = -4/3,\ a_2 = 1/3,\ b_0 = 2/3,\ b_1 = 0,\ b_2 = 0$$

Ellenőrizzük, hogy mely feltételek teljesülnek!

$$\begin{aligned}
\sum a_k: &\qquad 1 - 4/3 + 1/3 = 0\\
j = 1: &\qquad 1 \cdot (-4/3) + 2 \cdot (4/3) + 2/3 = 0\\
j = 2: &\qquad \frac{1}{2}(1^2 \cdot (-4/3) + 2^2 \cdot 1/3) + 2/3 = 0\\
j = 3: &\qquad \frac{1}{3}(1^3 \cdot (-4/3) + 2^3 \cdot 1/3) + 0 \neq 0
\end{aligned}$$

Azaz az $y_n - \frac{4}{3}y_{n-1} + \frac{1}{3}y_{n-2} = \frac{2}{3}hf_n$ módszer másodrendben konzisztens.

(b) Határozzuk meg a többlépéses módszer együtthatóit!

$$a_0 = 1,\ a_1 = -1,\ a_2 = 0,\ b_0 = 0,\ b_1 = 3/2,\ b_2 = -1/2$$

Ellenőrizzük, hogy mely feltételek teljesülnek!

$$\begin{aligned}
\sum a_k: &\qquad 1 - 1 + 0 = 0\\
j = 1: &\qquad 1 \cdot (-1) + 2 \cdot 0 + 3/2 - 1/2 = 0\\
j = 2: &\qquad \frac{1}{2}(1^2 \cdot (-1) + 2^2 \cdot 0) + (1^1 \cdot (3/2) + 2^1 \cdot (-1/2)) = 0\\
j = 3: &\qquad \frac{1}{3}(1^3 \cdot (-1) + 2^3 \cdot 0) + (1^2 \cdot (3/2) + 2^2 \cdot (-1/2)) \neq 0
\end{aligned}$$

Azaz az $y_n - y_{n-1} = h\left(\frac{3}{2}f_{n-1} - \frac{1}{2}f_{n-2}\right)$ módszer másodrendben konzisztens.

A további feladatrészek eredményei az Útmutatások, végeredmények fejezetben megtalálhatóak.

8.27. A módszer alakjából leolvasható, hogy másodrendű és az alábbi együtthatók adottak:

$$a_0 = 1,\ b_0 = 0.$$

Az utóbbi érték azt jelenti, hogy a módszer explicit. Ugyanakkor egy m-lépéses explicit módszer lineáris többlépéses módszer maximális rendje 2m-1. Azaz a módszer legfeljebb harmadrendben lehet konzisztens. Írjuk fel a harmadrendű konzisztencia meghatározásának feltételeit.

$$\begin{aligned}
\sum a_k: &\qquad 1 + a_1 + a_2 = 0\\
j = 1: &\qquad a_1 + 2a_2 + b_1 + b_2 = 0\\
j = 2: &\qquad 1/2(a_1 + 4a_2) + b_1 + b_2 = 0\\
j = 3: &\qquad 1/3(a_1 + 8a_2) + b_1 + 4b_2 = 0
\end{aligned}$$

Ekkor az együtthatókra az alábbi egyenletrendszert írhatjuk fel:

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 1 \\ 1 & 4 & 2 & 4 \\ 1 & 8 & 3 & 12 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Az egyenletrendszert megoldva az alábbi együtthatókat kapjuk:

$$a_1 = 4,\ a_2 = -5,\ b_1 = 4,\ b_2 = 2.$$

A kétlépéses módszer ismeretlen paramétereit meghatározva az

$$y_n + 4y_{n-1} - 5y_{n-2} = h(4f_{n-1} + 2f_{n-2})$$

maximális, harmadrendben konzisztens módszert nyerjük.

8.29. Az differenciálegyenletet felhasználva kapjuk az

$$f(t_{n-2}, y(t_{n-2})) = y'(t_{n-2}) = -y(t_{n-2})$$

összefüggést. Ezt az eredményt az eredeti módszerbe helyettesítve nyerjük, hogy:

$$y_n - 4y_{n-1} + 3y_{n-2} = 2hy_{n-2} = y_n - 4y_{n-1} + (3 - 2h)y_{n-2} = 0.$$

A többlépéses módszer elindításához az $y_0 = 1$ és $y_1 = e^{-h}$ kezdeti értékeket használjuk. Ekkor $h = 1/10$ lépésközökkel a módszerre megírt program az alábbi eredményeket adja vissza:

t	0	1/10	2/10	3/10	4/10	...	9/10	1
y	1	0.9048	0.8193	0.7439	0.6812	...	3.7702	10.7856
e	1	0.9048	0.8187	0.7408	0.6703	...	0.4066	0.3679
hiba	0	0	0.0006	0.0031	0.0109	...	3.3636	10.4177

10.12. táblázat. Hibaértékek az aktuális rácspont esetén.

Könnyen látható módon a feladatra alkalmazott módszer nem konvergens. Mégis mivel magyarázható a fenti táblázat eredménye? A válaszhoz írjuk fel a többlépéses módszer karakterisztikus egyenletét.

$$\varrho(\xi) = \xi^2 - 4\xi + 3 - 2h = 0.$$

Az egyenlet gyökei $\xi_{1,2} = 2 \pm \sqrt{1 + 2h}$. Ekkor a numerikus megoldás $y_n \equiv c_1\xi_1^n + c_2\xi_2^n$ alakú. A $\xi_1^n = (2 - \sqrt{1 + 2h})^n \to 0$, ha $n \to \infty$. Ezzel szemben a $\xi_2^n = (2+\sqrt{1+2h})^n \to \infty$, ha $n \to \infty$. Azaz a második gyök dominál (körülbelül 3 értékét hatványozzuk) és a numerikus megoldás nem követi a pontos megoldás lecsengését.

8.30. A lineáris többlépéses módszer kielégíti a gyökkritériumot, ha a

$$\varrho(\xi) = \sum_{k=0}^{m} a_k \xi^{m-k} = 0$$

karakterisztikus egyenlet $\xi_k \in \mathbb{C}$ gyökeire $|\xi_k| \leq$ minden $k = 1, \ldots, m$-re és $|\xi_k| = 1$ tulajdonságú gyökök egyszeresek. Az egyes feladatrészeknél a karakterisztikus egyenlet felírása után célunk a gyökök tulajdonságainak ellenőrzése a fenti definíció értelmében.

(a) A módszer karakterisztikus egyenlete:

$$\varrho(\xi) = \xi^2 - 6\xi + 5 = 0.$$

A két gyök: $\xi_1 = 1$ és $\xi_2 = 5$. Így a módszer nem teljesíti a gyökkritériumot, lévén van egynél nagyobb abszolút értékű gyöke.

(b) A módszer karakterisztikus egyenlete:

$$\varrho(\xi) = \xi^2 - 1 = 0.$$

A két gyök: $\xi_1 = 1$ és $\xi_2 = -1$. Így a módszer teljesíti a gyökkritériumot, lévén az egy abszolút értékű gyökök egyszeresek.

A további feladatrészek eredményei az Útmutatások, végeredmények fejezetben megtalálhatóak.

8.31. Egy lineáris többlépéses módszer erős stabilitásának ellenőrzéséhez szükséges a gyökkritérium teljesülése és az, hogy csak a $\xi = 1$ az egyetlen 1 abszolút értékű gyöke. Az egyes feladatrészeknél a karakterisztikus egyenlet felírása után célunk a gyökök tulajdonságainak ellenőrzése a fenti definíció értelmében.

Tekintsük a 8.25. feladatot.

(a) A módszer karakterisztikus egyenlete:

$$\varrho(\xi) = \xi^2 - 4/3\xi + 1/3 = 0.$$

A két gyök: $\xi_1 = 1$ és $\xi_2 = 1/3$. A módszer teljesíti a gyökkritériumot és csak a $\xi_1 = 1$ az egyetlen 1 abszolút értékű gyöke. Azaz a módszer erősen stabil.

(b) A módszer karakterisztikus egyenlete:

$$\varrho(\xi) = \xi^2 - 4\xi + 3 = 0.$$

A két gyök: $\xi_1 = 1$ és $\xi_2 = 1/3$. A módszer teljesíti a gyökkritériumot és csak a $\xi_1 = 1$ az egyetlen 1 abszolút értékű gyöke. Azaz a módszer erősen stabil.

(c) A módszer karakterisztikus egyenlete:

$$\varrho(\xi) = \xi^2 + 4\xi - 5 = 0.$$

A két gyök: $\xi_1 = -2 + i$ és $\xi_2 = -2 - i$. A módszer nem teljesíti a gyökkritériumot, lévén a gyökei abszolút értékben nagyobbak egynél ($|\xi_1| = |\xi_2| = \sqrt{5}$). Azaz a módszer nem erősen stabil.

A további feladatrészek eredményei az Útmutatások, végeredmények fejezetben megtalálhatóak.

8.32. Akár az explicit (Adams–Bashforth), akár az implicit (Adams–Moulton) Adams-módszerek karakterisztikus egyenletét írjuk fel, az alábbihoz jutunk:

$$\varrho(\xi) = \sum_{k=0}^{m} a_k \xi^{m-k} = \xi^m - \xi^{m-1} = \xi^{m-1}(\xi - 1) = 0.$$

Azaz a gyökök: $\xi_1 = 1$ és $\xi_{2,\dots,m} = 0$. A módszerek teljesítik a gyökkritériumot és csak a $\xi_1 = 1$ az egyetlen 1 abszolút értékű gyöke a karakterisztikus egyenletnek. Azaz az Adams-típusú módszerek erősen stabilak.

A közönséges differenciálegyenletek peremérték-feladatainak numerikus módszerei

Peremérték-feladatok megoldhatósága

9.1. A másodrendű, állandó együtthatós differenciálegyenletek megoldását szokásos módon $u(x) = e^{\lambda x}$ alakban keressük. Ekkor az eredeti egyenletre az alábbi karakterisztikus egyenletet nyerjük:

$$k(\lambda) = \lambda^2 - 1 = 0.$$

Ennek gyökei a $\lambda_1 = 1$ és $\lambda_2 = -1$. Ekkor a peremérték figyelembevétele nélkül a megoldás $u(x) = c_1 e^x + c_2 e^{-x},\ c_1, c_2 \in \mathbb{R}$ alakban áll elő. Fejezzük ki a peremértékek segítségével a c_1 és c_2 konstansokat!

$$u(0) = c_1 + c_2 = 2/3 \Rightarrow c_1 = 2/3 - c_2$$

$$u(1) = (2/3 - c_2)e + c_2(1/e) = 3/8$$

Ezekből az egyenletekből kifejezve a konstansokat a feladat megoldása:

$$u(x) = \Big(\frac{2}{3} - \frac{\frac{3}{8}e - \frac{2}{3}e^2}{1 - e^2}\Big)e^x + \Big(\frac{\frac{3}{8}e - \frac{2}{3}e^2}{1 - e^2}\Big)e^{-x}.$$

9.3. A másodrendű, állandó együtthatós differenciálegyenletek megoldását szokásos módon $u(x) = e^{\lambda x}$ alakban keressük. Ekkor az eredeti egyenletre az alábbi karakterisztikus egyenletet nyerjük:

$$k(\lambda) = \lambda^2 + 4\lambda = 0.$$

Ennek gyökei a $\lambda_1 = 2i$ és $\lambda_2 = -2i$. Ekkor a peremérték figyelembevétele nélkül a megoldás $u(x) = c_1 \sin(2x) + c_2 \cos(2x),\ c_1, c_2 \in \mathbb{R}$ alakban áll elő. Fejezzük ki a peremértékek segítségével a c_1 és c_2 konstansokat!

$$u(0) = c_2 = 1$$

$$u(\pi/2) = -1 = -1$$

Ezekből az egyenletekből kifejezve a konstansokat a feladat megoldása:

$$u(x) = \cos(2x) - c_1 \sin(2x), \;\; c_1 \in \mathbb{R}.$$

Azaz a feladatnak van megoldása, az nem egyértelmű és nem elemi függvények körében található.

9.7. Alkalmazzuk a jegyzet 10.3.2 tételének következményét (lineáris esetre) a feladatok egyértelmű megoldásának létezésére! A tétel elégséges feltételt ad.

(a) A tételt alkalmazva nyerjük, hogy $p(x) = 0, \;\; q(x) = 1, \;\; r(x) = \sin(x)$. Ellenőrizzük a tételben szereplő feltételeket:

- $f(x, u, u') \in C[T]$,
- $q(x), r(x) \in C[T]$ és $q(x) > 0$ minden $t - re$,
- létezik $M \geq 0 : |p(x)| \leq M$, például $M = 1$.

Azaz a tétel feltételei teljesülnek, így a peremérték-feladatnak létezik egyértelmű megoldása.

A további feladatrészek eredményei az Útmutatások, végeredmények fejezetben megtalálhatóak.

9.8. Tekintsük a lineáris peremérték-feladatot:

$$u'' = f(x, u, u') \equiv p(x)u' + q(x)u + r(x), x \in [a, b]$$

$$u(a) = \alpha, \;\; u(b) = \beta,$$

ahol $p, q, r \in C[a, b]$ adott folytonos függvények. A fenti lineáris peremérték-feladat elsőrendű rendszer alakjában az alábbi módon írható fel:

$$\mathbf{u}' = A(x)\mathbf{u} + \mathbf{r}(x),$$

ahol

$$A(x) = \begin{pmatrix} 0 & 1 \\ q(x) & p(x) \end{pmatrix}, \;\; \mathbf{r}(x) = \begin{pmatrix} 0 \\ r(x) \end{pmatrix}.$$

A peremfeltételek felírásához vezessük be a

$$B_a = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, B_b = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \mathbf{v} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

jelöléseket! Ekkor a fenti feladat peremfeltétele

$$B_a\mathbf{u}(a) + B_b\mathbf{u}(b) = \mathbf{v}$$

alakban írható fel.

(b) Alkalmazzuk ezen ismereteket a konkrét feladatra:

$$\begin{cases} u''(x) = \lambda u'(x) + \lambda^2 u(x),\ x \in [0,1],\ \lambda \in [0.5,1] \\ u(0) = 5,\ u(1) = 8! \end{cases}$$

Először írjuk át a peremérték-feladat differenciálegyenletet tartalmazó sorát a kívánt alakba! Ehhez a szokásos helyettesítést hajtjuk végre:

$$\begin{cases} u_1 = u \Rightarrow u_1' = u_2, \\ u_2 = u' \Rightarrow u_2' = u'' = \lambda u_2 + \lambda^2 u_1. \end{cases}$$

Azaz az elsőrendű rendszer:

$$\mathbf{u}'(x) = A\mathbf{u}(x) = \begin{pmatrix} 0 & 1 \\ \lambda^2 & \lambda \end{pmatrix} \begin{pmatrix} u_1(x) \\ u_2(x) \end{pmatrix}.$$

A bevezetett jelölésekkel a peremfeltételek: $u_1(0) = 5$ és $u_1(1) = 8$. Ekkor a feladat peremfeltétele

$$B_a\mathbf{u}(0) + B_b\mathbf{u}(1) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1(0) \\ \mathbf{u}_2(0) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1(1) \\ \mathbf{u}_2(1) \end{pmatrix} = \begin{pmatrix} 5 \\ 8 \end{pmatrix} = \mathbf{v}.$$

(c) Alkalmazzuk ezen ismereteket a konkrét feladatra:

$$\begin{cases} u'''(x) = -2\lambda^3 u(x) + \lambda^2 u'(x) + 2\lambda u''(x), \quad x \in (0,1) \\ u(0) = \beta_1,\ u(1) = \beta_2,\ u'(1) = \beta_3! \end{cases}$$

Először írjuk át a peremérték-feladat differenciálegyenletet tartalmazó sorát a kívánt alakba! Ehhez a szokásos helyettesítést hajtjuk végre:

$$\begin{cases} u_1 = u \Rightarrow u_1' = u_2, \\ u_2 = u' \Rightarrow u_2' = u_3, \\ u_3 = u'' \Rightarrow u_3' = -2\lambda^3 u_1 + \lambda^2 u_2 + 2\lambda u_3. \end{cases}$$

Azaz az elsőrendű rendszer:

$$\mathbf{u}'(x) = A\mathbf{u}(x) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2\lambda^3 & \lambda^2 & 2\lambda \end{pmatrix} \begin{pmatrix} u_1(x) \\ u_2(x) \\ u_3(x) \end{pmatrix}.$$

A bevezetett jelölésekkel a peremfeltételek: $u_1(0) = \beta_1,\ u_1(1) = \beta_2$ és $u_2(1) = \beta_3$. Ekkor a feladat $B_0\mathbf{u}(0) + B_1\mathbf{u}(1) = \mathbf{v}$ peremfeltétele:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1(0) \\ \mathbf{u}_2(0) \\ \mathbf{u}_3(0) \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1(1) \\ \mathbf{u}_2(1) \\ \mathbf{u}_3(1) \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Az (a) feladatrész eredménye az Útmutatások, végeredmények fejezetben megtalálható.

9.9. A lineáris peremérték-feladat elsőrendű $\mathbf{u}' = A(x)\mathbf{u} + \mathbf{r}(x)$ rendszerének megoldása felírható a következő módon. Legyen $\mathbf{Y}(x) \in \mathbb{R}^{n\times n}$ az egyenlet alapmegoldása. Ekkor az eredeti egyenlet általános megoldása

$$\mathbf{u}(x) = \mathbf{Y}(x)\Big(\mathbf{c} + \int_a^x \mathbf{Y}^{-1}(s)\mathbf{r}(s)ds\Big),$$

ahol $\mathbf{c} \in \mathbb{R}^n$ egy tetszőleges vektor. Célunk olyan $\mathbf{c}$ megválasztása, amely mellett a fenti $\mathbf{u}(x)$ függvény kielégíti a $B_a\mathbf{u}(a) + B_b\mathbf{u}(b) = \mathbf{v}$ peremfeltételt. Ez pontosan az alábbi esetben teljesül:

A lineáris peremérték-feladatnak pontosan akkor létezik egyértelmű megoldása, amikor a $Q = B_a + B_b\mathbf{Y}(b)$ mátrix reguláris. Emellett a keresendő $\mathbf{c}$ vektor az alábbi:

$$\mathbf{c} = Q^{-1}\Big(\mathbf{v} - B_b\mathbf{Y}(b)\int_a^b \mathbf{Y}^{-1}(s)\mathbf{r}(s)ds\Big).$$

(a) Alkalmazzuk a fenti módszert a konkrét feladatra!

$$\begin{cases} u''(x) = -u(x), \quad t \in (0, b) \\ u(0) = \alpha,\ u(b) = \beta \end{cases}$$

A 9.8. feladatban ismertetett módszer segítségével felírt $A(x)$ mátrix:

$$A(x) \equiv A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

A feladat peremfeltételének peremmátrixai:

$$B_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B_b = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

A Q mátrix felírásához szükséges az $\mathbf{Y}(x)$ alapmegoldás meghatározása is. Ennek meghatározása 2×2 mátrixok esetében a Hermite-féle interpolációs polinom segítségével számítható. Az A mátrix sajátértékei: $\lambda_1 = i$ és $\lambda_2 = -i$. Mivel a két sajátérték különböző, így az interpolációs polinomot az alábbi módon határozzuk meg:

$$\begin{cases} p(\lambda_1) = a_1(x)i + a_0(x) = e^{ix} = \cos(x) + i\sin(x) \\ p(\lambda_2) = -a_1(x)i + a_0(x) = e^{-ix} = \cos(x) - i\sin(x) \end{cases}$$

Ekkor az $a_0(x)$ és $a_1(x)$ polinomokra az alábbi adódik:

$$a_0(x) = \cos(x), \quad a_1(x) = \sin(x), \ x \in (0, b)$$

Ekkor az $\mathbf{Y}(x)$ alapmegoldás:

$$\mathbf{Y}(x) = \sin(x)A + \cos(x)I = \begin{pmatrix} \cos(x) & \sin(x) \\ -\sin(x) & \cos(x) \end{pmatrix}.$$

Ennek segítségével már meghatározható a Q mátrix is. Nevezetesen:

$$Q = B_0 + B_b\mathbf{Y}(b) = \begin{pmatrix} 1 & 0 \\ \cos(b) & \sin(b) \end{pmatrix}.$$

A peremérték-feladat megoldásának egyértelműségéhez a Q mátrix regularitása szükséges. Ez pontosan akkor teljesül, ha $\det Q \neq 0 \Leftrightarrow \sin(b) \neq 0 \Leftrightarrow b \neq k\pi, \ k \in \mathbb{Z}$. Azaz a feladatnak tetszőleges $b \neq k\pi, \ k \in \mathbb{Z}$ esetén létezik egyértelmű megoldása.

A (b) feladatrész eredménye az Útmutatások, végeredmények fejezetben megtalálható.

Véges differenciák módszere és a belövéses módszer

9.10. A numerikus módszer megadásához definiálunk egy rácshálót. Legyen $\omega_h \subset [0, l]$ a h lépésközű ekvidisztáns rácsháló:

$$\omega_h = \{x_i = ih, \ i = 1, 2, \ldots, N-1, \ h = l/N\}!$$

Az intervallum két végpontját hozzávéve:

$$\overline{\omega}_h = \{x_i = ih, \ i = 0, 1, \ldots, N, \ h = l/N\}.$$

Tekintsük az eredeti feladatot az ω_h rácsháló pontjaiban. Ekkor

$$-u''(x_i) = f(x_i),\ x_i \in \omega_h$$

$$u(x_0) = \mu_1,\ u(x_N) = \mu_2.$$

Jelölje $\mathbb{F}(\overline{\omega}_h)$ az $\overline{\omega}_h$-n értelmezett függvények vektorterét és legyen y_h egy adott rácsfüggvény! A második deriváltat a standard másodrendű differenciahányadossal helyettesítve az eredeti feladat az alábbi alakba írható át:

$$-\frac{y_h(x_i+h) - 2y_h(x_i) + y_h(x_i-h)}{h^2} = f(x_i),\ x_i \in \omega_h$$

$$y_h(x_0) = \mu_1,\ y_h(x_N) = \mu_2.$$

Mivel az y_h rácsfüggvény és a jobboldali vektor is azonosítható egy $\mathbb{R}^{N+1}$-beli vektorral, nevezetesen:

$$\vec{y}_h \in \mathbb{R}^{N+1}:\ (\vec{y}_h)_i = y_h(x_i),\ \text{illetve}\ (\vec{f}_h)_i = f(x_i),\ x_i \in \overline{\omega}_h.$$

Így az eredeti feladat egy $A_h\vec{y}_h = \vec{f}_h$ lineáris algebrai egyenletrendszerként írható fel, ahol $A_h \in \mathbb{R}^{(N+1)\times(N+1)}$. Ennek alakja:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ \frac{-1}{h^2} & \frac{2}{h^2} & \frac{-1}{h^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{-1}{h^2} & \frac{2}{h^2} & \frac{-1}{h^2} & 0 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{-1}{h^2} & \frac{2}{h^2} & \frac{-1}{h^2} & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} (\vec{y}_h)_0 \\ (\vec{y}_h)_1 \\ (\vec{y}_h)_2 \\ \vdots \\ \vdots \\ (\vec{y}_h)_{N-1} \\ (\vec{y}_h)_N \end{pmatrix} = \begin{pmatrix} \mu_1 \\ f(x_1) \\ f(x_2) \\ \vdots \\ \vdots \\ f(x_{N-1}) \\ \mu_2 \end{pmatrix}.$$

9.11. Legyen L egy függvényekhez függvényt rendelő operátor! Pontosabban $[0,l]$ intervallumon értelmezett függvényhez $[0,l]$ intervallumon értelmezett függvényt rendeljen hozzá. Ekkor

$$L : C^2[0,l] \to C(0,l) \cap C[\{0.l\}].$$

Azaz az L operátor egy tetszőleges $w \in C^2[0,l]$ függvény esetén az alábbi módon hat:

$$(Lw)(x) = \begin{cases} -w''(x) + c(x)w(x), & x \in (0,l) \\ w(x), & x \in \{0,l\}. \end{cases}$$

Ha feltesszük, hogy $f \in C[0,l]$, akkor a jobb oldal és peremértékek az alábbi alakban írhatóak:

$$\tilde{f}(x) = \begin{cases} f(x), & x \in (0,l) \\ \mu_1, & x = 0 \\ \mu_2, & x = l. \end{cases}$$

Azaz az eredeti feladat $Lu = \tilde{f}$ operátoregyenletes alakban írható.

9.13. A 9.12. feladat diszkretizációjából származó A_h együtthatómátrix az alábbi:

$$A_h = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ \frac{-1}{h^2} & \frac{2}{h^2} & \frac{-1}{h^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{-1}{h^2} & \frac{2}{h^2} & \frac{-1}{h^2} & 0 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{-1}{h^2} & \frac{2}{h^2} & \frac{-1}{h^2} & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}.$$

Ismeretes, hogy egy $B \in \mathbb{R}^{s\times s}$ mátrix M-mátrix, ha

- $b_{ij} \leq$ tetszőleges $i \neq j$-re,
- létezik olyan pozitív $g \in \mathbb{R}^s$ vektor, amelyre Bg is pozitív vektor.

Ellenőrizzük ezen feltételek teljesülését az A_h mátrix esetében! A mátrix előjelstruktúrája megfelelő. Célunk egy olyan pozitív $g_h \in \mathbb{R}^{N+1}$ vektor megadása, amelyre $A_h g_h$ is pozitív vektor lesz.

Ha $c(x) \geq c_0 > 0$, akkor A_h egy szigorúan diagonálisan domináns mátrix lesz. Ezért ebben az esetben a $g_h = [1, \dots, 1]^T$ vektor egy jó választás.

Tegyük fel, hogy $c(x) \geq 0$. Legyen $(g_h)_i = 1 + ih(l - ih),\ i = 0, \dots, N,\ h = l/N$! Ekkor igaz, hogy $i = 1, \dots, N-1$-re $(g_h)_i \geq 1$ és $(g_h)_0 = (g_h)_N = 1$. Továbbá könnyen látható, hogy

$$-(g_h)_{i+1} + 2(g_h)_i - (g_h)_{i-1} = 2h^2,\ i = 1, \dots, N-1.$$

A fenti összefüggés felhasználásával kapjuk, hogy

$$(A_h g_h)_i \geq \begin{cases} 2, & i = 1, \dots, N-1 \\ 1, & i = 0, N. \end{cases}$$

Azaz az $(A_h g_h)_i \geq 1$ minden $i = 0, \dots, N$-re. Ez pontosan azt jelenti, hogy a bevezetett g_h majoráló vektorral A_h egy M-mátrix.

9.15. A numerikus módszer megadásához definiálunk egy rácshálót. Legyen $\omega_h \subset [0.l]$ a h lépésközű ekvidisztáns rácsháló:

$$\omega_h = \{x_i = ih,\ i = 1, 2, \ldots, N-1,\ h = l/N\}.$$

Az intervallum két végpontját hozzávéve:

$$\overline{\omega}_h = \{x_i = ih,\ i = 0, 1, \ldots, N,\ h = l/N\}.$$

Tekintsük az eredeti feladatot az ω_h rácsháló pontjaiban. Ekkor

$$u''(x_i) + a(x_i)u'(x_i) + b(x_i)u(x_i) = f(x_i),\ x_i \in \omega_h$$

$$u(x_0) = \mu_1,\ u(x_N) = \mu_2.$$

Jelölje $\mathbb{F}(\overline{\omega}_h)$ az $\overline{\omega}_h$-n értelmezett függvények vektorterét és legyen y_h egy adott rácsfüggvény. Az első és a második deriváltat a standard másodrendű differenciahányadossal helyettesítve az eredeti feladat az alábbi alakba írható át:

$$\frac{y_h(x_i+h) - 2y_h(x_i) + y_h(x_i-h)}{h^2} + a(x_i)\frac{y_h(x_i+h) - y_h(x_i-h)}{2h} + b(x_i)u(x_i) = f(x_i),$$

ha $x_i \in \omega_h$, és

$$y_h(x_0) = \mu_1,\ y_h(x_N) = \mu_2.$$

Mivel az y_h rácsfüggvény és a jobboldali vektor is azonosítható egy $\mathbb{R}^{N+1}$-beli vektorral, nevezetesen:

$$\vec{y}_h \in \mathbb{R}^{N+1}:\ (\vec{y}_h)_i = y_h(x_i),\ a_i = a(x_i),\ b_i = b(x_i),\ \textit{illetve}\ (\vec{f}_h)_i = f(x_i),\ x_i \in \overline{\omega}_h.$$

Így az eredeti feladat egy $A_h\vec{y}_h = \vec{f}_h$ lineáris algebrai egyenletrendszerként írható fel, ahol az $A_h \in \mathbb{R}^{(N+1)\times(N+1)}$ együtthatómátrix az alábbi:

$$\begin{pmatrix}
1 & 0 & 0 & 0 & \ldots & 0 & 0 \\
\frac{1}{h^2} - \frac{a_1}{2h} & -\frac{2}{h^2} + b_1 & \frac{1}{h^2} + \frac{a_1}{2h} & 0 & 0 & \ldots & 0 \\
0 & \frac{1}{h^2} - \frac{a_2}{2h} & -\frac{2}{h^2} + b_2 & \frac{1}{h^2} + \frac{a_2}{2h} & 0 & \ldots & 0 \\
\vdots & & \ddots & \ddots & \ddots & & \vdots \\
\vdots & & & \ddots & \ddots & \ddots & \vdots \\
0 & \ldots & 0 & \frac{1}{h^2} - \frac{a_{N-1}}{2h} & -\frac{2}{h^2} + b_{N-1} & \frac{1}{h^2} + \frac{a_{N-1}}{2h} & 0 \\
0 & 0 & 0 & \ldots & 0 & 0 & 1
\end{pmatrix}.$$

9.16. A $[0, 1]$ intervallum esetén $h = 1/5$ lépésköz mellett az alábbi rácshálót definiáljuk:

$$\overline{\omega}_h = \{x_i = ih,\ i = 0, 1, \ldots, 5,\ h = 1/5\} = \{0, 0.2, 0.4, 0.6, 0.8, 1\}.$$

Azaz a lineáris algebrai egyenletrendszer mátrixának mérete 6×6. Határozzuk meg az együtthatómátrixot! A feladat kitűzése alapján $a(x) = x$, $b(x) = x^2$. Felhasználva a 9.15. feladatban leírt véges differenciás közelítéseket, valamint az a_i és b_i értékek meghatározásához szükséges ismereteket a keresendő A_h együtthatómátrix az alábbi:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 24.5 & -49.96 & 25.5 & 0 & 0 & 0 \\ 0 & 24 & -49.84 & 26 & 0 & 0 \\ 0 & 0 & 23.5 & -49.64 & 26.5 & 0 \\ 0 & 0 & 0 & 23 & -49.36 & 27 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

A feladat szerint $f(x) = -10x$. Ekkor a neki megfelelő jobb oldali vektor a peremértékkel együtt $\vec{f_h} = [1\ 2\ 4\ 6\ 8\ 2]^T$. Ekkor a numerikus megoldást $\vec{y_h} = A_h^{-1}\vec{f_h}$ alakban kapjuk. Ennek értéke:

$$\vec{y_h} = [0.9999\ 0.9280\ 0.9358\ 1.0910\ 1.44032.0000]^T.$$

Azaz a közelítő megoldás értéke az $x = 0.8$ pontban 1.4403.

A kézzel kiszámított A_h együtthatómátrixot és $\vec{f_h}$ jobb oldali vektort, valamint a belőlük származtatható $\vec{y_h}$ megoldásvektort a MATLAB-ban az alábbi módon írhatjuk be, illetve számíthatjuk ki:

```
>> A_h=[1 0 0 0 0 0; 24.5 -49.96 25.5 0 0 0;
0 24 -49.84 26 0 0; 0 0 23.5 -49.64 26.5 0;
0 0 0 23 -49.36 27; 0 0 0 0 0 1];
>> f_h=[1 2 4 6 8 2]';
>> y_h=A_h\f_h

y_h =

   0.999999999999997
   0.927977802490991
   0.935755725978431
   1.091023004730047
   1.440306505445524
   2.000000000000000
```

9.17. A 9.15. feladatban ismertetett eljárást kellene beprogramozni és a konkrét feladatra alkalmazni. A megírt kpep2.m fájl forráskódja az alábbi:

```
function [y_h]=kpep2(a,b,alpha,beta,N)
%% Kétpontos peremérték-feladat megoldása
%
%     u''(t)+c(t)u(t)+d(t)u'(t)=f(t) c\in\mathbb{R}, f(t)\inC[a,b]
%     u(a)=\alpha u(b)=\beta
%
```

```
%% Bemen■ paramatérek listája:

%     a    intervallum kezdete
%     b    intervallum vége
%     N    intervallumok száma


%% El■készületek

% Lépésköz

h=(b-a)/N;

% A diszkretizáló mátrix összerakása

for i=1:N-1
    c(i)=c1(a+i*h);
end

for i=1:N-1
    d(i)=d1(a+i*h);
end
e=ones(N-1,1);
A_h=(1/h^2)*spdiags([e-0.5*h*d' -2*e+h^2*c' e+0.5*h*d'],-1:1,N-1,N-1);
%% A numerikus megoldás meghatározása és plottolása

b_h=zeros(N-1,1);
b_h(1)=f(a+h)-alpha*(1/h^2-d(1)/(2*h));
b_h(N-1)=f(a+(N-1)*h)-beta*(1/h^2+d(1)/(2*h));
for i=2:N-2
    b_h(i)=f(a+i*h);
end
y=A_h\b_h;
y_i=linspace(alpha,beta,N+1);
y_i(2:N)=y;
y_h=y_i';

x_i=a:h:b;
plot(x_i,y_i,'r+')
hold on;
```

```
%% Az eredeti feladat jobb oldala
function ered=f(t)
ered=0;
%% Az eredeti feladat baloldalának c(t) függvénye
function ered2=c1(t)
ered2=t*cos(t);
%% Az eredeti feladat baloldalának d(t) függvénye
function ered2=d1(t)
ered2=0;
```

A programot a feladat adait, kérését figyelembe véve az alábbi paraméterekkel futtatjuk le, úgy, hogy a forráskódban egy *y_h(98)* sort pluszban beírtunk:

```
>> [y_h]=kpep2(0,1,0,1,100)
```

A feladat véges differenciás közelítő értéke az $x = 0.98$ pontban 0.983433. Más feladat esetén a paramétereket magától értetődő módon lehet változtatni.

9.21. A feladatban szereplő két pontos peremérték-feladat könnyen integrálható, ezért a feladat megoldása közvetlenül kiszámítható:

$$Y(x) = \frac{gx}{2v^2}(L - x).$$

Ezért a kilövés α szögét a

$$\tan\alpha = Y'(0) = \frac{gL}{2v^2}$$

összefüggésből határozhatjuk meg. A belövéses módszer programjának megírásához tanulmányozzuk a jegyzet 10.4.1.-es fejezetét.

A feladatra megírt programok az agyu.m és a belovesesmodszer.m fájlok. A módszer bemenő paramétere a kezdetiérték-feladatot explicit Eulerrel megoldó módszer h lépésköze lesz. A programban megadhatjuk továbbá az L intervallum végpontjának értékét, az yL végpontban felvett értéket és a v konstans sebességi értéket is.

A feladatban megadott h lépésközök mellett válasszuk meg a fenti paramétereket például az alábbi módon: $L = 10$, $yL = 0$ és $v = 1$. Ezek az értékek a 10 méterre becsapódó egységnyi sebességgel haladó ágyúgolyó kilövésének szögét adja vissza. A fenti összefüggés alapján a pontos értéke (a gravitációs állandó legyen $9.8m/s^2$):

$$\tan\alpha = Y'(0) = \frac{9.8\frac{m}{s^2} \cdot 10m}{2 \cdot \left(1 \cdot \frac{m}{s}\right)^2} = 49.$$

Adott h lépésközök mellett a program eredményét a pontos megoldással összevetve az alábbi hibaértékeket kapjuk:

h	$\lvert Y'(0) - Y'_{\text{beloveses}}(0)\rvert$
$1m$	$4.900 \cdot 10^{0}$
$0.1m$	$4.900 \cdot 10^{-1}$
$0.01m$	$4.900 \cdot 10^{-2}$
$0.001m$	$4.900 \cdot 10^{-3}$

10.13. táblázat. Hibaértékek adott lépésköz mellett.

A program eredményei az első három h lépésközre vonatkozóan az $Y'(0) = 20,\ 30,\ 70$ szemléltető próbálkozások mellett a 10.10 ábrán láthatók. Az ábrákon a piros értékek rendre azt mondják meg, hogy az $Y(40) = 0$ második peremfeltételt a belövéses módszer mely kezdeti értékre cseréli le az ágyúgolyó feladat megoldásához. Az adott h értékekre ez az alábbi lesz:

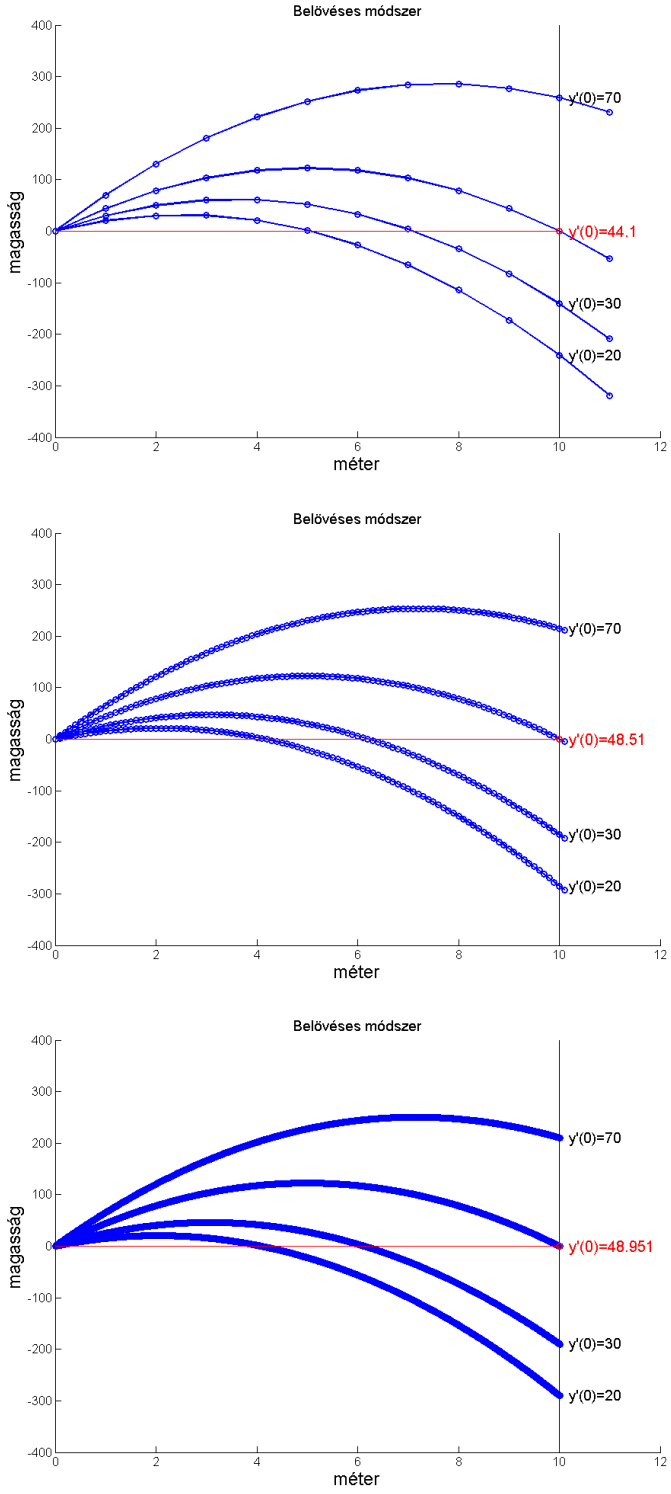
h	$Y'(0)$
$1m$	$4.900 \cdot 10^{0}$
$0.1m$	$4.900 \cdot 10^{-1}$
$0.01m$	$4.900 \cdot 10^{-2}$
$0.001m$	$4.900 \cdot 10^{-3}$

10.14. táblázat. A belövéses módszer $Y'(0)$ kezdeti érték javaslatai az $Y(40) = 0$ második peremfeltétel helyett adott h lépésköz mellett.

Megjegyzendő, hogy a belövéses módszer $h(c) = 0$ (lásd jegyzet 10.4.1.-es fejezet) egyenletének megoldására a megírt program a szelőmódszert alkalmazza. A belovesesmodszer.m fájl idevágó részlete:

```
%% Szel■módszer a gyökkereséshez
    function x = szelomodszer(x1,x2,tol,yL)
    y1 = agyu(x1)-yL;
    y2 = agyu(x2)-yL;
    while abs(x2-x1)>tol
fprintf('(%g,%g) (%g,%g)', x1, y1, x2, y2);
x3 = x2-y2*(x2-x1)/(y2-y1);
y3 = agyu(x3)-yL;
x1 = x2;
```

```
y1 = y2;
x2 = x3;
y2 = y3;
    end
    x = x2;
    return;
```



10.10. ábra. A belövéses módszer eredménye $h = 1,\ h = 0.1$ és $h = 0.01$ esetekben.

Parciális differenciálegyenletek

Elméleti feladatok

10.1. Tekintsük a kétváltozós, másodrendű, lineáris parciális differenciálegyenlet főrészének alakját $\Omega \subset \mathbb{R}^2$ tartományon:

$$(Lu)(x,y) = a(x,y)\frac{\partial^2 u(x,y)}{\partial x^2} + 2b(x,y)\frac{\partial^2 u(x,y)}{\partial x \partial y} + c(x,y)\frac{\partial^2 u(x,y)}{\partial y^2},$$

ahol a, b, c együtthatófüggvények. Azt mondjuk, hogy az L operátor

- *elliptikus típusú* az $(x,y) \in \Omega$ pontban, ha $a(x,y)c(x,y) - b^2(x,y) > 0$,
- *parabolikus típusú* az $(x,y) \in \Omega$ pontban, ha $a(x,y)c(x,y) - b^2(x,y) = 0$,
- *hiperbolikus típusú* az $(x,y) \in \Omega$ pontban, ha $a(x,y)c(x,y) - b^2(x,y) < 0$.

Azt mondjuk, hogy elliptikus (parabolikus, hiperbolikus) típusú az $\Omega_1 \subset \Omega$ tartományon, ha elliptikus (parabolikus, hiperbolikus) típusú az Ω_1 tartomány mindegyik pontjában.

Ezen definíciók mellett vizsgáljuk meg a konkrét feladat $\mathbb{R}^2$ egyes részein milyen típusú. Esetünkben $a(x,y) = x,\ b(x,y) = 0$ és $c(x,y) = y$. A definíció értelmében az operátor típusát xy előjele határozza meg.

Nevezetesen, ha

(a) $(x,y) \in \mathbb{R}^+ \times \mathbb{R}^+$, akkor L elliptikus ezen a tartományon,

(b) $(x,y) \in \mathbb{R}^+ \times \mathbb{R}^-$ vagy $(x,y) \in \mathbb{R}^- \times \mathbb{R}^+$, akkor L hiperbolikus ezen a tartományon,

(c) x vagy y valamelyik értéke 0, akkor L parabolikus típusú operátor az adott tartományon.

10.2. A 10.1. feladat gondolatmenetéhez hasonlóan állapítjuk meg, hogy a megadott L operátor $\mathbb{R}^2$ egyes részein milyen típusú.

A feladatunk esetében $a(x,y) = (x+y),\ b(x,y) = \sqrt{xy}$ és $c(x,y) = (x+y)$. A típusok meghatározása előtt érdemes megállapítanunk azt a tényt, hogy az L operátor csak $xy \geq$

0 esetén értelmes. Kibontva az $a(x,y)c(x,y)-b^2(x,y)$ alakot kapjuk, hogy $(x+y)^2-xy = x^2+xy+y^2$. Ekkor két eset lehetséges:

(a) $x^2+xy+y^2=0$,

(b) $x^2+xy+y^2>0$.

Az (a) eset csak $(x,y)=(0,0)$ pont esetén állhat fent. Azaz az origóban az L operátor parabolikus típusú.

A (b) eset az értelmezési tartomány figyelembe vételével $(xy \geq 0)$ pontosan akkor teljesül, ha x és y előjele megegyezik. Azaz az első és harmadik síknegyedben az L operátor elliptikus típusú.

10.4. Vegyük észre, hogy az u függvény $v=(1,-1)$ irányban vett iránymenti deriváltja 0, azaz

$$\Big(\frac{\partial u(x,y)}{\partial x}, \frac{\partial u(x,y)}{\partial x}\Big) \cdot v = \partial_v u(x,y) = 0.$$

Ez adja az alapötletünket arra vonatkozóan, hogy koordináta-transzformációt hajtsunk végre. Nevezetesen a fenti vektor iránya a koordinátarendszer 45°-os negatív irányú forgatását és kétszeres nyújtását motiválja. Ehhez térjünk át a (ξ,η) koordinátákra az alábbi módon:

$$\xi = x+y, \quad \eta = x-y.$$

Ekkor $u(x,y)=U(\xi,\eta)=U(\xi(x,y),\eta(x,y))$. Írjuk fel az eredeti egyenletet a bevezetett U függvény segítségével. Ehhez tekintsük:

$$u'(x,y) = U'(\xi,\eta) \cdot \begin{pmatrix} \frac{\partial \xi}{\partial x} & \frac{\partial \xi}{\partial y} \\ \frac{\partial \eta}{\partial x} & \frac{\partial \eta}{\partial y} \end{pmatrix} = \Big(\frac{\partial U(\xi,\eta)}{\partial \xi}, \frac{\partial U(\xi,\eta)}{\partial \eta}\Big) \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Továbbá fennáll, hogy

$$u'(x,y) = \Big(\frac{\partial u(x,y)}{\partial x}, \frac{\partial u(x,y)}{\partial x}\Big),$$

ekkor kapjuk, hogy:

$$\frac{\partial u(x,y)}{\partial x} = \frac{\partial U(\xi,\eta)}{\partial \xi} + \frac{\partial U(\xi,\eta)}{\partial \eta},$$
$$\frac{\partial u(x,y)}{\partial y} = \frac{\partial U(\xi,\eta)}{\partial \xi} - \frac{\partial U(\xi,\eta)}{\partial \eta}.$$

Így az eredeti egyenlet az új koordinátarendszerben az alábbi alakot ölti:

$$\frac{\partial u(x,y)}{\partial x} - \frac{\partial u(x,y)}{\partial y} = 0.$$

$$\Updownarrow$$

$$2\frac{\partial U(\xi,\eta)}{\partial \eta} = 0.$$

Ennek megoldása pedig $U(\xi,\eta) = C(\xi)$, azaz az eredeti feladat megoldása:

$$u(x,y) = C(x,y), \quad C \in C^1(\mathbb{R}).$$

10.6. Keressük a megoldást ún. szétválasztható alakban, azaz

$$u(x,y) = X(x)\cdot Y(y),$$

ahol $X \in C^2(\mathbb{R})$, $Y \in C^1(\mathbb{R})$. Továbbá $X(x)$ és $Y(y)$ nem az azonosan nulla függvény $\mathbb{R}$-en. Ezt behelyettesítve az eredeti egyenlet kapjuk, hogy

$$X''(x)Y(y) - X(x)Y'(y) = 0,$$

azaz

$$\frac{X''(x)}{X(x)} = \frac{Y'(y)}{Y(y)}.$$

Mivel mindkét oldal csak az adott változótól függ, ezért a fenti egyenlet megoldása egy $\lambda \in \mathbb{R}$ szám meghatározását jelenti. A bal oldal egy másodrendű, míg a jobb oldal egy elsőrendű állandó együtthatós közönséges differenciálegyenlet megoldását igényeli. Ezek általános megoldásait a karakterisztikus egyenletek gyökeivel határozhatjuk meg. Nevezetesen:

$$X(x) = \begin{cases} c_1 e^{\sqrt{\lambda}x} + c_2 e^{-\sqrt{\lambda}x}, \text{ ha } \lambda > 0, \ c_1, c_2 \in \mathbb{R}, \\ c_1 x + c_2, \text{ ha } \lambda = 0, \ c_1, c_2 \in \mathbb{R}, \\ c_1 \sin(\sqrt{|\lambda|}x) + c_2 \cos(\sqrt{|\lambda|}x), \text{ ha } \lambda < 0, \ c_1, c_2 \in \mathbb{R}. \end{cases}$$

$$Y(y) = ce^{\lambda y}, \ c \in \mathbb{R}.$$

Az így kapott $u(x,y) = X(x)Y(y)$ alakú függvények mind kielégítik az eredeti egyenletet. Fontos megjegyeznünk, a feladat nem állítja, hogy csak ilyen alakú megoldásai vannak a feladatnak. Például jó megoldás az $u(x,y) = x^3/6 + xy$, amely nem $X(x)Y(y)$ alakú.

Elliptikus és parabolikus feladatok megoldása véges differenciákkal

10.7. A diszkretizáció felírásához olvassuk át a példatárhoz tartozó jegyzet 11.2. Lineáris, másodrendű, elliptikus parciális differenciálegyenletek című részéből a 11.2.2 fejezetet!

Az $N_x = 3$ és $N_y = 2$ osztásrészek egyértelműen meghatározzák az egységnégyzet rácspontjainak számát, s így az együtthatómátrix méretét is.

A rácspontok száma $(N_x + 1)(N_y + 1) = 12$, míg az együtthatómátrix mérete 12×12. Könnyen meggondolható, hogy az x irányban a lépésköz $h_x = 1/(N_x + 1) = 1/4$, míg az y irányban $h_y = 1/(N_y + 1) = 1/3$. Bevezetve a $H_x = 1/(h_x)^2$ és a $H_y = 1/(h_y)^2$ jelöléseket az A_h diszkretizáló mátrix alakja az alábbi lesz:

$$A_h = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -H_y & 0 & 0 & -H_y & H_x + H_y & -H_x & 0 & 0 & -H_y & 0 & 0 \\ 0 & 0 & -H_y & 0 & 0 & -H_y & H_x + H_y & -H_x & 0 & 0 & -H_y & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

A mátrix struktúrájából jól kivehető, hogy a belső rácspontok (2db) közötti összefüggéseket a 6. és 7. sorok írják le. A többi sor a peremértékeket tárolják el.

10.8. A program megíráshoz olvassuk át a példatárhoz tartozó jegyzet 11.2. Lineáris, másodrendű, elliptikus parciális differenciálegyenletek című részéből a 11.2.5 fejezetet!

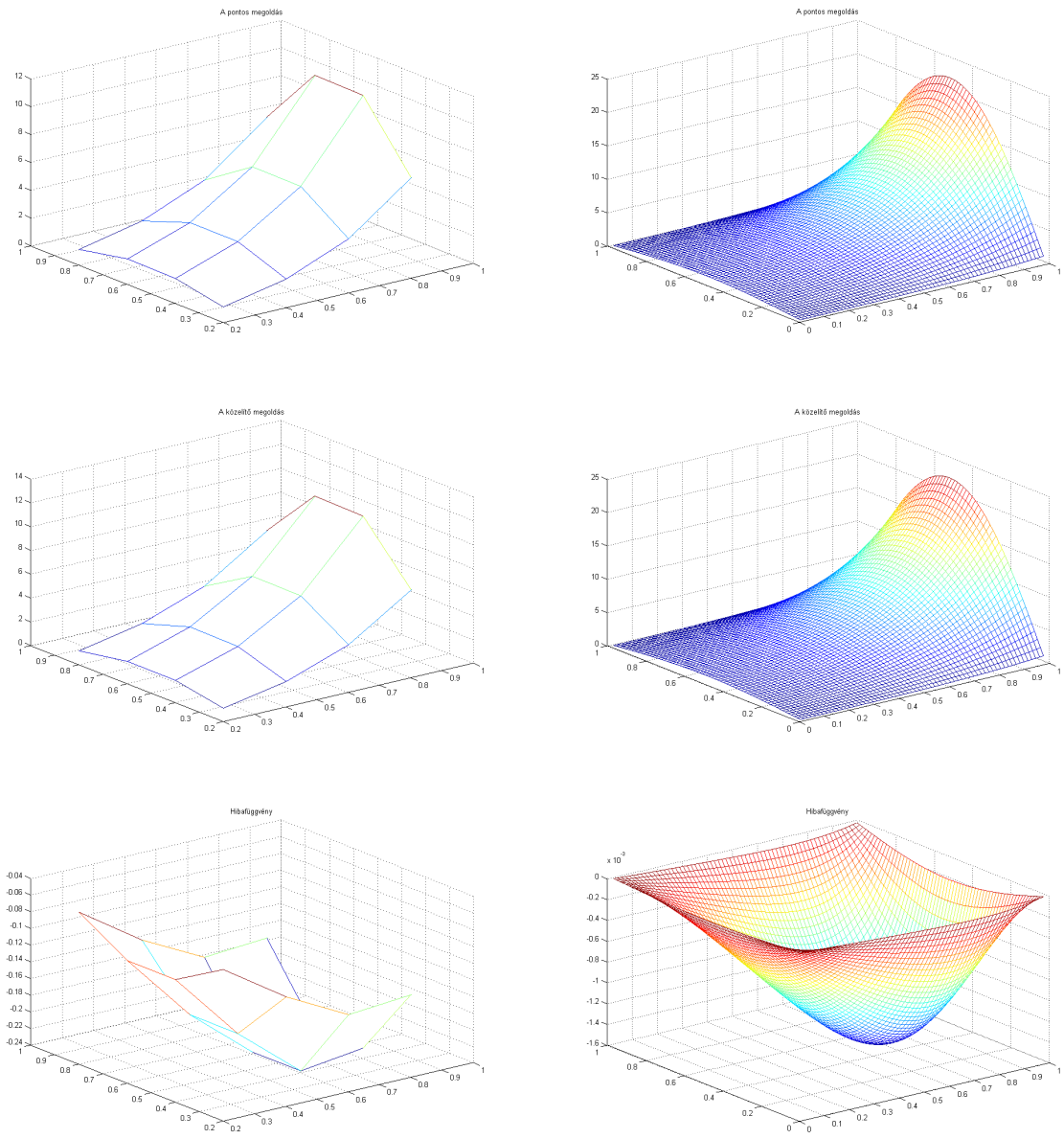
A megírt ellvdm1.m program bemenő paramétere a két irányban egyenlő részre történő osztások száma: n.

A módszer eredményének szemléltetésére kiragadjuk az $n = 4$ és $n = 64$ eseteket és rendre ábrázoljuk (10.11 ábra) a pontos, közelítő megoldásokat valamint a hiba nagyságát.

Ekkor az intervallumszámok növelésével a pontos és numerikus megoldás maximumnormájára a 10.15 táblázatban látható értékek figyelhetőek meg.

Az ellvdm1.m fájl forráskódja:

```
function [maximumnorma]=ellvdm1(n)
%% A Laplace u=x^2+y^2 egyenlet megoldása
%
% A feladatot az egységnégyzeten oldjuk meg az alábbi peremfeltétellel
%
```



10.11. ábra. A pontos, közelítő megoldások és a hibák nagysága $n = 4,\ n = 64$ esetekben.

```
% u(x,0)=0
% u(x,1)=x^2/2
% u(0,y)=sin(pi y)
% u(1,y)=e^(pi)sin(pi y)+y^2/2
%
% A feladat pontos megoldása: u(x,y)=e^(pi x)sin(pi y)+0.5x^2y^2.
```

n	maximumnorma értéke
4	$2.3746735250 \cdot 10^{-1}$
16	$2.3157280671 \cdot 10^{-2}$
64	$1.5958950751 \cdot 10^{-3}$
256	$1.0214327981 \cdot 10^{-4}$

10.15. táblázat. A maximumnorma értéke adott n részre történő osztás mellett.

A táblázat adataiból megállapítható az elméletből ismert tény, nevezetesen az, hogy a véges differenciás közelítés a maximumnormában másodrendű módszer.

```
%% A feladat bemen■ paraméterei
%
% n+1 - Az egy irányú intervallumok száma
%
% Megj.: Azaz (n)^2 besl■ pontom lesz.

%% A feladat kimen■ paraméterei
%
% A nemrikus megoldásvektor
% A maximumnormában mért hiba
% Ábra

%% A diszkretizációs mátrix felépítése
%
% Nagysága n^2, mert a perempontokat a jobb oldalban tároljuk majd el.

h=1/(n+1);

% F■átló
a1=-4*ones(n^2,1);
% F■átlóhoz legközelebbi fels■ átló
a21=sparse(1,1);
a22=ones(n^2-1,1);
for i=1:n-1;
    a22(i*(n),1)=0;
end
a2=[a21;a22];
% F■átlóhoz távolabbi fels■ átló
```

```
a3=ones(n^2,1);
% F■átlóhoz legközelebbi alsó átló
a42=ones(n^2-1,1);
for i=1:n-1;
    a42(i*(n),1)=0;
end
a41=sparse(1,1);
a4=[a42;a41];
% F■átlóhoz távolabbi alsó átló
a5=ones(n^2,1);

% A mátrix összerakása
V=[a1,a2,a3,a4,a5];
d=[0,1,n,-1,-n];
A=spdiags(V,d,n^2,n^2);
A_h=A'*(1/h^2);

%% A jobb oldali vektor felépítése

% Perem elkészítése
% Az u(x,1) perem
    for i=1:n
        g21(i)=(i/(n+1))^2/2;
    end
  g2=[zeros(1,n^2-n) g21]';
% Az u(0,y) perem
    for i=1:n;
        g31(1,1+(i-1)*n)=sin(pi*i/(n+1));
    end
  g3=[g31 zeros(1,n-1)]';
% Az u(1,y) perem
    for i=1:n;
        g1(1,i*n)=exp(pi)*sin(pi*i/(n+1))+((i/(n+1))^2)/2;
    end
% Megj.: Most az u(x,0) perem nullával egyenl■.

% Jobboldal elkészítése

F=[];
for i=1:n
    for j=1:n
```

```
        F(i,j)=(i/(n+1))^2+(j/(n+1))^2;
    end;
end;
nincsperem1=reshape(F,1,n^2);
nincsperem=nincsperem1';

f=nincsperem-g1'/(h^2)-g2/(h^2)-g3/(h^2);

%% A LAER megoldása, azaz y numerikus megoldás megadása
y=A_h\f;

%% A pontos megoldás betöltése
G=[];
for i=1:n
    for j=1:n
  G(i,j)=exp(pi*i/(n+1))*sin(pi*j/(n+1))+(1/2)*(i/(n+1))^2*(j/(n+1))^2;
    end;
end;
ered1=reshape(G,1,n^2);
ered=ered1';

%% Hibaszámítás és plottolás

maximumnorma=norm(ered-y,'inf');

% A pontos megoldás, a numerikus megoldás és a hiba kirajzolása
ugrid = reshape(ered,n,n);
mesh(h:h:n*h',h:h:n*h',ugrid')
title('A pontos megoldás')
pause
apgrid = reshape(y,n,n);
mesh(h:h:n*h',h:h:n*h',apgrid')
title('A közelít■ megoldás')
pause
errgrid = reshape(ered-y,n,n);
mesh(h:h:n*h',h:h:n*h',errgrid')
title('Hibafüggvény')
```

A MATLAB-ban például az $n = 4$ esetén a lenti parancsot beírva a 3 ábra mellett az alábbi maximumnorma értéket kapjuk vissza:

```
>> [maximumnorma]=ellvdm1(4)
```

```
maximumnorma =

    0.237467352500951
```

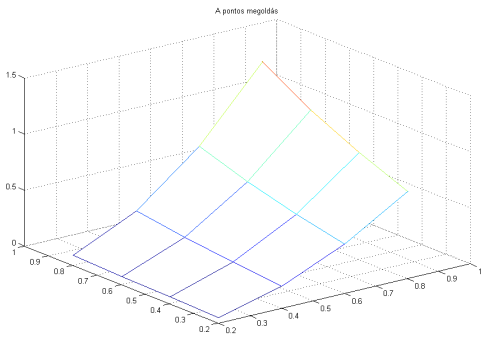
10.9. A feladat pontos megoldása $u(x, y) = x^2 e^y$. Ekkor a megírt ellvdm2.m program az intervallumszámok növelésével a pontos és numerikus megoldás maximumnormájára az alábbi értékeket adja vissza:

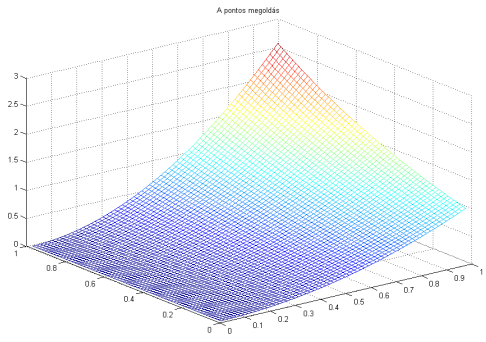
n	maximumnorma értéke
4	$1.3132312883 \cdot 10^{-4}$
16	$1.2514964504 \cdot 10^{-5}$
64	$8.6106173369 \cdot 10^{-7}$
256	$5.5104494078 \cdot 10^{-8}$

10.16. táblázat. A maximumnorma értéke adott n részre történő osztás mellett.

A táblázat adataiból megállapítható az elméletből ismert tény, nevezetesen az, hogy a véges differenciás közelítés a maximumnormában másodrendű módszer.

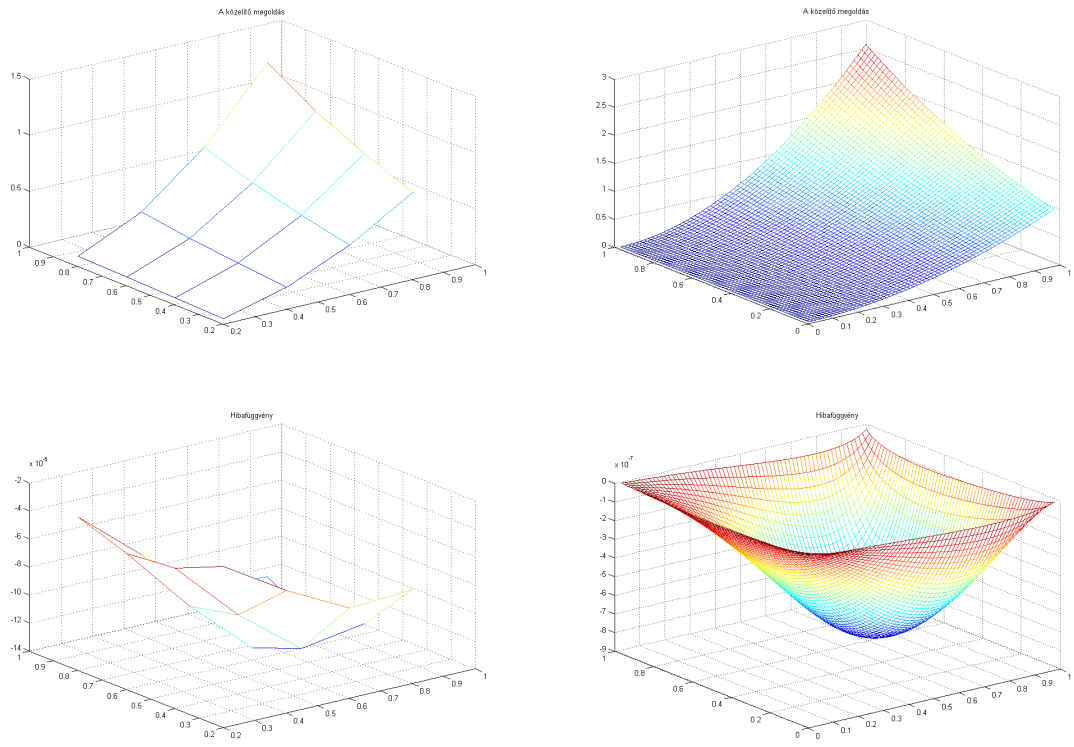
A módszer szemléltetésére ábrázoljuk az $n = 4$ és $n = 64$ eseteket.





Az *ellvdm2.m* fájl forráskódja:

```
function [maximumnorma]=ellvdm2(n)
%% A Laplace u=e^y(x^2+2) egyenlet megoldása
%
```



10.12. ábra. A pontos, közelítő megoldások és a hibák nagysága $n = 4,\ n = 64$ esetekben.

```
% A feladatot az egységnégyzeten oldjuk meg az alábbi peremfeltétellel
%
% u(x,0)=x^2
% u(x,1)=ex^2
% u(0,y)=0
% u(1,y)=e^(y)
%
% A feladat pontos megoldása: u(x,y)=x^2e^y.

%% A feladat bemen■ paraméterei
%
% n+1 - Az egy irányú intervallumok száma
%
% Megj.: Azaz (n)^2 besl■ pontom lesz.

%% A feladat kimen■ paraméterei
%
```

```
% A nemrikus megoldásvektor
% A maximumnormában mért hiba
% Ábra

%% A diszkretizációs mátrix felépítése
%
% Nagysága n^2, mert a perempontokat a jobb oldalban tároljuk majd el.

h=1/(n+1);

% F■átló
a1=-4*ones(n^2,1);
% F■átlóhoz legközelebbi fels■ átló
a21=sparse(1,1);
a22=ones(n^2-1,1);
for i=1:n-1;
    a22(i*(n),1)=0;
end
a2=[a21;a22];
% F■átlóhoz távolabbi fels■ átló
a3=ones(n^2,1);
% F■átlóhoz legközelebbi alsó átló
a42=ones(n^2-1,1);
for i=1:n-1;
    a42(i*(n),1)=0;
end
a41=sparse(1,1);
a4=[a42;a41];
% F■átlóhoz távolabbi alsó átló
a5=ones(n^2,1);

% A mátrix összerakása
V=[a1,a2,a3,a4,a5];
d=[0,1,n,-1,-n];
A=spdiags(V,d,n^2,n^2);
A_h=A'*(1/h^2);

%% A jobb oldali vektor felépítése

% Perem elkészítése
```

```
% Az u(x,0) perem
    for i=1:n
        g31(i)=(i/(n+1))^2;
    end
  g3=[g31 zeros(1,n^2-n)]';
% Az u(x,1) perem
    for i=1:n
        g21(i)=exp(1)*(i/(n+1))^2;
    end
  g2=[zeros(1,n^2-n) g21]';
% Az u(1,y) perem
    for i=1:n;
        g1(1,i*n)=exp(i/(n+1));
    end
% Megj.: Most az u(0,y) perem nullával egyenl■.

% Jobboldal elkészítése

F=[];
for i=1:n
    for j=1:n
        F(i,j)=exp(j/(n+1))*((i/(n+1))^2+2);
    end;
end;
nincsperem1=reshape(F,1,n^2);
nincsperem=nincsperem1';

f=nincsperem-g1'/(h^2)-g2/(h^2)-g3/(h^2);

%% A LAER megoldása, azaz y numerikus megoldás megadása
y=A_h\f;

%% A pontos megoldás betöltése
G=[];
for i=1:n
    for j=1:n
  G(i,j)=(i/(n+1))^2*exp(j/(n+1));
    end;
end;
ered1=reshape(G,1,n^2);
ered=ered1';
```

```
%% Hibaszámítás és plottolás

maximumnorma=norm(ered-y,'inf');

% A pontos megoldás, a numerikus megoldás és a hiba kirajzolása
ugrid = reshape(ered,n,n);
mesh(h:h:n*h',h:h:n*h',ugrid')
title('A pontos megoldás')
pause
apgrid = reshape(y,n,n);
mesh(h:h:n*h',h:h:n*h',apgrid')
title('A közelít■ megoldás')
pause
errgrid = reshape(ered-y,n,n);
mesh(h:h:n*h',h:h:n*h',errgrid')
title('Hibafüggvény')
```

A MATLAB-ban például $n = 64$ esetén a lenti parancsot beírva a 3 ábra mellett az alábbi maximumnorma értéket kapjuk vissza:

```
>> [maximumnorma]=ellvdm2(64)

maximumnorma =

    8.610617336923809e-007
```

10.10. Tanulmányozzuk alaposan a jegyzetben található forráskódot! Ekkor a numerikus megoldás előállításához az alábbi sorokat kell megváltoztatnunk:

```
init=exp(x);\\ bdry=[exp(t) exp(1+t)];
```

A feladat pontos megoldása az $u(x,t) = e^{x+t}$ függvény. Ekkor a pontos megoldást a forráskódban az alábbi módon írhatjuk be:

```
upontos=zeros(N,J);
for i=1:N
for n=1:J
upontos(i,n)=exp(x(i)+t(n));
end
end
```

A hibafüggvény ábrájának megjelenítéséhez cseréjük le a

```
%hibamatrix=upontos-appgrig;
```

sort az alábbira:

```
hibamatrix=upontos-appgrig;
```

Ahhoz, hogy a kívánt tartományon tudjuk futtatni a programot az endx, endt értékeket rendre 1,1-nek kell megválasztanunk a heatexp(endx,endt,Nx,q) függvény futtatása során.

10.11. A feladatot megoldó parvdm.m fájl forráskódja az alábbi:

```
function parvdm(a,b,n,T,r,theta)
%% Reakció-diffúzió feladat 1D-ben adott kezdeti és Dir. feltétellel
%
%       d_t u(t,x)=d_xx u(t,x), x\in[a,b], t\in[0,T]
%           u(t,a)=u(t,b)=0 Dirichlet peremfeltétel
%           u(0,x)=u0 kezdeti feltétel
%
%% Bemen■ paraméterek listája
%
%       a   Az intervallum kezd■pontja
%       b   Az intervallum végpontja
%       n   A rács bels■ pontjainak száma
%       T   Az id■intervallum végpontja
%       r   A delta/h^2 értéke (EE esetén stabilitáshoz r=<0.5 kell)
%   theta   A theta-módszer paramétere (=0 EE =1 IE =0.5 CN)

%% El■készületek

h=(b-a)/(n+1);
x=h*1:n;
delta=r*h^2;
kmax=round(T/delta);
u0=sin(x);

%% A lépésmátrix konstrukciója

N=sparse(2:n,1:n-1,ones(n-1,1),n,n);
I=speye(n);
```

```
Q=-2*speye(n)+N+N';
Q1=I-theta*((delta/h^2)*Q);
Q2=I+(1-theta)*((delta/h^2)*Q);
kezdeti=u0';

%% Az egyes id■lépések plottolása
for k=1:kmax
kezdeti=Q1\(Q2*kezdeti);
plot(a:h:b,[0,kezdeti',0],'bo')
axis([a,b,-1.2,1.2])
xlabel('x','FontSize',14)
ylabel('u(t,x)','FontSize',14)
title(['A t= ',num2str(k*delta,'%2.4f id■pillanatban a numerikus megoldás')
    ],'Color','r','FontSize',14)
pause(delta*50)
end;
```

A program a megoldás időbeli fejlődését mutatja be. Például a 10.11. hővezetési feladatot a $[0,1]$ időintervallumon, a $[0,\pi]$ intervallumon 9 rácsponttal, $r = 0.4$ hányadossal, Crank–Nicolson-módszerrel megoldó feladatot az alábbi módon írhatjuk be:

```
>> parvdm(0,pi,9,1,0.4,0.5)
```

Irodalomjegyzék

[1] Asher, U., Petzold, L. Computer Methods for Ordinary Differential Equations. SIAM, Philadelphia, 1998.

[2] Ayyub, B. M., McCuen, R. H., Numerical Methods for Engineers, Prentice Hall, Upper Saddle River, NJ, 1996.

[3] Burden, R.L., Burden, Faires, J.D., Numerical Analysis, PWS-Kent, fifth edition, 1993.

[4] Faragó I., Horváth R., *Numerikus módszerek*, 2011.

[5] Jain, M. K., Ivengar, S. R. K., Jain, R. K., Numerical Methods, Problems and Solutions, New Age International Publishers, 1984.

[6] Kincaid, D., Cheney, W. Numerical Analysis. Mathematics of Scientific Computing, American Mathematical Society, 2009.

[7] Kreyszig, E., Advanced Engineering Mathematics, 9th edition, Wiley, 2006

[8] Lambert, J. D. Numerical Methods for Ordinary Differential Systems,Wiley Publ., 2000.

[9] Logan, J. D. A First Course in Differential Equations, Springer, 2006.

[10] Quarteroni, A., Sacco, R., Saleri, F., Numerical Mathematics, Springer, 2007

[11] Shampine, L. F., Gladwell, I., Thompson, S. Solving ODEs with MATLAB, Cambrigde University Press, Cambridge, 2003.

[12] Simon P., Tóth J. Differenciálegyenletek. Bevezetés az elméletben és az alkalmazásokba. TypoTech, Budapest, 2005.

[13] Stoyan G., Takó G. Numerikus módszerek 1,2,3. TypoTeX, Budapest, 1995.

[14] Stoer, J., Bulirsch, R., Introduction to Numerical Analysis. Heidelberg, Springer, 1980.