NUMERIKUS ANALÍZIS

Műszaki Földttudományi alapszak, levelező tagozatos hallgatók számára

Dr. Házy Attila

TARTALOMJEGYZÉK

1 HIBASZÁMÍTÁS

1.1 HIBAFORRÁSOK

A feladatok megoldása során különféle hibaforrásokkal találkozunk:

- **Modellhiba**, amikor a valóságnak egy közelítését használjuk a feladat matematikai alakjának felírásához. (Pl. egy fizikai törvényekkel leírt modellt.)
- **Mérési vagy öröklött hiba**, amikor a modell adatai a pontos értékeknek csak közelítő értékei. Általában a mérés pontosságától függnek.
- **Műveleti (kerekítési-) és input hiba**, amely az adatok számítógépen való ábrázolásából adódnak. A racionális számoknak is csak egy részhalmaza ábrázolható a lebegőpontos aritmetikában. A műveletvégzés során kerekítés, túl- illetve alulcsordulás léphet fel.
- **Képlethiba**, amikor egy végtelen eljárást véges számú lépés után leállítunk, közelítő algoritmusokat alkalmazunk.

1.2 SZÁMÁBRÁZOLÁS

1.2.1 AZ EGÉSZ SZÁMOK

Az egész számokat a számítógépben előjeles vagy előjel nélküli bináris számként képzelhetjük el, így jellemezhetőek a használt bináris jegyek számával. Ez utóbbi nem rögzített, hanem bizonyos mértékben választható. A szokásos az, hogy 2- és 4-byte-os egész számok állnak rendelkezésre, ahol a byte nyolc bitet tartalmaz, azaz nyolc bináris jeggyel rendelkezik (sok gépnél a byte a legkisebb elérhető, címezhető tárolási egység).

Az egész számokkal való aritmetikai műveletek nagyságrenddel gyorsabbak a lebegőpontos számokénál és hibamenteseknek tekinthetők, ezért használatuk döntő mértékben felgyorsíthatja egy adott algoritmus futását a számítógépen.

Az egész számokkal való számítás minden lépését viszont figyelmesen át kell gondolni, mert ilyenkor valójában maradékosztályokban dolgozunk.

1.2.2 A LEBEGŐPONTOS SZÁMOK

A számítógépek egy véges számhalmazt ábrázolnak és a számításokat is ezekkel a számokkal végzik. Leggyakrabban a lebegőpontos aritmetikát használják. Nézzük ennek a modelljét:

1.2.1. DEFINÍCIÓ. *A nemnulla lebegőpontos számok általános alakja:*

$$\pm a^k \left(\frac{m_1}{a} + \frac{m_2}{a^2} + \cdots + \frac{m_t}{a^t} \right),$$

ahol $a > 1$ a számábrázolás alapja, $\pm$ az előjel, $t > 1$ a számjegyek száma, $k \in \mathbb{Z}$ a kitevő.

Az m_1 számjegy normalizált, azaz $1 \leq m_1 \leq a-1$ (ez garantálja a számábrázolás egyértelműségét).

A többi számjegyre: $0 \leq m_i \leq a-1$ $(i = 2, ..., t)$

A nulla nem normalizált! Ebben az esetben $k = 0, m_1 = m_2 = \cdots = m_t = 0$, előjele általában $+$.

A számábrázolás alapja lehet $2, 10, 16, \ldots$ (általában a programozási nyelven múlik, hogy melyiket használja)

$t = 8$: egyszeres pontosság, $t = 16$: dupla pontosság.

A lebegőpontos számokat $[\pm, k, m_1, m_2, \ldots, m_t]$ alakban tároljuk (a valóságban ettől eltérhet...), ahol $m := (m_1, m_2, \ldots, m_t)$ a mantissza, míg k a szám karakterisztikája.

A géptől és a pontosságtól függően m tárolására 4,8,16 byte áll rendelkezésre. Ezzel párhuzamosan nő a k értékkészlete. Adott pontosság mellett:

$$L \leq k \leq U,$$

ahol $L < 0, U > 0$ és $|L| \approx |U|$.

A legnagyobb ábrázolható szám:

$$\begin{aligned} M^\infty &= a^U \cdot \sum_{i=1}^{t} \frac{a-1}{a^i} = a^U \cdot \left(\frac{a-1}{a} + \frac{a-1}{a^2} + \cdots + \frac{a-1}{a^t} \right) \\ &= a^U \left(\frac{a-1}{a} \cdot \frac{1 - \left(\frac{1}{a}\right)^t}{1 - \left(\frac{1}{a}\right)} \right) = a^U \left(\frac{a-1}{a} \cdot \frac{a^t - 1}{a^t} \cdot \frac{a}{a-1} \right) \\ &= a^U \left(1 - a^{-t}\right) \end{aligned}$$

A legkisebb ábrázolható szám: $-M^\infty$.

A lebegőpontos számok a $[-M^\infty, M^\infty]$-beli számok diszkrét (racionális) részhalmazát alkotják és ez a részhalmaz a 0-ra nézve szimmetrikus.

A 0-hoz legközelebbi pozitív lebegőpontos számot ε_0-val jelöljük.

$$\varepsilon_0 = a^L \left(\frac{1}{a} + 0 + 0 + \cdots + 0 \right) = a^{L-1}.$$

Így a 0-n kívül a $(-\varepsilon_0, \varepsilon_0)$ intervallumban nincs más lebegőpontos szám (lehetnek nem normalizált számok, de azokkal nem foglalkozunk).

Az ε_0-hoz legközelebbi pozitív lebegőpontos szám:

$$a^L \left(\frac{1}{a} + 0 + 0 + \cdots + \frac{1}{a^t} \right) = \varepsilon_0 + a^{L-t} = \varepsilon_0(1 + a^{1-t}).$$

Az 1 mindig lebegőpontos szám: $1 = [+, 1, 1, 0, 0, \ldots 0]$.

Az 1 után a $[+, 1, 1, 0, 0, \ldots 0, 1]$ lebegőpontos szám következik, ez $1 + a^{1-t} = 1 + \varepsilon_1$, ahol $\varepsilon_1 = a^{1-t}$.

1.2.2. DEFINÍCIÓ. *Ezt az ε_1-et a gép relatív pontosságának, vagy gépi epszilonnak nevezzük.*

Az $\varepsilon_0, \varepsilon_1$ számok abszolút és relatív hibakorlátot jelentenek az inputnál és a négy alapműveletnél.

Legyen adott

$$0 < x = [+, k, m] = a^k \left(\frac{m_1}{a} + \frac{m_2}{a^2} + \cdots + \frac{m_t}{a^t} \right) < M^\infty$$

Az x-hez legközelebb eső, x-nél nagyobb lebegőpontos szám: $x + a^{k-t}$, ugyanis

$$\overline{x} = x + a^k \left(0 + 0 + 0 + \cdots + \frac{1}{a^t} \right) = x + a^{k-t},$$

tehát $\delta_x = \overline{x} - x = a^{k-t}$. Mivel k karakterisztikájú számok közül a legkisebb lehetséges érték $a^k \cdot \dfrac{1}{a}$, ezért (mivel $a^{k-1} \leq x$)

$$\delta_x = \overline{x} - x = a^{k-t} = a^{k-1+1-t} = a^{k-1} \cdot a^{1-t} = a^{k-1} \cdot \varepsilon_1 \leq x \cdot \varepsilon_1,$$

1.2.3 KEREKÍTÉS, LEVÁGÁS

Az Input hibája Legyen $x \in \mathbb{R}$, $|x| \leq M^\infty$ és legyen $\mathrm{fl}(x)$ az x-hez hozzárendelt lebegőpontos szám (ez lehet kerekítéssel vagy levágással).

Kerekítés esetén:

$$\text{fl}(x) = \begin{cases} 0, & \text{ha } |x| < \varepsilon_0 \\ \text{az } x\text{-hez legközelebbi lebegőpontos szám}, & \text{ha } \varepsilon_0 \le |x| \le M^\infty \end{cases}$$

Ekkor kerekítés esetén

$$|\,\text{fl}(x) - x| \le \begin{cases} \varepsilon_0, & \text{ha } |x| < \varepsilon_0 \\ \frac{1}{2}\varepsilon_1|x|, & \text{ha } |x| \ge \varepsilon_0 \end{cases}$$

Levágás esetén $\frac{1}{2}\varepsilon_1|x|$ helyett $\varepsilon_1|x|$ áll (ez pontatlanabb, de könnyebb levágni, mint kerekíteni).

Alapműveletek hibája Legyen a $\Diamond$ az alapműveletek bármelyike $(+, -, \cdot, /)$. Ekkor kerekítés esetén

$$|\,\text{fl}(x \Diamond y) - x \Diamond y| \le \begin{cases} \varepsilon_0, & \text{ha } |x \Diamond y| < \varepsilon_0 \\ 1/2 \cdot \varepsilon_1 \cdot |x \Diamond y|, & \text{ha } |x \Diamond y| \ge \varepsilon_0 \end{cases}$$

vagy az ε_0-lal kapcsolatos eseteket elhagyva:

$$|\,\text{fl}(x \Diamond y) - x \Diamond y| \le \varepsilon_1 |x \Diamond y| \begin{cases} 1, & \text{levágás esetén} \\ 1/2, & \text{kerekítés esetén} \end{cases}$$

Levágás esetén:

$$-\varepsilon_1|x \Diamond y| \le \text{fl}(x \Diamond y) - x \Diamond y \le \varepsilon_1|x \Diamond y|$$

ebből adódik,hogy

$$\text{fl}(x \Diamond y) - x \Diamond y = \varepsilon_1 \cdot |x \Diamond y| \cdot s \qquad \text{ahol } -1 \le s \le 1.$$

Ekkor viszont

$$\text{fl}(x \Diamond y) = x \Diamond y + \varepsilon_1 \cdot |x \Diamond y| \cdot s = x \Diamond y(1 + \text{sgn}(x \Diamond y) \cdot \varepsilon_1 \cdot s)$$

Legyen $\varepsilon_\Diamond := \text{sgn}(x \Diamond y) \cdot \varepsilon_1 \cdot s \le \varepsilon_1$. Ekkor

$$\text{fl}(x \Diamond y) = x \Diamond y(1 + \varepsilon_\Diamond),$$

ahol

$$|\varepsilon_\Diamond| \le \varepsilon_1 \cdot \begin{cases} 1, & \text{levágás esetén} \\ 1/2, & \text{kerekítés esetén} \end{cases}$$

Ez az összefüggés a 0 körüli hézagban nem érvényes! Továbbá akkor sem, ha a művelet eredménye $> M^\infty$ (azaz túlcsordulás esetén)

Ha a művelet eredménye $\neq 0$, de eleme a $(-\varepsilon_0, \varepsilon_0)$ intervallumnak, akkor alulcsordulást kapunk (általában 0-nak veszi a gép hibajelzés nélkül!)

1.3 KLASSZIKUS HIBAANALÍZIS

1.3.1. DEFINÍCIÓ. *Legyen A pontos érték, a pedig annak valamilyen közelítése. A $\Delta a = A - a$ mennyiséget az a közelítés hibájának nevezzük, a $|\Delta a| = |A - a|$ számot pedig az abszolút hibájának. Azt a δa értéket pedig, amelyre fennáll, hogy $|A - a| = |\Delta a| \leq \delta a$, az a abszolút hibakorlátjának mondjuk.*

A definíció értelmében használjuk az $A = a \pm \delta a$ hivatkozást is, ami annyit jelent, hogy $A \in [a - \delta a, a + \delta a]$. Nyilván annál jobb a közelítés, más szóval annál élesebb a becslés (és erre törekedni kell), minél kisebb a δa.

A közelítés jóságát ezért az abszolút hiba és az abszolút hibának a pontos érték egységére eső része – a relatív hiba – együtt jellemzi.

1.3.2. DEFINÍCIÓ. *Az A szám valamely a közelítő értékének relatív hibája a $\frac{\delta a}{|A|}$ mennyiség.*

Minthogy az A pontos érték általában nem ismeretes, ezért a $\frac{\delta a}{|A|}$ helyett a $\frac{\delta a}{|a|}$ közelítést használjuk. Az így elkövetett hiba mértéke:

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

Szokás a relatív hiba helyett annak százalékos érték ét megadni, azaz $\frac{\delta a}{|A|} \Leftrightarrow 100\frac{\delta a}{|A|}$

Jelölések:

A következő jelöléseket és elnevezéseket használjuk: x, y pontos értékek, a és b a közelítő értékeik, δa és δb hibakorlátokkal, azaz $|x - a| = |\Delta a| \leq \delta a$ és $|y - b| = |\Delta b| \leq \delta b$.

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

$$|\Delta (a \Diamond b)| \leq \delta (a \Diamond b),$$

illetve a

$$\frac{|\Delta (a \Diamond b)|}{|(x \Diamond y)|} \leq \frac{\delta (a \Diamond b)}{|(x \Diamond y)|} \approx \frac{\delta (a \Diamond b)}{|(a \Diamond b)|}$$

becsléseket keressük, ahol $\Delta (a \Diamond b)$ a művelet hibáját, $\delta(a \Diamond b)$ pedig abszolút hibakorlátját jelöli. Az additív műveletek (összeadás, kivonás) hibaszámítás szempontjából egymás között hasonlóságot mutatnak, ezért egyetlen tételben adjuk meg a megfelelő hibakorlátokat.

1.3.1 ABSZOLÚT HIBA

1.3.3. TÉTEL. *Az additív műveletek abszolút hibakorlátjai a következők:*

$$\begin{aligned} \delta(a+b) &\leq \delta a + \delta b, \\ \delta(a-b) &\leq \delta a + \delta b. \end{aligned}$$

B i z o n y í t á s.

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

amiből a fenti állításunk következik.

1.3.4. MEGJEGYZÉS. Mivel mindkét művelet esetén ugyanazt az eredményt kaptuk, valójában az előjelükre semmilyen kikötést nem kellett tenni. Az eredmény akárhány, tetszőleges előjelű tagra kiterjeszthető. Tekintsük a

$$\sum_{i=1}^{n} x_i \approx \sum_{i=1}^{n} a_i, \qquad (x_i = a_i \pm \delta a_i, \quad i = 1, 2, \ldots, n)$$

összegzést. Könnyen belátható, hogy $\delta\left(\sum_{i=1}^{n} a_i\right) = \sum_{i=1}^{n} \delta a_i$. Természetesen ez az esetek nagy részében jelentősen túlbecsli a tényleges abszolút hibát, hiszen azt tételezi fel, hogy az egyes tagok hibáinak előjele a legkedvezőtlenebbül alakul. Valószínűségszámítási eszközökkel élesebb becslés is adható, jó megbízhatósággal.

1.3.5. TÉTEL. *A multiplikatív műveletek abszolút hibakorlátjai a következők:*

$$\begin{aligned} \delta(ab) &\approx |a|\,\delta b + |b|\,\delta a, \\ \delta(a/b) &\approx \frac{|a|\,\delta b + |b|\,\delta a}{|b|^2}. \end{aligned}$$

B i z o n y í t á s. A szorzat abszolút hibakorlátjára kapjuk, hogy

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

Ha $|a| \gg |\Delta a|$ és $|b| \gg |\Delta b|$, akkor a $|\Delta a|\,|\Delta b|$ másodrendű hibatagot elhanyagolhatjuk és azzal éppen az állításunkat kapjuk.

Az osztás esetén természetesen feltesszük, hogy a nevező nem zérus és azt kapjuk, hogy

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

Itt pedig hasonló meggondolással a $\frac{\Delta b}{b}$ tagot hanyagolhatjuk el az 1 mellett, amivel állításunk kiadódik.

1.3.2 RELATÍV HIBA

1.3.6. TÉTEL. *Az aritmetikai műveletek relatív hibakorlátjai a következők (feltéve, hogy nevező sehol sem lehet zérus, és az additív műveleteknél az operandusok előjele megegyező):*

$$\frac{\delta(a+b)}{|a+b|} = \max\left\{\frac{\delta a}{|a|},\frac{\delta b}{|b|}\right\},$$

$$\frac{\delta(a-b)}{|a-b|} = \frac{\delta a+\delta b}{|a-b|},$$

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

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

Bizonyítás. Csak az összeadás relatív hibáját bizonyítjuk.

$$\begin{aligned}\frac{\delta(a+b)}{|a+b|} &= \frac{\delta a+\delta b}{|a+b|} = \frac{\left(\frac{|a|\delta a}{|a|}+\frac{|b|\delta b}{|b|}\right)}{|a+b|} \leq \\ &\leq \max\left\{\frac{\delta a}{|a|},\frac{\delta b}{|b|}\right\}\frac{|a|+|b|}{|a+b|} = \max\left\{\frac{\delta a}{|a|},\frac{\delta b}{|b|}\right\}.\end{aligned}$$

Az utolsó egyenlőség az a és b azonos előjeléből következik.

A kivonásra adott összefüggés megegyezik a definícióval. A szorzás és osztás relatív hibája behelyettesítés után azonnal adódik.

2 MÁTRIXOK

2.0.7. DEFINÍCIÓ. *Az $m \times n$ típusú (méretű) valós A mátrixon valós a_{ij} számok alábbi táblázatát értjük:*

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

Az m és az n értelemszerűen pozitív egész számok.

Beszélünk a mátrix i-edik soráról és j-edik oszlopáról.

A sorok és oszlopok metszéspontjában vannak a mátrix a_{ij} elemei ($i = 1, \dots, m,\ j = 1, \dots, n$).

Komplex számokból (sőt, más – absztrakt – struktúrák elemeiből) is felépíthetünk mátrixokat.

Az $m \times n$ típusú valós mátrixok halmazát $\mathbb{R}^{m \times n}$ jelöli, ennek megfelelően például az $A \in \mathbb{R}^{k \times l}$ azt jelenti, hogy A egy $k \times l$ típusú (méretű) valós mátrix.

2.0.8. DEFINÍCIÓ. *Az egyetlen sorból vagy egyetlen oszlopból álló mátrixot vektornak nevezzük.*

A sorvektor szokásos megadási módja: $x = [x_1, \dots, x_n]$. Az x_i a vektor egy eleme, másképpen komponense. Az oszlopvektorokat az

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

formában szoktuk megadni, ahol $\mathbb{R}^n$ az n komponensű oszlopvektorok halmaza tulajdonképpen $\mathbb{R}^n \equiv \mathbb{R}^{n \times 1}$).

2.0.9. MEGJEGYZÉS. Ha egyszerűen csak vektort említünk és nem tesszük hozzá, hogy az sor- vagy oszlopvektor-e, akkor azon mindig oszlopvektort értünk. A komponensek kettős indexelése egyrészt felesleges (az egyik mindig 1), másrészt ez is mutatja, hogy valójában a mátrixoktól függetlenül is értelmezhető matematikai objektum.

A definíciókban látható szögletes zárójelek helyett szokás kerek zárójeleket is használni, valamint gyakran célszerű a következő, tömörebb jelölést választani:

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

Ha a sorok és oszlopok száma megegyezik, akkor a mátrixot ***négyzetesnek*** nevezzük, tömörebb jelölése is némileg egyszerűsödik. Például az $n \times n$ méretű C mátrix esetén:

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

2.1 MÁTRIXMŰVELETEK

2.1.1 ÖSSZEADÁS

Csak azonos méretű mátrixok között értelmezzük, a következőképpen: ha $A, B \in \mathbb{R}^{m\times n}$, akkor

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

Az összeadás két fontos tulajdonsága: kommutatív és asszociatív (a tagok sorrendje felcserélhető és csoportosítható)

$$A + B = B + A \qquad\qquad (A + B) + C = A + (B + C).$$

2.1.2 SZÁMMAL VALÓ SZORZÁS

Legyen $A \in \mathbb{R}^{m\times n}$ és legyen λ valós szám (azaz $\lambda \in \mathbb{R}$). Ekkor

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

Nyilvánvaló, hogy $\lambda(\mu A) = (\lambda\mu)A$, továbbá a fenti két művelet értelmezéséből következik az alábbi két – a disztributivitást kimondó – szabály:

$$\lambda(A + B) = \lambda A + \lambda B \quad \text{és} \quad (\lambda + \mu)A = \lambda A + \mu A.$$

Megállapodás szerint számmal jobbról és balról is szorozhatunk: $\lambda A = A\lambda$.

Az $\mathbb{R}^{m\times n}$ halmazt (vagy az $\mathbb{R}^n$ halmazt) a fenti két művelettel algebrai struktúrának tekinthetjük és mint ilyen, rendelkezik mindazon tulajdonságokkal, amelyek a lineáris teret (más szóval – itt ugyan szokatlanul hangzik – vektorteret) definiálják; például van zéruselem, az összeadás invertálható, stb. Ezért nem lesz meglepetés, ha az általunk érzékelt háromdimenziós térben definiált vektorok bizonyos jellemzőit (pl. hosszúság) általánosítjuk többdimenzióban is, sőt ezt az általánosítást kiterjesztjük mátrixokra is.

2.1.3 TRANSZPONÁLÁS (TÜKRÖZÉS)

Az $A \in \mathbb{R}^{m\times n}$ transzponáltját jelölje A^T, amit a következőképpen definiálunk:

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

Úgy is fogalmazhatjuk: a sorokat és az oszlopokat felcseréljük.

A transzponálás definíciójából adódik, hogy

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

Négyzetes mátrixok esetén a transzponálás a főátlóra való tükrözést jelent. (A főátlót azon elemek alkotják, melyek sor– és oszlopindexe megegyezik, azaz az $a_{11}, a_{22}, \ldots, a_{nn}$ elemek.)

A transzponálás felhasználásával az oszlopvektorokat meg lehet adni még az $x = [x_1, \ldots, x_n]^T$, a sorvektorokat pedig az

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}^T$$

formában is.

2.1.4 SZORZÁS

Ha $A \in \mathbb{R}^{m\times k}$, és $B \in \mathbb{R}^{k\times n}$, akkor a szorzatukat a

$$\begin{aligned} C &= (c_{ij})_{i,j=1}^{m,n} = AB \in \mathbb{R}^{m\times n} \Leftrightarrow c_{ij} = \sum_{t=1}^{k} a_{it}b_{tj} \\ (i &= 1, \ldots, m,\ j = 1, \ldots, n). \end{aligned}$$

előírással definiáljuk.

Látható, hogy amint az összeadásnál, úgy itt is lényeges a műveletben szereplő operandusok (tényezők) mérete, csak itt más a követelmény: az első tényező oszlopainak száma és a második tényező sorainak a száma kell, hogy egyenlő legyen; az eredmény mérete a tényezőkéből következik. Fontos megjegyezni, hogy a szorzás nem kommutatív, tehát általában

$$AB \neq BA.$$

A két mátrix között többnyire nem is értelmezhető mindkét sorrendben a szorzás.

Ha $x, y \in \mathbb{R}^n$ akkor az $x^T y$ és az xy^T szorzás végezhető el; előbbi neve ***skalár szorzás***, utóbbié ***diadikus szorzás***. A skalár szorzat eredménye egy skalár szám (1×1-es mátrix), szokták külön is definiálni.

2.1.1. DEFINÍCIÓ. *Az $x, y \in \mathbb{R}^n$ skaláris szorzata a*

$$x^T y = \sum_{i=1}^{n} x_i y_i.$$

A skalár szorzás definíciójából látható, hogy a szorzatmátrix bármelyik (i, j) indexű elemét úgy kapjuk, mint egy skalárszorzatot: az első tényező i-edik sorát (mint sorvektort) szorozzuk a második tényező j-edik oszlopával, azaz

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

Gyakran kell alkalmaznunk a mátrixszorzásnak – az értelmezésből közvetlenül adódó – alábbi tulajdonságait:

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

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

2.2 SPECIÁLIS MÁTRIXOK, VEKTOROK

2.2.1. DEFINÍCIÓ. *Az A mátrix szimmetrikus, ha $A^T = A$.*

Nyilvánvaló, hogy csak négyzetes mátrix lehet szimmetrikus és ekkor $a_{ij} = a_{ji}$ $(i,\ j = 1, \ldots, n)$

2.2.2. DEFINÍCIÓ. *A $0 \in \mathbb{R}^{m\times n}$ mátrix zérusmátrix, ha az összes eleme zérus, azaz*

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

A zérusmátrix az összeadásra nézve a zéruselem, azaz minden A mátrix esetén

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

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

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

Az egységmátrix (amit gyakran a magyar elnevezésének kezdőbetűjével, E-vel is jelölünk) a szorzásra nézve egységelem, azaz minden $A \in \mathbb{R}^{n\times n}$ esetén

$$AI = IA = A.$$

2.2.4. MEGJEGYZÉS. Szokás az egységmátrix fogalmát a nem négyzetes mátrixokra is kiterjeszteni, a definícióját úgy adva meg, hogy az azonos indexpárú elemei 1-esek, a többi zérus (például

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

Ilyenkor persze az egység elnevezés már nem jogos, félrevezető. Mindenesetre nagyon hasznos kiterjesztés, több matematikai szoftver (a Matlab is) él vele.

Az egységmátrix mellett fontos fogalom az egységvektor.

2.2.5. DEFINÍCIÓ. *Az $e_i \in \mathbb{R}^n$ vektort (i-edik) egységvektornak nevezzük, ha az i-edik komponense 1-es, a többi pedig zérus.*

A transzponáltját felírva, tehát:

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

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

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

A diagonális elemek indexei azonosak, ezért jelöljük azokat rendszerint egyetlen indexszel. Magát a D diagonálmátrixot gyakran a $\mathrm{diag}(d_1, \ldots, d_n)$ vagy $\mathrm{diag}(d_i)$ $(i = 1, \ldots n)$ formában is jelöljük.

2.2.7. DEFINÍCIÓ. *A $P \in \mathbb{R}^{n\times n}$ mátrix permutációmátrix, ha minden sorában és oszlopában pontosan egy darab 1-es van és a többi elem zérus.*

Például az alábbi mátrix egy 4×4-es permutációmátrix:

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

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

Az alsó– és felsőháromszögmátrixok alakja sematikusan a következő:

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

Beszélni szoktunk valamely négyzetes $A \in \mathbb{R}^{n\times n}$ mátrix alsó és felső háromszögrészéről, amelyekre a $tril(A)$, illetve $triu(A)$ jelöléssel hivatkozunk. Ezeket értelemszerűen úgy kapjuk, hogy az eredeti mátrix főátlója feletti, illetve alatti elemeit kicseréljük zérusokra. Például az $A = \begin{bmatrix} 3 & -2 & 2 \\ 9 & 0 & 0 \\ -5 & 3 & 1 \end{bmatrix}$ mátrix esetén

$$tril(A) = \begin{bmatrix} 3 & 0 & 0 \\ 9 & 0 & 0 \\ -5 & 3 & 1 \end{bmatrix} \quad triu(A) = \begin{bmatrix} 3 & -2 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(az l az angol *lower*, u pedig az *upper* szó kezdőbetűje.)

2.2.9. DEFINÍCIÓ. *Az $A \in \mathbb{R}^{n\times n}$ mátrix sávmátrix p alsó sávszélességgel és q felső sávszélességgel, ha teljesül, hogy*

$$a_{ij} = 0, \quad \textit{ha } i > j + p, \textit{ vagy } i + q < j.$$

Más szóval: a $-p$-edik diagonálisuk alatti és q-adik diagonálisuk feletti elemeik zérusak. A sávot, amelynek elemeit nem kötelezően zérusokként kezeljük, azon a_{ij} elemek definiálják, amelyek indexeire teljesül, hogy $i - p \leq j \leq i + q$, vagy ekvivalens módon $j - q \leq i \leq j + p$. Sematikusan ábrázolva:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & a_{1,1+q} & 0 & \dots & \dots & 0 \\ a_{21} & a_{22} & & & & \ddots & & & \vdots \\ \vdots & & \ddots & & & & \ddots & & \vdots \\ a_{1+p,1} & & & \ddots & & & & \ddots & 0 \\ 0 & \ddots & & & \ddots & & & & a_{n-q,n} \\ \vdots & & \ddots & & & \ddots & & & \vdots \\ \vdots & & & \ddots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & & & \ddots & a_{n-1,n} \\ 0 & \dots & \dots & \dots & 0 & a_{n,n-p} & \dots & a_{n,n-1} & a_{nn} \end{bmatrix}.$$

Általában ritka mátrixoknak nevezzük azokat a mátrixokat, amelyek viszonylag sok, ismert pozíciójú zérust tartalmaznak. Ilyenek pédául a sávmátrixok, vagy lehetnek szabálytalan (de ismert, rögzített) elhelyezkedésű sok zérust tartalmazó mátrixok is. Az ilyen mátrixok tárolása a zérusok figyelmen kívül hagyásával helytakarékosan oldható meg, és a velük való különböző manipulációkat végrehajtó programok is gazdaságosan, művelettakarékosan írhatók meg. (Persze, tudni kell, hogy közben nem változnak-e meg a zérusok, illetve azt, hogy legfeljebb hol változhatnak.) Ebből a szempontból ritkának tekinthetők a háromszögmátrixok is, a nemzérus elemek tárolása történhet vektorban, például oszlopfolytonos sorrendben. A szimmetria is tekintetbe vehető; ilyen esetben elég csak az alsó (vagy a felső) háromszögrészt tárolni. (Itt is ügyelni kell a programozás során, hogy hol romlik el esetleg az algoritmus végrehajtása során a szimmetria.)

2.3 VEKTOR- ÉS MÁTRIXNORMÁK

2.3.1. DEFINÍCIÓ. *Az $f : \mathbb{R}^{n\times k} \to \mathbb{R}$ függvényt mátrixnormának (vagy vektornormának, ha $k = 1$) nevezzük, ha*

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

A norma szokásos jelölése: $\|x\|$.

Normát nyilván nagyon sokféleképpen adhatunk meg. Bár a vektorokat is (speciális méretű) mátrixokkal azonosítottuk, vannak szempontok, amelyek alapján mégis csak különböző objektumokról van szó, ezért konkrét normákat különböző

módon vezetünk be. Vektorok esetén a normák fontos osztályát alkotják, az ún. hatványnormák:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}},$$

ahol $p \geq 1$ egész szám.

A leggyakrabban használt vektornormák a következők:

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

A legutóbbit a $p \to \infty$ határátmenettel kapjuk. (Az $\|x\|_2$ norma megfogalmazásánál természetesen a jobboldalon az abszolút érték jel elhagyható valós vektorok esetén, de akkor nem, ha az elemek komplexek.)

2.3.2. PÉLDA. Legyen $v = [1, -7, 0, 3]^T$. Ekkor

$$\|v\|_1 = 1 + 7 + 0 + 3 = 11,$$

$$\|v\|_2 = \sqrt{1 + 49 + 0 + 9} = 7.681 \text{ (kerekítve)},$$

$$\|v\|_\infty = \max\{1, 7, 0, 3\} = 7.$$

A háromkomponensű vektorok által bezárt szög skalár szorzás segítségével való kiszámításának szabályát kiterjesztve, értelmezhetjük az akárhány komponensű vektorok közötti szöget is.

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

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

összefüggés definiálja.

Legyen $A = [a_{ij}]_{i,j=1}^{m,n}$. A leggyakrabban használt mátrixnormák a következők:

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

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

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

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

(A Frobenius-normánál itt sem kell csak komplex elemek esetén a jobboldalon az abszolút érték jel.)

2.3.4. PÉLDA. Legyen

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

Az említett normái:

$$\|A\|_1 = \max\{2+4;\ 1+1; 0+1\} = 6,$$

$$\|A\|_\infty = \max\{2+1+0;\ 4+1+1\} = 6,$$

$$\|A\|_F = \sqrt{4+1+0+16+1+1} = 4.796.$$

2.3.1 INDUKÁLT MÁTRIXNORMA

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

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

Az indukált mátrixnorma geometriai jelentése: az egységnormájú x vektorok megnyújtásának (Ax) maximális mértéke. Könnyen igazolható, hogy $\|A\|_1$ az $\|x\|_1$, $\|A\|_2$ az $\|x\|_2$, $\|A\|_\infty$ pedig az $\|x\|_\infty$ vektornorma által indukált mátrixnorma.

2.3.6. PÉLDA. Felhasználva az indukált mátrixnorma definícióját, igazoljuk, hogy $a, b \in R^n$ esetén $\left\|ab^T\right\|_2 = \|a\|_2 \|b\|_2$.
Megoldás: Az értelmezés szerint

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

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

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

Bizonyítás. Először igazoljuk, hogy indukált mátrixnormában

$$\|Ax\|_V \leq \|A\|_M \|x\|_V \quad \left(A \in \mathbb{R}^{n\times n},\ x \in \mathbb{R}^n\right).$$

(A továbbiakban az M és V normaindexeket elhagyjuk.)
Ha $x \neq 0$, az indukált mátrixnorma definíciója alapján

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

ahonnan

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

Azt az 1 normájú x-et választva melynél az $\|ABx\|$ maximális, éppen az állításunk adódik.

2.4 DETERMINÁNS, INVERZ

Jelölje $A(i)$ azt az $(n-1)\times(n-1)$-es mátrixot, amelyet az $A \in \mathbb{R}^{n\times n}$ az i-edik sora és *első* oszlopa elhagyásával kapunk:

$$A(i) = \begin{bmatrix} a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{i-1,2} & \cdots & a_{i-1,n} \\ a_{i+1,2} & \cdots & a_{i+1,n} \\ \vdots & \ddots & \vdots \\ a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

2.4.1. DEFINÍCIÓ. *A*

$$\begin{aligned} &\det(A) = a_{11}, && \text{ha } n = 1 \\ &\det(A) = a_{11}a_{22} - a_{12}a_{21}, && \text{ha } n = 2 \\ &\det(A) = \sum_{i=1}^{n} (-1)^{i+1} \cdot a_{i1} \cdot \det(A(i)), && \text{ha } n \geq 3 \end{aligned}$$

előírásokkal számított valós számot a négyzetes, $A \in \mathbb{R}^{n\times n}$ mátrix determinánsának nevezzük.

2.4.2. DEFINÍCIÓ. *Az $X \in \mathbb{R}^{n\times n}$ mátrixot a négyzetes, $A \in \mathbb{R}^{n\times n}$ mátrix inverzének nevezzük, ha*

$$AX = XA = I,$$

ahol I az egységmátrix.

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

2.4.3. TÉTEL. *Az inverz mátrixra fennállnak az alábbi tulajdonságok:*

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

Azokat a mátrixokat, melyeknek létezik inverze, nemszinguláris mátrixoknak nevezzük.

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

Ha $\det(A) = 0$, akkor a mátrixot szinguláris nak nevezzük.

2.5 KONDÍCIÓSZÁM

2.5.1. DEFINÍCIÓ. *A $cond(A) = \|A\| \, \|A^{-1}\|$ mennyiséget az $A \in \mathbb{R}^{n\times n}$ mátrix kondíciószámának nevezzük.*

Külön foglalkozunk az egy- és a többváltozós esetekkel.

Egyváltozós eset Legyen $f : \mathbb{R} \to \mathbb{R}$ legalább kétszer folyonosan differenciálható függvény, $x = a \pm \delta a$. Az $f(x)$ helyett $f(a)$-t számoljuk. Az

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

másodrendű Taylor-formulából kapjuk, hogy

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

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

$$\delta\left(f\left(a\right)\right) \approx |f'(a)|\,\delta a.$$

Többváltozós eset: Legyen $f : \mathbb{R}^n \to \mathbb{R}$ legalább kétszer folytonosan differenciálható függvény és $x, a \in \mathbb{R}^n, \Delta a = x - a$, valamint $x_i = a_i \pm \delta a_i$, ahol $x_i, a_i, \delta a_i \in \mathbb{R}$. A többváltozós Taylor-formulából az egyváltozós esethez hasonlóan a másodrendű tagot elhanyagolva (megjegyezzük, hogy nem mindig lehet) kapjuk:

$$f(x) \approx f(a) + \nabla f(a)^T (x - a),$$

ahol $\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]^T$. Ebből pedig adódik a

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

becslés.

2.5.2. DEFINÍCIÓ. *(Függvények relatív hibája)*

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

Egy függvény kiszámítása rendszerint egy algoritmussal történik, ezért érdekes megvizsgálni, hogy az input adat relatív hibáját az algoritmus hányszorosra nagyítja fel, amit a

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

mennyiség fejez ki.

Egyszerű átalakításokkal adódik, hogy

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

2.5.3. DEFINÍCIÓ. *(Függvények kondíciószáma) A*

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

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

2.5.4. DEFINÍCIÓ. *Egy függvényt numerikusan instabilnak, vagy rosszul kondicionáltnak nevezünk, ha nagy a kondíciószáma. A függvény stabil, vagy jól kondicionált, ha a kondíciószám kicsi.*

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

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

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

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

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

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

A kondíciószámot értelmezhetjük az $F = [f_1, \dots f_n]^T : \mathbb{R}^m \to \mathbb{R}^n$ többváltozós (ún. vektor-vektor) függvényre is. A levezetést mellőzve adódik, hogy

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

ahol $F'(x) = \left[\frac{\partial f_i}{\partial x_j}\right]_{i,j=1}^{n,m}$, az ún. Jacobi-mátrix.

2.5.7. MEGJEGYZÉS. A kondíciószám normafüggő.

A mátrixok kondíciószáma bevezetésének motivációja:

Legyen $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix, $x, y \in \mathbb{R}^n$ és tekintsük az $F : \mathbb{R}^n \to \mathbb{R}^n$ függvényt, ahol

$$F(x) = y = A^{-1}x$$

(azaz y az $Ay = x$ lineáris egyenletrendszer megoldása a jobboldali vektor függvényében). Egyszerű számolással megmutatható, hogy ezen függvény kondíciószáma $c(F,x) \leq \|A\| \, \|A^{-1}\|$, és ez a becslés pontos.

Így cond$(A) = \|A\| \, \|A^{-1}\|$ kifejezés az A mátrix kondíciószáma.

2.6 DIREKT ÉS INVERZ HIBÁK

A függvényértékek számítása során – mint már említettük – hiba következhet be.

Jelölje x és y a pontos értékeket és legyen pontosan $y = f(x)$, a ténylegesen számított behelyettesítési érték pedig $\hat{y}$. Az eltérést, azaz a $\Delta y = \hat{y} - y$ értéket ***direkt hibának*** nevezzük. Amennyiben az $\hat{y}$-ra valamely $\hat{x}$ értékkel pontosan fennáll, hogy $\hat{y} = f(\hat{x}) = f(x + \Delta x)$, akkor a Δx értéket ***inverz hibának*** mondjuk.

Az $x \to x+\Delta x = \hat{x}$ és az $y \to y+\Delta y = \hat{y}$ megváltozást (vagy megváltoztatást) perturbációnak is szoktuk említeni. Az inverz hiba elemzését és becslését ***inverz hibaanalízisnek*** nevezzük. Ha több inverz hiba is létezik, akkor a (valamilyen normában) legkisebb inverz hiba meghatározása az érdekes. (Gondoljunk például arra, hogy ha $x, \hat{y} \in \mathbb{R}^n$ és $A \in \mathbb{R}^{n\times n}$, akkor többféle $\Delta A \in \mathbb{R}^{n\times n}$ is szolgáltathatja ugyanazt az $\hat{y} = (A + \Delta A)x$ eredményt.)

A direkt és az inverz hiba kapcsolatának vizsgálatához tegyük fel, hogy f kétszer folytonosan differenciálható. Ekkor tehát felírható a következő Taylor-polinom:

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

ahol $\vartheta \in (0, 1)$ Így a számított megoldás abszolút hibája

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

A relatív hiba pedig

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

Innen kapjuk, hogy

$$\frac{\delta(\hat{y})}{|y|} \leq c(f, x)\,\frac{\delta(x)}{|x|}$$

közelítő egyenlőtlenséget, amely szóban kifejezve a következő:

relatív direkt hiba $\leq$ ***kondíciószámszám*** $\times$ ***relatív inverz hiba***

Az egyenlőtlenség azt mutatja, hogy egy rosszul kondicionált probléma számított megoldásának nagy lehet a (relatív) direkt hibája. Egy $y = f(x)$ értéket számító algoritmust ***direkt stabilnak*** nevezünk, ha a direkt hiba kicsi és ***inverz stabilnak*** nevezzük, ha bármely x értékre olyan $\hat{y}$ számított értéket ad, amelyre a Δx inverz hiba kicsi. A kicsi jelző környezetfüggő. Egy direkt stabil módszer nem feltétlenül inverz stabil. Ha az inverz hiba és a kondíciószám kicsi, akkor az algoritmus direkt stabil.

A gyakorlatban természetesen a számítás végeredményének a hibája, a direkt hiba a fontos. Az inverz hibaanalízis jelentősége abban áll, hogy sokszor az inverz hibát tudjuk becsülni. Az alkalmazott számítógép számábrázolási pontossága rendszerint ismert, gyakran annak (később tárgyalt) mérőszámát vagy az azzal arányos mennyiséget tekinthetjük inverz hibának. Az arányossági tényező megállapítása tapasztalatok alapján történik, szakkönyvek is ajánlanak értékeket. Jól kondicionált feladat esetén pedig az inverz hibából következtethetünk a direkt hibára.

3 LINEÁRIS EGYENLETRENDSZEREK MEGOLDÁSI MÓDSZEREI

3.1 LINEÁRIS EGYENLETRENDSZEREK

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

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

Az egyenletrendszert megadhatjuk a tömörebb

$$Ax = b$$

formában is, ahol

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

Ha $m < n$, akkor ***alulhatározott***,

ha $m > n$, akkor ***túlhatározott*** egyenletrendszerről beszélünk.

Az $m = n$ esetben pedig az egyenletrendszert ***négyzetesnek*** nevezzük. Az egyenletrendszerek geometriai tartalmát a következőképpen írhatjuk le:

Az $\mathbb{R}^n$ euklideszi tér $d \in \mathbb{R}^n$ normálvektorú és $x_0 \in R^n$ ponton átmenő hipersíkját az

$$(x - x_0)^T d = 0$$

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

A hipersík egyenlete más formában kifejezve:

$$x^T d = x_0^T d.$$

Az $Ax = b$ egyenletrendszert felírhatjuk az ekvivalens

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

alakban, ahol $a_i^T = [a_{i1}, \ldots, a_{in}]$,

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

- (i) az egyenletrendszernek nincs megoldása,
- (ii) az egyenletrendszernek pontosan egy megoldása van,
- (iii) az egyenletrendszernek végtelen sok megoldása van.

3.1.1. DEFINÍCIÓ. *Ha az $Ax = b$ lineáris egyenletrendszernek legalább egy megoldása van, akkor az egyenletrendszert **konzisztens**nek nevezzük. Ha az egyenletrendszernek nincs megoldása, akkor az egyenletrendszer **inkonzisztens**.*

Az $Ax = b$ egyenletrendszert felírhatjuk az ekvivalens

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

alakban is, ahol $a_i \in \mathbb{R}^n$ az A mátrix i-edik oszlopát jelöli (x_i pedig skalár: a megoldásvektor i-edik komponense). A $\sum_{i=1}^{n} x_i a_i$ összeg az $\{a_i\}_{i=1}^{n}$ vektorok lineáris kombinációja. Az egyenletrendszer akkor és csak akkor oldható meg, ha b kifejezhető az A oszlopvektorainak lineáris kombinációjaként.

A megoldhatóságot megállapíthatjuk a rang fogalmának segítségével is:

az $Ax = b$ egyenletrendszernek akkor és csak akkor van megoldása, ha $rank\,(A) = rank\,([A, b])$. Ha $rank\,(A) = rank\,([A, b]) = n$, akkor az $Ax = b$ egyenletrendszernek pontosan egy megoldása van.

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

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

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

3.2 A GAUSS-MÓDSZER

A Gauss-módszer két fázisból áll.

I. Azonos (a megoldást őrző) átalakításokkal az $Ax = b$ egyenletrendszert felső háromszög alak úra hozzuk:

II. A kapott felső háromszögmátrixú egyenletrendszert megoldjuk.

Az azonos átalakítást itt most úgy végezzük el, hogy egyik egyenletből kivonjuk a másik egyenlet alkalmasan megállapított konstansszorosát, így ott a szóban forgó ismeretlen együtthatója zérussá válik. Kiiktatjuk – idegen szóval eliminálju k – az ismeretlent, ezért is nevezzük a módszert eliminációs eljárásnak. Az első lépésben az

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

alakot kell tehát kapnunk, amit ha $a_{11} \neq 0$,akkor elérhetünk úgy, hogy az i-edik sorból kivonjuk ($i = 2, \ldots, n$) az első sor l_{i1}-szeresét:

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

Az $a_{i1} - l_{i1}a_{11} = 0$ feltételből kapjuk, hogy $l_{i1} = \frac{a_{i1}}{a_{11}}$. Ha ezt a l_{i1} értéket behelyettesítjük az egyenletbe, akkor könnyen ellenőrizhető, hogy itt tényleg arról van szó, hogy az első egyenletből kifejezzük az x_1-et, és a kapott kifejezést behelyettesítjük a maradék többibe. A számítások során nem kell az x_i szimbólumokat magunkkal cipelni, elég, ha az együtthatómátrix elemein hajtjuk végre a megfelelő módosítást. Így könnyen programozható a kinullázás alábbi algoritmusa:

Legyen $a_{ij}^{(1)} = a_{ij}$. Ekkor minden $i = 2, \ldots, n$ és $j = 1, \ldots, n$ esetén

$$l_{i1} = a_{i1}^{(1)}/a_{11}^{(1)}$$
$$a_{ij}^{(2)} = a_{ij}^{(1)} - l_{i1}a_{1j}^{(1)}$$
$$b_i^{(2)} = b_i^{(1)} - l_{i1}b_1^{(1)}$$

A következő lépésben minden $i = 3, \ldots, n$ és $j = 2, \ldots, n$ esetén

$$l_{i2} = a_{i2}^{(2)}/a_{22}^{(2)}$$
$$a_{ij}^{(3)} = a_{ij}^{(2)} - l_{i2}a_{2j}^{(2)}$$
$$b_i^{(3)} = b_i^{(2)} - l_{i2}b_2^{(2)}$$

Általános (k-adik) lépés: minden $i = k+1, \ldots, n$ és $j = k, \ldots, n$ esetén

$$l_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$$
$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)}$$
$$b_i^{(k+1)} = b_i^{(k)} - l_{ik}b_k^{(k)}$$

Az I. lépés: felső felső háromszög alakúra hozás Általános (k-adik) lépés: minden $i = k+1, \ldots, n$ és $j = k, \ldots, n$ esetén

$$l_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$$
$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)}$$
$$b_i^{(k+1)} = b_i^{(k)} - l_{ik}b_k^{(k)}$$

Tekintsük az $Ax = b$ egyenletrendszert, ahol $A = [a_{ij}]_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$ felső háromszögmátrix. Ekkor

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

Az egyenletrendszer akkor és csak akkor oldható meg egyértelműen, ha $a_{11} \neq 0, \dots, a_{nn} \neq 0$, azaz $\det(A) \neq 0$. Ezen egyenletrendszer megoldását adja a következő algoritmus:

$$\begin{array}{ll} 1 & x_n = b_n/a_{nn} \\ 2 & \textbf{FOR } i \leftarrow n-1 \textbf{ DOWNTO } 1 \\ 3 & \qquad x_i \leftarrow \left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j \right) / a_{ii} \end{array}$$

3.2.1 A GAUSS-MÓDSZER ALGORITMUSA

I. (eliminációs) fázis:

$$\begin{array}{ll} 1 & \textbf{FOR } k \leftarrow 1 \textbf{ TO } n-1 \\ 2 & \qquad \textbf{FOR } i \leftarrow k+1 \textbf{ TO } n \\ 3 & \qquad\qquad l_{ik} = a_{ik}/a_{kk} \\ 4 & \qquad\qquad b_i = b_i - l_{ik}b_k \\ 5 & \qquad\qquad \textbf{FOR } j \leftarrow k \textbf{ TO } n \\ 6 & \qquad\qquad\qquad a_{ij} \leftarrow a_{ij} - l_{ik}a_{kj} \end{array}$$

II. (visszahelyettesítő) fázis:

$$\begin{array}{ll} 1 & x_n = b_n/a_{nn} \\ 2 & \textbf{FOR } i \leftarrow n-1 \textbf{ DOWNTO } 1 \\ 3 & \qquad s \leftarrow 0 \\ 4 & \qquad \textbf{FOR } j \leftarrow i+1 \textbf{ TO } n \\ 5 & \qquad\qquad s \leftarrow s + a_{ij}x_j \\ 6 & \qquad x_i \leftarrow (b_i - s)/a_{ii} \end{array}$$

Alkalmazás a determináns kiszámítására. Ismeretes, hogy a determináns értéke nem változik, ha bármely sorához hozzáadjuk egy másik sor akárhányszorosát. Mivel a fenti eljárás közben csak ilyen átalakításokat hajtunk végre a mátrixon (és

így a determinánsán is), a determináns marad az eredeti. Azt tehát a végén a főátlóbeli elemek szorzata adja.

3.2.2 A GAUSS-MÓDSZER MŰVELETIGÉNYE

A Gauss-módszer véges sok lépésben, véges sok aritmetikai alapművelet ($+, -, *, /$) elvégzése után megadja az $Ax = b$ ($A \in \mathbb{R}^{n\times n}$) egyenletrendszer megoldását. A szükséges aritmetikai műveletszám (műveletigény) az egyenletrendszer-megoldó eljárások fontos minőségi jellemzője, mert az ilyen algoritmusok számítógépideje nagyjából arányos az aritmetikai műveletigénnyel (a kerekítési hibák halmozódása is – bár ez konkrét esetekben nem törvényszerű – kedvezőtlenebb lehet nagyobb műveletszám esetén).

A műveletigényt szokás az algoritmus költségének is nevezni, ennek megfelelően beszélhetünk *olcsó*, illetve *drága* eljárásokról.

Számláljuk össze a Gauss-módszer által igényelt additív és multiplikatív műveleteket. Jelöljük az additív műveleteket A-val, a multiplikatívakat pedig M-mel. (A hagyományos felépítésű számítógépeknél az összeadás és a kivonás nagyjából egyező időigényű, tőlük nagyobb, de egymás között szintén azonosnak vehető a szorzás és az osztás időigénye; indokolt tehát a műveletek ilyen szempontból két csoportba való sorolása.)

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

1	**FOR** $k \leftarrow 1$ **TO** $n-1$	
2	**FOR** $i \leftarrow k+1$ **TO** n	
3	$l_{ik} = a_{ik}/a_{kk}$	M
4	$b_i = b_i - l_{ik}b_k$	$(M+A)$
5	**FOR** $j \leftarrow k$ **TO** n	
6	$a_{ij} \leftarrow a_{ij} - l_{ik}a_{kj}$	$(M+A)$

A legbelső ciklusmag $(n-k+1)$-szor, a második $(n-k)$-szor fut le kapjuk:

$$((n-k)(n-k+1) + 2(n-k))\, M$$

$$((n-k)(n-k+1) + (n-k))\, A.$$

Minthogy ez a belső ciklusok a külső ciklus $k = 1, \ldots, n-1$ lépéseire fut le, ki kell számolni a

$$\sum_{k=1}^{n-1} ((n-k)(n-k+1) + 2(n-k))\, M$$

és

$$\sum_{k=1}^{n-1} ((n-k)(n-k+1) + (n-k))\, A$$

összegeket. Ismert azonosságok felhasználásával adódik az

$$\left(\frac{n^3}{3}+n^2-\frac{4}{3}n\right)M+\left(\frac{n^3}{3}+\frac{n^2}{2}-\frac{5}{6}n\right)A.$$

A II. fázis műveletigénye hasonló számítással:

1	$x_n = b_n/a_{nn}$	M
2	**FOR** $i \leftarrow n-1$ **DOWNTO** 1	
3	$\quad s \leftarrow 0$	
4	$\quad$ **FOR** $j \leftarrow i+1$ **TO** n	
5	$\quad\quad s \leftarrow s + a_{ij}x_j$	$(M+A)$
6	$\quad x_i \leftarrow (b_i - s)/a_{ii}$	$(M+A)$

Összegezve az $i = n-1, n-2, \ldots, 1$ lépések műveletigényét, kapjuk, hogy a II. fázis összköltsége

$$M+\sum_{i=1}^{n-1}\left((n-i)(M+A)+(M+A)\right)=\begin{cases}\left(\frac{n^2}{2}+\frac{n}{2}+1\right)M\\ \left(\frac{n^2}{2}+\frac{n}{2}\right)A.\end{cases}$$

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

$$\left(\frac{n^3}{3}+n^2-\frac{4}{3}n+\frac{n^2}{2}+\frac{n}{2}+1\right)M=\left(\frac{n^3}{3}+\frac{3}{2}n^2-\frac{11}{6}n+1\right)M$$

$$\left(\frac{n^3}{3}+\frac{n^2}{2}-\frac{5}{6}n+\frac{n^2}{2}+\frac{n}{2}\right)A=\left(\frac{n^3}{3}+n^2-\frac{1}{3}n\right)A.$$

Aritmetikai műveleteket használó akármely véges algoritmus lényeges jellemzője a végrehajtásához szükséges aritmetikai műveletek száma, természetesen a feladat paramétereinek (pl. ismeretlenek száma, együttható mátrix mérete, stb.) függvényében. A régebbi számítógépek esetén a multiplikatív műveletek végrehajtási ideje lényegesen nagyobb volt, mint az additívaké. Ezért ezeket külön határozták meg. Később megfigyelték, hogy a lineáris algebra számítási eljárásaiban az additív és a multiplikatív műveletek száma nagyon gyakran közel azonos. Ezért alkották meg a számítási igény mérésére a *flop* fogalmát.

3.2.1. DEFINÍCIÓ. *1* ***régi flop*** *az a számítási munka, amely az* $s = s + x * y$ *művelet* (*1 összeadás + 1 szorzás*) *elvégzéséhez kell. 1* ***új flop*** *pedig az a számítási munka, amely egyetlen* (*mindegy, hogy additív vagy multiplikatív*) *aritmetikai művelet elvégzéséhez szükséges.*

Az új flop bevezetését az indokolja, hogy az újabb számítógépeken a multiplikatív és az additív műveletek ideje azonosnak tekinthető. Tehát egy régi flop 2 új floppal azonos a mai számítógépeken.

Figyelmbe véve mindezeket, valamint, hogy nagy n-ekre a legmagasabb fokú tag válik dominánssá, kimondhatjuk a Gauss-módszer műveletigényéről szól ó tételt.

3.2.2. TÉTEL. *A Gauss-módszer $\frac{n^3}{3} + O(n^2)$ additív és ugyanennyi multiplikatív műveletet igényel, azaz összességében a műveletigénye $\frac{n^3}{3} + O(n^2)$ régi és $\frac{2n^3}{3} + O(n^2)$ új flop.*

3.2.3 A FŐELEMKIVÁLASZTÁSOS GAUSS-MÓDSZER

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

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

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

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

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

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

A sorok cseréjénél az egyenletek (és b megfelelő komponenseinek) sorrendje, az oszlopok cseréjénél pedig a változók sorrendje változik meg. Általában, így az előző példában is, több választási lehetőségünk is van sor-, vagy oszlopcserére. Ha azonban az $a_{kk} = 0$ elem alatt a többi együttható is zérus, akkor az $[a_{ij}]_{i,j=1}^{n,k}$ részmátrix oszlopai lineárisan összefüggők, A szinguláris és az eliminációs eljárás sorcserével sem folytatható. Hasonló a helyzet, ha a_{kk} és a sorában, tőle jobbra, minden együttható zérus, mert A ekkor is szinguláris.

A sorok felcserélését úgy, hogy az új pivot elem zérustól különböző legyen, ***pivotálásnak*** vagy **főelemkiválasztásnak** nevezzük. A pivot elem megválasztása nagymértékben befolyásolja az eredmények a numerikus stabilitást.

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

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

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

A részleges elnevezést az indokolja, hogy csak az aktuális oszlopbeli, szóba jöhető elemek közül választjuk ki a legnagyobb abszolút értékűt.

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

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

(megjegyezzük, hogy ***oszlopcsere*** esetén ***változócsere*** is történik.)

3.2.3. MEGJEGYZÉS. Ha determinánst számolunk pivotálással, akkor figyelemmel kell követnünk a sor- és az oszlopcserék számát. Egy sor- vagy oszlopcsere ugyanis a determináns előjelét megváltoztatja. Tehát a végső háromszögmátrix főátlóbeli elemeinek szorzatát még -1-gyel szorozni kell, amennyiben páratlan számú cserét hajtottunk végre.

3.2.4. MEGJEGYZÉS. A főelemkiválasztásos Gauss-módszer esetén az I. fázis minden lépésében pivotálást hajtunk végre. A teljes főelemkiválasztást biztonságos (numerikusan stabil) stratégiának tekinthetjük. A részleges főelemkiválasztás egyéb technikákkal kiegészítve ugyancsak biztonságosnak tekinthető. Éppen ezért a teljes főelem kiválasztás árát nem éri meg megfizetni. (A nagyság szerinti összehasonlítás is műveletet igényel; rendszerint különbségképzést és az eredmény előjel-bitjének figyelését. Ehhez jön még a többszöri abszolút érték-képzés és a nagyobb adatmozgatás.) A részleges főelemkiválasztás lényegesen kevesebb pótlólagos aritmetikai műveletet igényel mint a teljes főelemkiválasztás. A gyakorlatban tehát legtöbbször csak részleges főelem kiválasztást alkalmazunk, de azt – néhány kivételektől eltekintve – mindig!

Vannak a gyakorlatban többször előforduló, fontos esetek, amikor nem kell pivotálni, az eljárás főelemkiválasztás nélkül is stabilan viselkedik. Ezek a következők:

- Az A mátrix szimmetrikus és pozitív definit.

- Az A mátrix diagonálisan domináns a következő értelemben:

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

A pozitív definitség fogalma pedig a következő.

3.2.5. DEFINÍCIÓ. *Az $A \in \mathbb{R}^{n\times n}$ mátrix pozitív definit, ha $\forall x \in R^n$, $x \neq 0$ esetén $x^T Ax > 0$.*

3.3 AZ LU-FELBONTÁS

3.3.1. DEFINÍCIÓ. *Az $A \in \mathbb{R}^{n\times n}$ mátrix LU-felbontásán a mátrix*

$$A = L \cdot U$$

szorzatra bontását értjük, ahol $L \in \mathbb{R}^{n\times n}$ alsó, $U \in \mathbb{R}^{n\times n}$ pedig felső háromszögmátrix.

3.3.2. TÉTEL. *Az LU-felbontás nem egyértelmű.*

B i z o n y í t á s. Ha D egy nemszinguláris diagonális mátrix, akkor

$$L_1 \cdot U_1 = L_1 \cdot \left(DD^{-1}\right) \cdot U_1 = (L_1 D) \cdot \left(D^{-1} U_1\right) = L_2 \cdot U_2.$$

3.3.3. MEGJEGYZÉS.

- Ha A nemszinguláris, igazolható, hogy egyetlen LU-felbontásból az összes többi csak ilyen módon származtatható.

- Amennyiben A nemszinguláris és van egy $A = L \cdot U$ szorzatra bontása, akkor olyan $A = \tilde{L} \cdot \tilde{U}$ faktorizációt is találunk, melyben az $\tilde{L}$ vagy $\tilde{U}$ főátlóbeli elemeit tetszőlegesen, de 0-tól különböző számként előírjuk.

- Ilyenkor közbeszúrjuk azt a (DD^{-1})-et, amely $D = diag(d_{1,}, d_2, \ldots, d_n)$ elemeire $d_i = \tilde{u}_{ii}/u_{ii}$, ha ott az $\tilde{U}$ főátlóbeli eleme van $\tilde{u}_{ii}$-nek megadva, és $d_j = \tilde{l}_{jj}/l_{jj}$, ha a j-edik helyen az $\tilde{L}$ főátlóbeli elemét írjuk elő.

- Ebből az is következik, hogy ha van LU-felbontása A-nak, akkor egyértelműen van olyan is, ahol az L (vagy az U) minden főátlóbeli eleme 1-es. Az ilyen háromszögmátrixokat ***egység (alsó vagy felső) háromszögmátrixoknak*** nevezzük.

3.3.4. TÉTEL. *Egy $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrixnak akkor és csak akkor létezik LU-felbontása, ha összes főminor mátrixa is nemszinguláris, azaz*

$$\det\left(A_{(r)}\right) = det\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rr} \end{pmatrix} \neq 0 \qquad (r = 1, \ldots, n-1).$$

B i z o n y í t á s.

Emlékezzünk a Gauss-módszer k-adik lépésére Minden $i = k+1, \ldots, n$ és $j = k, \ldots, n$ esetén

$$l_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}$$
$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik} a_{kj}^{(k)}$$
$$b_i^{(k+1)} = b_i^{(k)} - l_{ik} b_k^{(k)}$$

Ezt úgy is írhatjuk, hogy a Gauss-elimináció során a k-adik oszlop kinullázása úgy történik, hogy elvégezzük a $A^{(k+1)} = L_k \cdot A^{(k)}$ szorzást, ahol

$$L_k = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \vdots & \cdots & -l_{kk} & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -l_{k+1,k} & 1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \cdots & -l_{nk} & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n\times n},$$

ahol $l_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}$

Ugyanis Az $L_k \cdot A^{(k)}$ szorzás az $A^{(k)}$ első k oszlopát változatlanul hagyja, a $k+1$-ediktől kezdődően az $i.$ sorhoz hozzáadja a $k.$ sor $-l_{ik}$-szorosát. A Gauss-elimináció alkalmazásával belátható, hogy az L_k mátrix inverze

$$L_k^{-1} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \vdots & \cdots & l_{kk} & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & l_{k+1,k} & 1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \cdots & l_{nk} & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n\times n},$$

A főelemkiválasztás nélküli Gauss-elimináció így felírható a következő alakban is.

$$L_{n-1} \cdot (L_{n-2} \cdot (\cdots L_1)) \cdot A = U,$$

és

$$L_1^{-1} \cdot (L_2^{-1} \cdot (\cdots L_{n-1}^{-1})) \cdot U = L \cdot U,$$

ahol L alsó háromszög mátrix, a főátlóban 1-esekkel, az U pedig felső háromszög mátrix.

Tehát az U mátrix nem más, mint a Gauss-módszer utolsó lépésében kapott $A^{(n)}$ mátrix, az L mátrix pedig az alábbi alakban irható:

$$L = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ l_{31} & l_{32} & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \cdots & \vdots & \vdots & \vdots \\ l_{k1} & l_{k2} & l_{k3} & \cdots & 1 & 0 & 0 & 0 \\ l_{k+1,1} & l_{k+1,2} & l_{k+1,3} & \cdots & l_{k+1,k} & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \cdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nk} & l_{n,k+1} & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n\times n},$$

Az L és U mátrix elemei megkaphatóak az alábbi képletekből:

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj} \qquad (i \le j)$$

$$l_{ij} = \frac{1}{u_{jj}} \left(a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj} \right) \qquad (i > j)$$

3.3.5. PÉLDA. Határozzuk meg az

$$A = \begin{bmatrix} 1 & -1 & 2 \\ 3 & -2 & 7 \\ 2 & -2 & 1 \end{bmatrix}$$

mátrix LU-felbontását!

Ha az r-edik lépésben elakad az elimináció, akkor $a_{rr}^{(r-1)} = 0$, ami azt jelenti, hogy $\det\left(A_{(r)}\right) = 0$.

He nem tesszük fel, hogy $\det\left(A_{(r)}\right) \neq 0$ minden $r = 1, ..., n$ estén, akkor van olyan nemszinguláris mátrix, amelynek nincs LU-felbontása. Például az

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

mátrixnak nincs LU-felbontása.

Viszont részleges főelemkiválasztást használva minden nemszinguláris A mátrix esetén sikeresen elvégezhető a Gauss-elimináció.

Mivel a sorcserék alkalmazása a sorok permutálásának felel meg, ezért kimondhatjuk a következő tételt.

3.3.6. TÉTEL. *Minden $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrixhoz létezik olyan P permutációmátrix, hogy a PA mátrixnak van LU-felbontása.*

3.4 A CHOLESKY-FELBONTÁS

Egy alsó háromszögmátrix transzponáltja felső háromszögmátrix és ez fordítva is igaz. Felmerül a kérdés, hogy milyen mátrixoknak van olyan LU-felbontása, melyben a két tényező egymás transzponáltja.

3.4.1. TÉTEL. *Ha az $A \in \mathbb{R}^{n\times n}$ mátrix szimmetrikus és pozitív definit, akkor létezik $A = L \cdot L^T$ szorzatra bontása, ahol L alsó háromszögmátrix.*
Ezt a felbontást nevezzük ***Cholesky-felbontásnak****.*

Néha a Cholesky-felbontást $A = U^T \cdot U$ alakban írják és vannak szoftverek (például a Matlab is ilyen), amelyek az U mátrixot állítják elő.

A Cholesky-felbontást úgy is megkaphatjuk ezen mátrixoknál, hogy elkészítjük Gauss-eliminációval az LU-felbontást, majd megkeressük azt a D diagonális mátrixot, amelynél az $(LD)\,(D^{-1}U)$ szorzatban az LD és az $D^{-1}U$ diagonális elemei egyenlők.

Ez azonban a műveletszámot növeli jelentősen, hisz a Cholesky-felbontásban elég az egyik tényezőt kiszámítani, ami durván fele annyi költséggel jár.

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

egyenletből

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

és

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

azaz

$$l_{kk} = \sqrt{a_{kk} - \left(l_{k1}^2 + l_{k2}^2 + \ldots + l_{k,k-1}^2\right)}$$

és

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

A Cholesky-felbontás algoritmusa:

1 **FOR** $k \leftarrow 1$ **TO** n **DO**

2 $\quad l_{kk} = \left(a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2\right)^{1/2}$

3 $\quad$ **FOR** $i \leftarrow k+1$:**TO** n **DO**

4 $\quad\quad l_{ik} = \left(a_{ik} - \sum_{j=1}^{k-1} l_{ij}l_{kj}\right)/l_{kk}$

A Cholesky-felbontás költsége a fenti implementálásban

$$\frac{1}{6}n^3 + O(n^2).$$

3.5 ALKALMAZÁSOK

Tekintsük az $Ax = b$ megoldását, ahol A nemszinguláris. Van olyan P permutációmátrix, hogy létezik a $PA = LU$ faktorizáció és innen azt kapjuk, hogy

$$PAx = LUx = Pb.$$

Bevezetve az $y = Ux$ új változót, végrehajtható tehát a következő:

Az egyenletrendszer megoldása LU-módszerrel

1. Határozzuk meg a $PA = LU$ felbontást.
2. Oldjuk meg az $Ly = Pb$ egyenletrendszert y-ra.
3. Oldjuk meg az $Ux = y$ egyenletrendszert x-re.

A 2. és a 3. lépésben háromszögmátrixú egyenletrendszert kell megoldani, tehát az összköltséget dominánsan az 1. lépés határozza meg. Szimmetrikus, pozitív definit A esetén az 1. lépésben természetesen a Cholesky-felbontást alkalmazzuk.

3.5.1. MEGJEGYZÉS. A P mátrixot nem kell előre ismernünk. Az 1. lépésben

a faktorizációt a főelemkiválasztást alkalmazó Gauss-eliminációval végezzük el, és ott regisztráljuk a sorcseréket. Ekkor egy $PAP_1 = LU$ faktorizációt végzünk

és az x megoldásnál vesszük figyelembe a P_1-et. Így a fentebb leírt LU-módszer lényegében ugyanaz, mint a főelemkiválasztásos Gauss-módszer. A különbség ott jelentkezik, hogy az A-val nem párhuzamosan transzformáljuk a jobboldali b vektort, hanem azt elhalasztjuk. A 2. lépés pontosan a b transzformációját hajtja

végre.

Az LU-módszert akkor különösen előnyös használni, ha egynél több, előre nem ismert jobboldalú

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

alakú egyenletrendszert kell megoldani. Ekkor elég az A mátrix LU-felbontását egyszer meghatározni, majd rendre az $Ly_i = b_i$, $Ux_i = y_i$ ($i = 1, \ldots, k$), összesen $2k$ darab háromszögmátrixú egyenletrendszert megoldani.

Alkalmazhatjuk például mátrix invertálására, az előre rögzített $b_i = e_i$ egységvektorokkal ($i = 1, 2, \ldots, n$). Ekkor az eljárás egyébként teljesen ekvivalens a Gauss-Jordan elimináció alkalmazásával.

3.6 GAUSS-JORDAN MÓDSZER

Ugyanazzal a technikával, mint ahogy a k-adik oszlopban az a_{kk} alatti elemeket kinulláztuk, a fölötte lévő elemeket is zérussá lehet tenni. Azaz az eliminációs fázisban k minden értékére az i ciklusváltozót nemcsak $k+1$-től n-ig, hanem 1-től n-ig futtathatjuk, kivéve az $i = k$ esetet. (Ez annak felel meg, mintha az x_k-nak az k-adik egyenletből való kifejezése után azt az összes többibe behelyettesítenénk.)

Az **I. fázis** végeredménye így egy diagonálmátrixú egyenletrendszer, vagyis a **II. fázis** ekkor csupán az $x_i = b_i/a_{ii}$ ($i = 1, 2, \ldots, n$) utasításokból áll (amiket menet közben, egy-egy oszlop teljes kinullázása után – vagy még előtte – azonnal is megtehetünk).

Persze, szekvenciálisan végrehajtva ez a módszer nem előnyös, hiszen jelentősen megnő a műveletek száma.

Ha viszont csak azután kezdünk a főátló fölötti elemek nullázásával foglalkozni, miután kialakítottuk a felső háromszögmátrixot, és ezt a nullázást a $k = n, n-1, \ldots, 2$ sorrendben végezzük (tehát az oszlopok szerint visszafelé haladva), akkor az A mátrix elemeihez már nem kell hozzányúlni.

Ugyanis az i-edik sor $-l_{ik}$-szor a k-adik sor $(i = 1, 2, \ldots, k-1)$ elvégzése során a k-adik sorban az a_{kk} elem kivételével minden elem (elvileg) már 0. A k-adik oszlopba sem kell a 0-át beírni. A **II. fázis** úgy tekinti, hogy ott zérus áll. A főátló fölötti elemek nullázása tehát nem más, mint a már tárgyalt Gauss-módszer **II. fázisa**.

Az algoritmus több, de ugyanolyan együttható mátrixú $Ax = b_j$ ($b_j \in \mathbb{R}^n$, $j = 1, 2, \ldots, m$) egyenletrendszert oldjon meg.

Főátló alatti nullázás (I. fázis):

Legyen $B = [b_1, b_2, \ldots, b_m]$
Legyen $A = [A, B]$, azaz kibővítjük az A-t a jobboldali b vektorokkal
FOR $k \leftarrow 1$ **TO** n-1 **DO**
// Határozzuk meg a t indexet, hogy $|a_{tk}| = \max_{k \le i \le n} |a_{ik}|$.
    IF $k \neq t$
        cseréljük fel a k-adik és t-edik sort
    FOR $i \leftarrow k+1$ **TO** n **DO**
        $l_{ik} = a_{ik}/a_{kk}$
        FOR $j \leftarrow k+1$ **TO** $n+m$ **DO**
            $a_{ij} = a_{ij} - l_{ik}a_{kj}$

Főátló fölötti nullázás (II. fázis):

FOR $k \leftarrow n$ **DOWNTO** 2 **DO**
    FOR $i \leftarrow 1$ TO $k-1$ **DO**
        $l_{ik} = A_{ik}/A_{kk}$
        FOR $j \leftarrow n+1$ TO $n+m$ **DO**
            $a_{ij} = a_{ij} - l_{ik}a_{kj}$
    FOR $j \leftarrow n+1$ TO $n+m$ **DO**
        $x_{k,j-n} = a_{kj}/a_{kk}$
FOR $j \leftarrow n+1$ TO $n+m$ **DO**
    $x_{1,j-n} = a_{1j}/a_{11}$

Végeredmény:

$$[x_1, x_2, \ldots, x_m] = X$$

3.6.1. MEGJEGYZÉS. A Gauss-Jordan eljárás I. fázisában a részleges főelemkiválasztás elhagyható. Fenti algoritmus alkalmas mátrixinvertálásra. Könnyen belátható

ugyanis, hogy az $Ax = e_i$ egyenletrendszer megoldása éppen az inverz mátrix i-edik oszlopvektora. Ha az algoritmusban B az egységmátrix, akkor a végeredmény: $X = A^{-1}$.

3.7 ITERATÍV-ELJÁRÁSOK

Tekintsük az

$$Ax = b \tag{1}$$

lineáris egyenletrendszert, ahol $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix, $b \in \mathbb{R}^n$.

Feladat: Keressük az egyenletrendszer közelítő megoldását.

A **direkt módszer-típus** a feladat paramétereinek felhasználásával előállítja, megkonstruálja a módszer által szolgáltatott közelítő megoldást.

Az **iterációs módszer-típus** a feladat paramétereinek felhasználásával előállít, képez valamilyenfajta iterációs képletet és bizonyos ún. kezdeti értékből kiindulva, az iterációs képlet alkalmazásával közelítő megoldások sorozatát állítja elő. Ekkor az alábbi kérdéseket kell megválaszolni:

- az előállított közelítő megoldások sorozata konvergens-e;
- ha a konvergencia teljesül, hová tart ez a sorozat;
- ha a sorozat az egzakt megoldáshoz tart, milyen a konvergencia sebessége.

Célunk az, hogy az $x^{(0)} \in \mathbb{R}^n$ kezdeti vektorból kiindulva olyan $x^{(r)} \in \mathbb{R}^n$, $r = 1, 2, \ldots$ vektor-sorozatot generáljunk, melyre

$$\lim_{r\to\infty} x^{(r)} = x^*$$

teljesül, ahol x^* az egyenletrendszer egzakt megoldása.

3.7.1 STACIONÁRIS ITERATÍV-ELJÁRÁSOK

Tekintsük az

$$Ax = b$$

lineáris egyenletrendszert, ahol $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix, $b \in \mathbb{R}^n, b \neq 0$.

A fenti lineáris egyenletrendszer. iterációs megoldásához tekintsük a

$$\begin{aligned}
&\Phi_0(A, b)\\
&\Phi_1(x^{(0)}, A, b)\\
&\Phi_2(x^{(1)}, x^{(0)}, A, b)\\
&\vdots\\
&\Phi_r(x^{(r-1)}, \ldots, x^{(1)}, x^{(0)}, A, b)
\end{aligned}$$

függvényeket, melyek révén az $x^{(r)}$, $r = 1, 2, \ldots$ sorozatot az alábbiak szerint definiáljuk:

$$\begin{aligned} x^{(0)} &= \Phi_0(A, b) \\ x^{(1)} &= \Phi_1(x^{(0)}, A, b) \\ x^{(2)} &= \Phi_2(x^{(1)}, x^{(0)}, A, b) \\ &\vdots \\ x^{(r)} &= \Phi_r(x^{(r-1)}, \ldots, x^{(1)}, x^{(0)}, A, b) \end{aligned}$$

Gyakran $x^{(0)}, x^{(1)}, \ldots, x^{(s)} \in \mathbb{R}^n$ vektorok az önkényesen választott $\Phi_0, \Phi_1 \ldots, \Phi_{s-1}$ függvényekre, valamely $s \geq 0$ esetén $\Phi = \Phi_s = \Phi_{s+1} = \ldots, \Phi_k$ teljesül. Ha minden $r \geq s$ esetén a Φ_r függvény r-től független, ahol s valamely pozitív egész szám, akkor a módszert **stacionárius**-nak nevezzük.

Stacionárius esetben legyen $\Phi = \Phi_s = \Phi_{s+1} = \ldots$. Ekkor $x^{(r+1)}$ legfeljebb az előző $x^{(r)}, x^{(r-1)}, \ldots, x^{(r-s+1)}$ darab vektortól függ.

$s = 1$ esetén:

$$\begin{aligned} x^{(0)} &= \Phi_0(A, b) \\ x^{(r)} &= \Phi(x^{(r-1)}, A, b) \qquad (r = 1, 2, \ldots) \end{aligned}$$

$s = 2$ esetén:

$$\begin{aligned} x^{(0)} &= \Phi_0(A, b) \\ x^{(1)} &= \Phi_0(x^{(0)}, A, b) \\ x^{(r)} &= \Phi(x^{(r-1)}, x^{(r-2)}, A, b) \qquad (r = 2, 3, \ldots) \end{aligned}$$

A fentiekben definiált módszerek foka legfeljebb 1, ill. 2.

Ha Φ_r az $(x^{(r-1)}, \ldots, x^{(1)}, x^{(0)}$ vektorok lineáris függvénye, akkor a módszert lineárisnak nevezzük. Egyébként a módszer nemlineáris.

Elsőfokú, lineáris stacionárius módszer esetén $x^{(r)} = \Phi(x^{(r-1)}, A, b)$ a következő alakot ölti:

$$x^{(r+1)} = Gx^{(r)} + k \tag{2}$$

ahol $G \in \mathbb{R}n \times n$ és $k \in \mathbb{R}$ az eredeti lineáris egyenletrendszer paramétereiből valamilyen módon képezett mátrixot, ill. vektort jelöl. Ezt az alakot **iteratív alak**nak is nevezik.

Megvizsgáljuk, hogy a fenti iteratív alak alkalmazásával nyert vektor-sorozatot milyen feltételek mellett konvergens, mikor konvergál a megoldáshoz, s milyen a konvergencia sebessége.

Vezessük be az ún. kapcsolt lineáris egyenletrendszert:

$$(E - G)x = k \tag{3}$$

ahol $E \in \mathbb{R}^{n\times n}$ az egységmátrix.

Megvizsgáljuk, hogy mi a kapcsolat az eredeti lineáris egyenletrendszer és a kapcsolt lineáris egyenletrendszer. között, s ezáltal elemezzük az iteratív alak által definiált módszert.

Jelölje $\zeta(A,b)$ az $Ax = b$ megoldás-halmazát és jelölje $\zeta(E-G,k)$ a kapcsolt $(E-G)x = k$ megoldás-halmazát. Legyen $x^* \in \mathbb{R}^n$ az $Ax = b$ egyenlet egzakt megoldása (azaz $x^* = A^{-1}b$).

3.7.1. DEFINÍCIÓ. *Azt mondjuk, hogy a $x^{(r+1)} = Gx^{(r)} + k$ képlettel definiált iterációs módszer az $Ax = b$ lineáris egyenletrendszerrel*

- *konzisztens, ha $\zeta(A,b) \subseteq \zeta(E-G,k)$,*
- *reciprok konzisztens, ha $\zeta(E-G,k) \subseteq \zeta(A,b)$,*
- *teljesen konzisztens, ha $\zeta(A,b) = \zeta(E-G,k)$.*

3.7.2. TÉTEL. *Legyen $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix, $b \in \mathbb{R}^n$, $b \neq 0$ és x^* az () egzakt megoldása. Ha a () által meghatározott $x^{(0)}, x^{(1)}, x^{(2)} \ldots \in \mathbb{R}^n$ sorozat minden $x^{(0)}$ esetén az x^* megoldáshoz konvergál, akkor () teljesen konzisztens. Másrészt, ha () teljesen konzisztens és az általa meghatározott $x^{(0)}, x^{(1)}, x^{(2)} \ldots$ sorozat konvergens, akkor ez az x^* megoldáshoz tart.*

3.7.3. TÉTEL. *Legyen $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix A () iterációs módszer akkor és csak akkor teljesen konzisztens az () egyenletrendszerrel, ha konzisztens és az $E - G \in \mathbb{R}^{n\times n}$ mátrix nemszinguláris. Ha $E - G \in \mathbb{R}^{n\times n}$ nemszinguláris, a teljes konzisztencia akkor és csak akkor áll fenn, ha a módszer reciprok konzisztens és $A \in \mathbb{R}^{n\times n}$ nemszinguláris.*

3.7.4. TÉTEL. *Legyen $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix, $b \in \mathbb{R}^n$. A () iterációs módszer akkor és csak akkor konzisztens az () lineáris egyenletrendszerrel, ha létezik olyan $M \in \mathbb{R}^{n\times n}$ mátrix, melyre*

$$G = E - MA \qquad k = Mb$$

teljesül, ahol $E \in \mathbb{R}^{n\times n}$ az egységmátrix.

3.7.5. TÉTEL. *Legyen $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix és $b \in \mathbb{R}^n$. Ekkor (**??**) iterációs módszer akkor és csak akkor konzisztens (**??**) lineáris egyenletrendszerrel, ha*

$$k = (E - G)A^{-1}b$$

teljesül.

3.8 JACOBI-MÓDSZER

Tekintsük

$$Ax = b$$

lineáris egyenletrendszert, ahol $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix és $b \in \mathbb{R}^n$, $b \neq 0$. Legyen $x^{(0)} \in \mathbb{R}^n$ valamely kezdeti érték.

Tekintsük az

$$A = L + D + U$$

felbontást, ahol

$L \in \mathbb{R}^{n\times n}$ alsó háromszög mátrix, melyre $l_{ij} = a_{ij}$ $(1 \leq i \leq n, 1 \leq j < i)$,
$D \in \mathbb{R}^{n\times n}$ diagonális mátrix, melyre $d_{ii} = a_{ii}$ $(1 \leq i \leq n)$,
$U \in \mathbb{R}^{n\times n}$ felső háromszög mátrix, melyre $u_{ij} = a_{ij}$ $(1 \leq i \leq n, i < j \leq n)$.

Ekkor

$$\begin{aligned}(L + D + U)x &= b\\ Dx &= b - (L + U)x\\ x &= D^{-1}(b - (L + U)x) = -D^{-1}(L + U)x + D^{-1}b\end{aligned}$$

adódik, amely a

$$x = Gx + k$$

előállítást eredményezi, ahol

$$G = -D^{-1}(L + U) \in \mathbb{R}^{n\times n}, \qquad k = D^{-1}b \in \mathbb{R}^n$$

Megmutatjuk, hogy teljesül a teljes konzisztencia, azaz, van olyan $M \in \mathbb{R}^{n\times n}$ mátrix, amelyre

$$G = E - MA \qquad k = Mb$$

teljesül.

Meghatározzuk az $M \in \mathbb{R}^{n\times n}$ mátrixot.

$$E - G = E + D^{-1}(L + U) = E + D^{-1}(A - D) = E + D^{-1}A - D^{-1}D = D^{-1}A$$

és

$$k = D^{-1}b$$

Fentiekből $M = D^{-1}$ következik. Így az $E - G$ nemszingularitásából következik, hogy a módszer teljesen konzisztens.

3.8.1. PÉLDA. Oldjuk meg iterációval (ha konvergens) az alábbi egyenletrendszert 0.05 pontossággal:

$$\begin{array}{rcrcrcr} 8x_1 & + & 2x_2 & - & 4x_3 & = & -2 \\ 2x_1 & - & 5x_2 & + & x_3 & = & 9 \\ 2x_1 & + & x_2 & + & 7x_3 & = & 15 \end{array}$$

Megoldás: A pontos megoldás: $[1, -1, 2]^T$, majd ezzel is összehasonlítjuk a közelítéseket Látható, hogy az együttható mátrix diagonálisan domináns, így az eljárás konvergens lesz.

Átírjuk az egyenletrendszert az alábbi alakba:

$$\begin{array}{rcrcrcrcr} x_1 & = & & & - & 0.25x_2 & + & 0.5x_3 & - & 0.25 \\ x_2 & = & 0.4x_1 & & & & + & 0.2x_3 & - & 1.8 \\ x_3 & = & -0.2857x_1 & + & -0.1429x_2 & & & & = & 2.1429 \end{array}$$

$\|G\|_\infty = 0.75 < 1$, tehát valóban konvergens az iteráció. Induljunk ki az $x^{(0)} = c = [-0.25, -1.8, 2.1429]^T$ vektorból. A következő közelítések adódnak:

$$\begin{array}{c} x^{(1)} = [1.2714, -1.4714, 2.4714]^T \\ x^{(2)} = [1.3536, -0.7971, 1.9898]^T \\ \vdots \\ x^{(8)} = [0.9923, -1.0024, 1.9987]^T \end{array}$$

Meglehetősen lassú a konvergencia, aminek a $\|G\|_\infty$ viszonylag nagy értéke az oka.

3.9 GAUSS-SEIDEL-MÓDSZER

Tekintsük

$$Ax = b$$

lineáris egyenletrendszert, ahol $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix és $b \in \mathbb{R}^n$, $b \neq 0$. Legyen $x^{(0)} \in \mathbb{R}^n$ valamely kezdeti érték.

Tekintsük az

$$A = L + D + U$$

felbontást, ahol

$L \in \mathbb{R}^{n\times n}$ alsó háromszög mátrix, melyre $l_{ij} = a_{ij}$ $(1 \leq i \leq n, 1 \leq j < i)$,

$D \in \mathbb{R}^{n\times n}$ diagonális mátrix, melyre $d_{ii} = a_{ii}$ $(1 \leq i \leq n)$,

$U \in \mathbb{R}^{n\times n}$ felső háromszög mátrix, melyre $u_{ij} = a_{ij}$ $(1 \leq i \leq n, i < j \leq n)$.

Ekkor

$$\begin{aligned}(L+D+U)x &= b\\ (L+D)x &= b-Ux\\ x &= (L+D)^{-1}(b-Ux) = -(L+D)^{-1}Ux + (L+D)^{-1}b\end{aligned}$$

adódik, amely a

$$x = Gx + k$$

előállítást eredményezi, ahol

$$G = -(L+D)^{-1}U \in \mathbb{R}^{n\times n}, \qquad k = (L+D)^{-1}b \in \mathbb{R}^n$$

Megmutatjuk, hogy teljesül a teljes konzisztencia, azaz, van olyan $M \in \mathbb{R}^{n\times n}$ mátrix, amelyre

$$G = E - MA \qquad k = Mb$$

teljesül.

Meghatározzuk az $M \in \mathbb{R}^{n\times n}$ mátrixot.

$$\begin{aligned}E - G &= E + (L+D)^{-1}U = E + (L+D)^{-1}(A - (L+D))\\ &= E + (L+D)^{-1}A - (L+D)^{-1}(L+D)\\ &= E + (L+D)^{-1}A - E = (L+D)^{-1}A\end{aligned}$$

és

$$k = (L+D)^{-1}b.$$

Fentiekből $M = (L+D)^{-1}$ következik. Így az $E-G$ nemszingularitásából következik, hogy a módszer teljesen konzisztens.

3.9.1. PÉLDA. Oldjuk meg most Seidel-iterációval az előbbi egyenletrendszert, 0.05 pontossággal:

$$\begin{array}{rcrcrcr} 8x_1 & + & 2x_2 & - & 4x_3 & = & -2\\ 2x_1 & - & 5x_2 & + & x_3 & = & 9\\ 2x_1 & + & x_2 & + & 7x_3 & = & 15 \end{array}$$

Megoldás: Ugyanazon átalakítás után, ugyancsak az $x^{(0)} = c = [-0.25, -1.8, 2.1429]^T$ választással a következőt kapjuk:

$$\begin{aligned}x^{(1)} &= [1.2714, -0.8629, 1.9029]^T\\ x^{(2)} &= [0.9171, -1.0526, 2.0312]^T\\ &\vdots\\ x^{(5)} &= [0.9999, -1.0001, 2.0001]^T\end{aligned}$$

3.10 HIBABECSLÉSEK

Tekintsük

$$Ax = b$$

lineáris egyenletrendszert, ahol $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix és $b \in \mathbb{R}^n$, $b \neq 0$. Legyen $x^{(0)} \in \mathbb{R}^n$ egy kezdeti érték és $x^* \in \mathbb{R}^n$ az egyenlet egzakt megoldása.

Tekintsük az

$$x^{(r+1)} = Gx^{(r)} + k \qquad r = 0, 1, 2 \dots$$

konzisztens iterációs módszert. Nyilvánvalóan

$$x^* = Gx^* + k$$

teljesül.

Ekkor az r-dik közelítés hibáját

$$e^{(r)} = x^{(r)} - x^*$$

képlettel definiáljuk.

Az

$$x^{(r+1)} - x^* = G(x^{(r)} - x^*)$$

összefüggésből

$$e^{(r+1)} = Ge^{(r)}$$

adódik. Tegyük fel, hogy a G iterációs mátrixra $\|G\| < 1$ teljesül. Ekkor

$$\left\|e^{(r+1)}\right\| = \left\|Ge^{(r)}\right\| \leq \|G\| \left\|e^{(r)}\right\| < \left\|e^{(r)}\right\|$$

adódik.

3.10.1. TÉTEL. *Legyen adott az $Ax = b$ lineáris egyenletrendszert, ahol $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix és $b \in \mathbb{R}^n$, $b \neq 0$ és a belőle származtatott*

$$x^{(r+1)} = Gx^{(r)} + k \qquad r = 0, 1, 2 \dots$$

iterációs módszer. Ha a $G \in \mathbb{R}^{n\times n}$ iterációs mátrixra $\|G\|_\infty < 1$, akkor tetszőleges $x^{(0)} \in \mathbb{R}^n$ kezdeti érték esetén az iterációs módszer konvergens vektorsorozatot eredményez, mely tart az $x^* \in \mathbb{R}^n$ egzakt megoldáshoz és a közelítés hibájára teljesül a következő egyenlőtlenség minden $r = 1, 2, \dots$ esetén:

$$\left\|x^{(r+1)} - x^*\right\|_\infty \leq \frac{\|G\|_\infty}{1 - \|G\|_\infty} \left\|x^{(r+1)} - x^{(r)}\right\|_\infty .$$

B i z o n y í t á s. Első lépésként számítsuk ki a következőt:

$$\begin{aligned}\left\|x^{(r+1)}-x^{(r)}\right\|_\infty &= \left\|Gx^{(r)}-Gx^{(r-1)}\right\|_\infty \leq \|G\|_\infty \left\|x^{(r)}-x^{(r-1)}\right\|_\infty \\ &= \|G\|_\infty \left\|Gx^{(r-1)}-Gx^{(r-2)}\right\|_\infty \\ &\leq \|G\|_\infty^2 \left\|x^{(r-1)}-x^{(r-2)}\right\|_\infty = \dots \\ &\leq \|G\|_\infty^{r-t} \left\|x^{(t+1)}-x^{(t)}\right\|_\infty\end{aligned}$$

minden $1 \leqq t < r$ esetén

Továbbá érvényes a következő:

$$\begin{aligned}\left\|x^{(r+p)}-x^{(r)}\right\|_\infty \leq{}& \left\|x^{(r+p)}-x^{(r+p-1)}\right\|_\infty + \left\|x^{(r+p-1)}-x^{(r+p-2)}\right\|_\infty \\ &+ \dots + \left\|x^{(r+2)}-x^{(r+1)}\right\|_\infty + \left\|x^{(r+1)}-x^{(r)}\right\|_\infty .\end{aligned}$$

Ebből az előzőt felhasználva azt kapjuk, hogy

$$\begin{aligned}\left\|x^{(r+p)}-x^{(r)}\right\|_\infty \leq{}& \|G\|_\infty^{p-1} \left\|x^{(r+1)}-x^{(r)}\right\|_\infty \\ &+ \|G\|_\infty^{p-2} \left\|x^{(r+1)}-x^{(r)}\right\|_\infty + \dots \\ &+ \|G\|_\infty \left\|x^{(r+1)}-x^{(r+1)}\right\|_\infty + \left\|x^{(r+1)}-x^{(r)}\right\|_\infty\end{aligned}$$

A mértani sor összegképletét felhasználva kapjuk, hogy

$$\|G\|_\infty^{p-1} + \|G\|_\infty^{p-2} + \dots + \|G\|_\infty + 1 = \frac{1-\|G\|_\infty^p}{1-\|G\|_\infty}$$

Ebből és az előző egyentlőtlenségből adódik, hogy

$$\begin{aligned}\left\|x^{(r+p)}-x^{(r)}\right\|_\infty &\leq \frac{1-\|G\|_\infty^p}{1-\|G\|_\infty} \left\|x^{(r+1)}-x^{(r)}\right\|_\infty \\ &\leq \frac{1}{1-\|G\|_\infty} \left\|x^{(r+1)}-x^{(r)}\right\|_\infty \\ &\leq \frac{\|G\|_\infty^r}{1-\|G\|_\infty} \left\|x^{(1)}-x^{(0)}\right\|_\infty\end{aligned}$$

.

Ha r-rel tartunk a végtelenhez, akkor az egyenlőtlenség jobboldala 0-hoz tart. Ebből adódik, hogy az $\{x^{(r)}\}_{r=1}^{\infty}$ sorozat Cauchy-sorozat, ami véges dimenzióban azt jelenti, hogy konvergens (mert $\mathbb{R}^n$ teljes metrikus tér a szokásos normával). Ebből következik, hogy csak egyetlen határértéke lehet: x^*.

Vegyük az egyenlőtlenség mindkét oldalán a $p \to \infty$ határátmenetet.

Ekkor $\lim_{p\to\infty} x^{(r+p)} = x^*$ és $\|G\|_\infty < 1$ miatt $\lim_{p\to\infty} \|G\|_\infty^p = 0$.

Emiatt és $\|G\|_\infty < 1$ kapjuk, hogy

$$\left\|x^* - x^{(r)}\right\|_\infty \leq \frac{1}{1-\|G\|_\infty} \left\|x^{(r+1)} - x^{(r)}\right\|_\infty,$$

azaz

$$\left\|e^{(r)}\right\|_\infty \leq \frac{1}{1-\|G\|_\infty} \left\|x^{(r+1)} - x^{(r)}\right\|_\infty.$$

Mivel $\left\|e^{(r+1)}\right\|_\infty \leq \|G\|_\infty \left\|e^{(r)}\right\|_\infty$, ezért

$$\left\|x^* - x^{(r+1)}\right\|_\infty \leq \|G\|_\infty \frac{1}{1-\|G\|_\infty} \left\|x^{(r+1)} - x^{(r)}\right\|_\infty,$$

azaz

$$\left\|x^* - x^{(r+1)}\right\|_\infty \leq \frac{\|G\|_\infty}{1-\|G\|_\infty} \left\|x^{(r+1)} - x^{(r)}\right\|_\infty,$$

ami az állításunk volt.

3.10.2. KÖVETKEZMÉNY. *A Jacobi-módzser konveregnciájának elégséges feltétele, hogy*

$$\left\|-D^{-1}(L+U)\right\|_\infty < 1.$$

3.10.3. KÖVETKEZMÉNY. *A Gauss-Seidel-módzser konveregnciájának elégséges feltétele, hogy*

$$\left\|-(L+D)^{-1}U\right\|_\infty < 1.$$

Könnyű belátni, hogy ha az együttható mátrix diagonálisan domináns, akkor mind a Jacobi, mind a Gauss-Seidel módszer esetén a $\|G\|_\infty < 1$ feltétel automatikusan teljesül.

A Gauss-Seidel módszer esetén

$$\begin{aligned}(L+D)x^{(r+1)} &= -Ux^{(r)} + b\\ Dx^{(r+1)} &= -Lx^{(r+1)} - Ux^{(r)} + b\\ x^{(r+1)} &= -D^{-1}Lx^{(r+1)} - D^{-1}Ux^{(r)} + D^{-1}b\end{aligned}$$

áll elő, mely a módszer közismert iterációs alakja.

Ha a konvergencia elégséges feltétele automatikusan teljesül, akkor ezen alak alkalmazása a hatékonyabb, mivel ekkor nem szükséges az inverz meghatározása. Ekkor, stop-kritériumként használható pl. olyan feltétel, hogy a két legutoljára képezett közelítés különbségének normája elegendően kicsinek bizonyul-e.

3.11 ALGORITMUSOK

A Jacobi-módszer algoritmusa: Adottak:

- Az $Ax = b$ lineáris egyenletrendszer, ahol $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix, $b \in \mathbb{R}^n, b \neq 0$ vektor
- valamely $x^{(0)}$ adott kezdeti vektor;
- ε a közelítés pontosságát definiáló paraméter $(\varepsilon > 0)$.

Határozzuk meg a $G = -D^{-1}(L+U)$ mátrixot és $\|G\|_\infty$ értékét.
IF $\|G\|_\infty < 1$
  THEN $h = 2\varepsilon$
    $r = 0$
    WHILE $h > \varepsilon$ **DO**
      $x^{(r+1)} = Gx^{(r)} + k$
      $h = \dfrac{\|G\|_\infty}{1-\|G\|_\infty} \left\|x^{(r+1)} - x^{(r)}\right\|_\infty$
      $r = r + 1$
    RETURN$(x^{(r)})$
  ELSE RETURN(„hiba”)

A Gauss-Seidel-módszer algoritmusa:
Adottak:

- Az $Ax = b$ lineáris egyenletrendszer, ahol $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix, $b \in \mathbb{R}^n, b \neq 0$ vektor
- valamely $x^{(0)}$ adott kezdeti vektor;
- ε a közelítés pontosságát definiáló paraméter $(\varepsilon > 0)$.

Határozzuk meg a $G = -(L+D)^{-1}U$ mátrixot és $\|G\|_\infty$ értékét.
IF $\|G\|_\infty < 1$
  THEN $h = 2\varepsilon$
    $r = 0$
    WHILE $h > \varepsilon$ **DO**
      $x^{(r+1)} = Gx^{(r)} + k$
      $h = \dfrac{\|G\|_\infty}{1-\|G\|_\infty} \left\|x^{(r+1)} - x^{(r)}\right\|_\infty$
      $r = r + 1$
    RETURN$(x^{(r)})$
  ELSE RETURN(„hiba”)

3.12 A KONVERGENCIA GYORSÍTÁSA

A konvergencia gyorsítása a következőképpen érhető el:

Tekintsük az elsőfokú, lineáris stacionárius iterációs módszer általános iterációs képletét:

$$x^{(r+1)} = Gx^{(r)} + k,$$

Legyen

$$x^{(r+1)} = \omega(Gx^{(r)} + k) + (1-\omega)x^{(r)},$$

ahol ω az ún. relaxációs tényező, amelyre $0 < \omega \leq 2$ teljesül.

Könnyű belátni, hogy $\omega = 1$ esetén ez a két képlet megegyezik. Ha $0 < \omega < 1$ esetén alulkorrigálásnak, $1 < \omega \leq 2$ esetén túlkorrigálásnak nevezzük a

Ekkor

$$x^{(r+1)} = (\omega G + (1-\omega)E)x^{(r)} + \omega k$$

érvényes, ami

$$x^{(r+1)} = G_\omega x^{(r)} + k_\omega,$$

iterációs képletet eredményezi.

Itt az iterációs mátrix

$$G_\omega = (\omega G + (1-\omega)E)$$

és

$$k_\omega = \omega k.$$

Ezek konkrét alakja a következők:

Jacobi-módszer: $G_\omega = -\omega D^{-1}(L+U) + (1-\omega)E.$

Gauss-Seidel-módszer: $G_\omega = -\omega(L+D)^{-1}U + (1-\omega)E.$

Az ω-t úgy kívánjuk választani, hogy ezáltal a konvergencia gyorsabbá váljon.

4 LEGKISEBB NÉGYZETEK MÓDSZERE

4.1 LEGKISEBB NÉGYZETEK MÓDSZERE, EGYENES ESET

Legyen $N \in \mathbb{N}$ és adottak az $x_1, x_2, \ldots, x_N \in \mathbb{R}$ alappontok és az $y_1, y_2, \ldots, y_N \in \mathbb{R}$ függvényértékek (pl. mérési eredmények). Keressük azt az egyenest $y = a_0 + a_1 x$, melyre a

$$\sum_{i=1}^{N} [y_i - (a_0 + a_1 x_i)]^2$$

kifejezés minimális.

A fenti feltételnek eleget tevő egyenest az (x_i, y_i) $i = i, \dots, N$, értékeket négyzetesen legjobban közelítő egyenesnek nevezzük.

A feladat megoldásához az

$$F(a_0, a_1) = \sum_{i=1}^{N} \big[y_i - (a_0 + a_1 x_i)\big]^2 : \mathbb{R}^2 \to \mathbb{R}$$

függvényt kell minimalizálnunk. A többváltozós függvények szélsőértékéről tanultak szerint az $F'_{a_0}(a_0, a_1) = 0$ és $F'_{a_1}(a_0, a_1) = 0$ feltételnek eleget tevő a_0, a_1-et keressük. A parciális deriváltakra

$$\sum_{i=1}^{N} -2[y_i - (a_0 + a_1 x_i)] = 0$$

$$\sum_{i=1}^{N} -2[y_i - (a_0 + a_1 x_i)]x_i = 0$$

egyenletrendszert kapjuk.

Ezt az egyenletrendszert az alábbi alakban írhatjuk:

$$\sum_{i=1}^{N} y_i - N a_0 - \sum_{i=1}^{N} a_1 x_i = 0$$

$$\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} a_0 x_i - \sum_{i=1}^{N} a_1 x_i^2 = 0$$

amelyből adódik, hogy

$$N a_0 + \left(\sum_{i=1}^{N} x_i\right) a_1 = \sum_{i=1}^{N} y_i$$

$$\left(\sum_{i=1}^{N} x_i\right) a_0 + \left(\sum_{i=1}^{N} x_i^2\right) a_1 = \sum_{i=1}^{N} x_i y_i$$

Vezessük be a következő jelöléseket:

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \in \mathbb{R}^{N \times 2}, \quad b = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N, \quad a = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \in \mathbb{R}^2.$$

Ekkor

$$A^TA = \begin{pmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{pmatrix} \qquad A^Tb = \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{pmatrix}$$

Így az egyenletrendszer

$$A^TAa = A^Tb$$

alakban írható.

A $det(A^TA) = 0$ csak akkor teljesülhet, ha $x_1 = x_2 = \ldots = x_N$ (érdektelen eset).

Tehát feltehetjük, hogy $det(A^TA) \neq 0$. Ekkor az egyenletrendszer egyértelműen megoldható. Például az A^TA invertálható, így

$$a = (A^TA)^{-1}A^Tb.$$

4.2 A LEGKISEBB NÉGYZETEK MÓDSZERE, POLINOM ESET

Legyen $n, N \in \mathbb{N}$ úgy, hogy n << N, adottak az $x_1, x_2, \ldots, x_N \in \mathbb{R}$ alappontok és az $y_1, y_2, \ldots, y_N \in \mathbb{R}$ függvényértékek (pl. mérési eredmények). Keressük azt a $P_n(x) = \sum_{j=0}^{n} a_j x^j$ polinomot, melyre a

$$\sum_{i=1}^{N} (y_i - P_n(x_i))^2$$

kifejezés minimális.

A fenti feltételnek eleget tevő P_n polinomot az (x_i, y_i) $i = i, \ldots, N$, értékeket négyzetesen legjobban közelítő n-ed fokú polinomnak nevezzük.

A feladat megoldásához az

$$F(a_0, a_1, \ldots, a_n) = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{n} a_j x_i^j \right)^2 : \mathbb{R}^{n+1} \to \mathbb{R}$$

függvényt kell minimalizálnunk. A többváltozós függvények szélsőértékéről tanultak szerint az $F'(a_0, a_1, \ldots, a_n) = 0$ feltételnek eleget tevő a_j-ket keressük. A

parciális deriváltakra

$$\frac{\partial F}{\partial a_j}(a_0, a_1, \ldots, a_n) = \sum_{i=1}^{N} 2(y_i - P_n(x_i)) \left(-\frac{\partial P_n}{\partial a_j}(x_i) \right) = 0$$

$(j = 0, 1, \ldots, n)$.

$$\sum_{i=1}^{N} P_n(x_i) \frac{\partial P_n}{\partial a_j}(x_i) = \sum_{i=1}^{N} y_i \frac{\partial P_n}{\partial a_j}(x_i) \qquad (j = 0, 1, \ldots, n).$$

Mivel $\dfrac{\partial P_n}{\partial a_j}(x_i) = (x_i)^j$, a fenti egyenlet a következő alakba írható:

$$\sum_{i=1}^{N} (x_i)^j \sum_{k=0}^{n} a_k (x_i)^k = \sum_{k=0}^{n} a_k \sum_{i=1}^{N} (x_i)^{j+k} = \sum_{i=1}^{N} y_i (x_i)^j$$

$(j = 0, 1, \ldots, n)$. Ezzel a_k -kra egy lineáris egyenletrendszert kaptunk ($n + 1$ darab egyenlet, $n + 1$ darab ismeretlennel).

Vezessük be a következő jelöléseket:

$$A = \begin{pmatrix} 1 & x_1 & \ldots & x_1^n \\ 1 & x_2 & \ldots & x_2^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & \ldots & x_N^n \end{pmatrix} \in \mathbb{R}^{N \times (n+1)},$$

$$b = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N, \qquad a = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} \in \mathbb{R}^{n+1}.$$

Ekkor az egyenletrendszer

$$A^T A a = A^T b$$

alakban írható.

4.3 A LEGKISEBB NÉGYZETEK MÓDSZERE, FÜGGVÉNY ESET

Az f függvény helyettesítésére (közelítésére) a szóba jöhető, előre rögzített H függvényosztályból azt a $h \in H$ függvényt keressük, amely az

$$\|f - h\| \to \min, \qquad h \in H$$

feltételes szélsőérték feladat megoldása. Tulajdonképpen minden $h \in H$ tekinthető közelítésnek, ezért a feladatot kielégítő függvényt szokás legjobb approximációnak nevezni.

Függvények $[a, b]$ intervallumon való legkisebb négyzetes közelítéséről akkor beszélünk, ha a norma diszkrét esetben ($a \leq x_1 < x_2 < \ldots < x_m \leq b$)

$$\|f\|_2 = \left(\sum_{i=1}^{m} f^2 (x_i)\, w (x_i) \right)^{\frac{1}{2}},$$

folytonos esetben pedig

$$\|f\|_2 = \left(\int_a^b f^2 (x)\, w (x)\, dx \right)^{\frac{1}{2}},$$

ahol a rögzített $w(x)$ súlyfüggvényre diszkrétnél a $w(x_i) > 0$ ($i = 1, 2, \ldots m$), folytonosnál pedig a $w(x) \in C[a, b]$, $w(x) > 0, \forall x \in [a, b]$ teljesülését megköveteljük. Fontos speciális eset a $w(x) \equiv 1$.

Lineáris eset: Legyen a H függvényhalmaz olyan, hogy ismert

$$\phi_i : [a, b] \to \mathbb{R} (i = 1, \ldots, n)$$

függvények valamennyi lineáris kombinációját tartalmazza, tehát a $h(x)$ függvény alakja

$$h(x) = a_1\phi_1(x) + a_2\phi_2(x) + \ldots + a_n\phi_n(x) = \sum_{i=1}^{n} a_i\phi_i(x).$$

A ϕ_i függvényeket ***alapfüggvényeknek*** vagy másképpen ***bázisfüggvényeknek*** nevezzük.

Diszkrét, lieáris eset: Fontos kérdés az approximációs feladat megoldásának létezése és egyértelműsége. Lineáris approximációra igaz az alábbi állítás.

4.3.1. TÉTEL. *Ha $\{\phi_i\}_{i=1}^n \subset C[a, b]$ lineárisan függetlenek, akkor bármilyen normában és minden $f \in C[a, b]$ esetén létezik legjobban közelítő $h(x) = \sum_{i=1}^n a_i\phi_i(x)$ függvény.*

Legyen $F = F(a_0, a_1 \ldots, a_n)$. Ekkor meg kell oldani a

$$F = \sum_{i=i}^{m} \left[f(x_i) - \left(a_1\phi_1(x_i) + \ldots + a_j\phi_j(x_i) + \ldots + a_n\phi_n(x_i) \right) \right]^2 \to \min$$

szélsőértékfeladatot. Ennek megoldása pedig $\dfrac{\partial F}{\partial a_j} = 0$, $(j = 1, 2, \ldots, n)$, vagyis a

$$-2\sum_{i=i}^{m} \left[f(x_i) - (a_1\phi_1(x_i) + \ldots + a_j\phi_j(x_i) + \ldots + a_n\phi_n(x_i)\right] \phi_j(x_i) = 0$$

lineáris egyenletrendszer megoldása. (Az egyenlet teljesülése az approximációs feladat megoldásának már említett egyértelmű létezése miatt elegendő.)

Egyszerűsítés és a szokásos alakra való rendezés után kapjuk, hogy

$$a_1 \sum_{i=i}^{m} \phi_1(x_i)\phi_j(x_i) + \ldots + a_n \sum_{i=i}^{m} \phi_n(x_i)\phi_j(x_i) = \sum_{i=i}^{m} f(x_i)\phi_j(x_i)$$

$(j = 1, 2, \ldots, n)$. Vezessük be az

$$\langle u, v\rangle = \sum_{i=i}^{m} u(x_i)v(x_i)w(x_i)$$

jelölést.

Ezzel az egyenletrendszer alakja a következő:

$$\begin{aligned}
a_1 \langle\phi_1, \phi_1\rangle + a_1 \langle\phi_2, \phi_1\rangle + \ldots + a_n \langle\phi_n, \phi_1\rangle &= \langle f, \phi_1\rangle \\
a_1 \langle\phi_1, \phi_2\rangle + a_1 \langle\phi_2, \phi_2\rangle + \ldots + a_n \langle\phi_n, \phi_2\rangle &= \langle f, \phi_2\rangle \\
&\vdots \\
a_1 \langle\phi_1, \phi_n\rangle + a_1 \langle\phi_2, \phi_n\rangle + \ldots + a_n \langle\phi_n, \phi_n\rangle &= \langle f, \phi_n\rangle
\end{aligned}$$

4.3.2. MEGJEGYZÉS. A

$$\langle u, v\rangle = \sum_{i=i}^{m} u(x_i)v(x_i)w(x_i)$$

összefüggéssel egy skaláris szorzatot definiáltunk a diszkrét pontokon értelmezett függvények között. Ez két $\mathbb{R}^n$-beli vektornak a szorzata (ha $w(x) \equiv 1$). Az egyenletrendszer az úgynevezett normálegyenletrendszer. A $G = \left[\langle\phi_j, \phi_i\rangle\right]_{i,j=1}^{n}$, $a = [a_1, \ldots, a_n]^T$ és a $b = \left[\langle f, \phi_1\rangle, \ldots, \langle f, \phi_n\rangle\right]^T$ jelölésekkel tömörebben:

$$Ga = b.$$

A $G \in \mathbb{R}^{n\times n}$ mátrixot Gram-mátrixnak nevezzük.

Legyen $A = [\phi_j(x_i)]_{i,j=1}^{m,n} \in \mathbb{R}^{m\times n}$, $a = [a_1, \ldots, a_n]^T \in \mathbb{R}^n$, $b = y = [y_1, \ldots, y_m]^T \in \mathbb{R}^m$ és $m > n$.

Keresünk olyan a^* paramétervektort, amely az $Aa-b$ hibát valamilyen normában minimalizálja. Ha létezik a $Aa = b$ egyenletnek megoldása, akkor a minimumfeladat egyenértékű vele. Az euklideszi normában megfogalmazott

$$\|Aa - b\|_2 \to \min .$$

minimumfeladat megoldása az alábbi tétel:

4.3.3. TÉTEL. *Az $a \in R^n$ akkor és csak akkor megoldása a feladatnak, ha*

$$A^T Aa = A^T b.$$

4.4 LEGKISEBB NÉGYZETEK MÓDSZERE, FOLYTONOS ESET

Legyen $f \in C[a, b]$és $h(x) = a_1\phi_1(x) + a_2\phi_2(x) + \ldots + a_n\phi_n(x)$. Ekkor tehát az

$$\begin{aligned} F(a_1, \ldots, a_n) &= \left\| f - \sum_{i=1}^{n} a_i\phi_i \right\|_2^2 = \\ &= \int_a^b \left(f(x) - \sum_{j=1}^{n} a_j\phi_j(x) \right)^2 dx \to \min \end{aligned}$$

($w(x) \equiv 1$) szélsőérték-feladatot kell megoldani. Említettük, hogy a feladatnak lineárisan független alapfüggvények esetén egyértelmű megoldása van.

$$\frac{\partial F(a_1, \ldots, a_n)}{\partial a_i} = 0 \quad (i = 1, \ldots, n)$$

egyenletrendszer megoldására redukálódik a feladat. A parciális deriválás során használjuk ki, hogy itt a deriválás és integrálás sorrendje felcserélhető és vezessük be itt is a skalárszorzatot az

$$\langle u, v\rangle = \int_a^b u(x)v(x)dx, \qquad u, v \in C[a, b]$$

értelmezéssel. Ekkor formailag ugyanahhoz az egyenletrendszerhez jutunk, mint a diszkrét esetben. (Természetesen, választhatunk valamilyen $w(x) > 0$ ($x \in [a, b]$) súlyfüggvényt itt is; most $w(x) \equiv 1$).

A Gram-mátrix ugyanúgy rosszul kondicionált lehet, mint a diszkrét esetben.

4.4.1. PÉLDA. Legyen $a = 0$, $b = 1$, $w(x) \equiv 1$és $\phi_i(x) = x^{i-1}$ $(i = 1, \ldots, n)$. Határozzuk meg a Gram-mátrixot. **Megoldás:** $\phi_i(x)\,\phi_j(x) = x^{i+j-2}$, $g_{ij} =$

$\int_0^1 x^{i+j-2}dx = 1/(i+j-1)$ és

$$G = \left[\frac{1}{i+j-1}\right]_{i,j=1}^{n},$$

ami nem más mint az ún. Hilbert-mátrix.

Ortogonális, ortonormált eset: Ha a $\langle \phi_i, \phi_j \rangle$ skalárszorzat zérust ad minden $i \neq j$ esetén, akkor itt is ortogonális függvényrendszerről beszélünk az adott $[a, b]$ intervallumon. Ha még $\langle \phi_i, \phi_i \rangle = 1$ is teljesül a szóbanforgó i indexekre, akkor ortonormált rendszer a neve. Ilyen bázisra való áttérésre folytonos esetben is van lehetőség. Például a $C[-1, 1]$, $w(x) \equiv 1$ esetén a

$$\sqrt{\frac{1}{2}}P_0, \sqrt{\frac{2n+1}{2}}P_n(x) \quad (n = 1, 2, \ldots)$$

ortonormált függvényrendszer, ahol

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left(x^2 - 1\right)^n$$

az ún. n-edik Legendre-polinom.

Az első néhány Legendre-polinom:

$$P_0(x) = 1,\ P_1(x) = x,\ P_2(x) = \frac{3}{2}x^2 - \frac{1}{2},\ P_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x.$$

Ha rendelkezésünkre áll egy ortonormált $\{\phi_i(x)\}_{i=1}^n$ függvényrendszer, akkor a legkisebb négyzetes approximáció meghatározásához (elvileg) csak az $a_i = \langle f, \phi_i \rangle$ $(i = 1, \ldots, n)$ ún. Fourier-együtthatókat kell kiszámolnunk.

4.4.2. PÉLDA. Az $f(x) = \sqrt{x}$ függvénynek adjuk meg az egyenessel való legkisebb négyzetes közelítését az $[1, 4]$ intervallumon. **Megoldás:** $\phi_1(x) \equiv$

$1, \phi_2(x) = x$ és $f(x) \approx h(x) = a_1\phi_1(x) + a_2\phi_2(x) = a_1 + a_2(x)$. A $[\langle \phi_j, \phi_i \rangle]_{i,j=1}^2$ együtthatómátrixot és az $[\langle f, \phi_1 \rangle, \langle f, \phi_2 \rangle]^T$ jobboldalt itt is kiszámolva a

$$\begin{array}{rcrcr} 3a_1 & + & 7.5a_2 & = & 4.6667 \\ 7.5a_2 & + & 30a_2 & = & 21 \end{array}$$

egyenletrendszerhez jutunk. Ezt megoldva a $\sqrt{x} \approx 0.7407 + 0.3259x$ egyenes egyenletét kapjuk.

4.4.3. PÉLDA. Itt is áttérhetnénk ortonormált rendszerre. Nem nehéz belátni, hogy a $\phi_1^*(x) \equiv \frac{1}{\sqrt{3}}$ és a $\phi_2^*(x) = \frac{2}{3}(x - 2.5)$ függvények ortonormált rendszert alkotnak az $[a, b] = [1, 4]$ intervallumon. Az

$$f(x) \approx h^*(x) = a_1^*\phi_1^*(x) + a_2^*\phi_2^*(x)$$

közelítést használva a normál egyenletrendszer együtthatómátrixa az egységmátrix és az eredmény természetesen ugyanaz az egyenes, mint amit az előbb már megkaptunk.

4.4.4. MEGJEGYZÉS. Az $f \in C[a, b]$ függvények körében az L_2-norma mellett többféle norma is fontos szerepet játszik a gyakorlatban. Ezek közül az egyik a Csebisev-norma. A Csebisev-norma értelmezése

$$\|f\|_C = \max_{x\in[a,b]} |f(x)|,$$

a Csebisev-féle approximációs feladat pedig adott f függvény és H függvényhalmaz esetén: $h \in H$ és

$$\|f - h\|_C = \max_{x\in[a,b]} |f(x) - h(x)| \to \min$$

A feladat megoldását az f legjobb egyenletes approximációjának nevezzük (természetesen az adott intervallumon és az adott H mellett). Az elnevezést az indokolja, hogy a $\max_{x\in[a,b]} |f(x) - h(x)|$ egyben az elkövetett hibának egy, az $[a, b]$-n x-től független – tehát egyenletes – korlátja. A lineáris Csebisev-approximációra nem ismeretes

olyan általános megoldási módszer, mint az L_2-normában való közelítésre.

5 AZ INTERPOLÁCIÓ

Az interpoláció alapfeladatát a következőképpen fogalmazhatjuk meg.

5.0.5. DEFINÍCIÓ. *Ismerjük egy $f : \mathbb{R} \to \mathbb{R}$ függvény*

$$a \le x_1 < x_2 < \ldots < x_n \le b$$

pontokban felvett értékeit, azaz az $y = f(x)$

$$y_i = f(x_i) \quad (i = 1, \ldots, n)$$

függvényértékeket. Az $f(x)$ függvényt, amely lehet a teljes $[a, b]$ intervallumon vagy csak $\{x_i\}_{i=1}^n$ pontokban ismert, egy olyan, általában könnyen számítható $h(x)$ függvénnyel közelítjük (vagy helyettesítjük), amelyre fennáll, hogy

$$y_i = h(x_i) \quad (i = 1, \ldots, n).$$

5.0.6. DEFINÍCIÓ. *Az* $\{x_i\}_{i=1}^n$ *pontokat* **interpolációs alappontoknak***, a feltételt* **interpolációs feltételnek** *vagy* **interpolációs alap-egyenletrendszernek** *nevezzük.*

Az interpolációs feltétel teljesülése esetén azt reméljük, hogy a

$$h(x) = h(x; \{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n)$$

interpoláló függvény az (x_i, x_{i+1}) intervallumokban jól közelíti az $f(x)$ függvényt.

Ezen a feltételen kívül további feltétel(eke)t is előírhatunk. Például, ha ismerjük az f függvény deriváltjait is az alappontokban, akkor megkövetelhetjük az

$$f'(x_i) = h'(x_i) \quad (i = 1, \ldots, n)$$

teljesülését is. A szóba jöhető $h(x)$ függvények H halmazának megválasztásától és a feltétel esetleges kibővítésétől függően beszélünk különböző típusú interpolációkról.

5.0.7. DEFINÍCIÓ. *Ha a* $h(x)$ *függvénnyel* $f(x)$*-et az* (x_1, x_n) *intervallumon kívül közelítjük, akkor* **extrapolációról** *beszélünk.*

Lineáris eset:

Ekkor tehát a H függvényhalmaz ismert $\phi_i : [a, b] \to \mathbb{R}$ $(i = 1, \ldots, n)$ függvények valamennyi lineáris kombinációja, tehát a $h(x)$ függvény alakja

$$h(x) = a_1\phi_1(x) + a_2\phi_2(x) + \ldots + a_n\phi_n(x) = \sum_{i=1}^n a_i\phi_i(x).$$

A ϕ_i függvényeket ***alapfüggvényeknek***, másképpen ***bázisfüggvényeknek*** nevezzük. Az ismeretlen $a_1, \ldots, a_n$ együtthatókat az interpolációs feltételből határozhatjuk meg. Tehát az alábbi egyenleteknek kell teljesülni:

$$\begin{array}{rcl} a_1\phi_1(x_1) + a_2\phi_2(x_1) + \ldots + a_n\phi_n(x_1) & = & f(x_1), \\ \ldots & \vdots & \ldots \\ a_1\phi_1(x_n) + a_2\phi_2(x_n) + \ldots + a_n\phi_n(x_n) & = & f(x_n). \end{array}$$

Ez egy lineáris egyenletrendszer az ismeretlen $a_1, \ldots, a_n$ együtthatókra nézve. Tömörebb felírása érdekében vezessük be a következő jelöléseket:

$$B = [\phi_j(x_i)]_{i,j=1}^n, a = [a_1, \ldots, a_n]^T, \text{ és } c = [f(x_1), \ldots, f(x_n)]^T.$$

A feltétel alakja így

$$Ba = c.$$

Ha $\det(B) \neq 0$, akkor az egyenletrendszernek pontosan egy megoldása van: $a = B^{-1}c$.

A gyakorlatban sokféle $\{\phi_i(x)\}_{i=1}^n$ bázisfüggvényt alkalmaznak. Az egyik legfontosabb a

$$\phi_1(x) = 1, \phi_2(x) = x, \ldots, \phi_n(x) = x^{n-1}$$

függvényrendszer. Ekkor beszélünk ***Lagrange-féle interpolációról***.

Az interpolációs feladat mátrixa ez esetben

$$B = \begin{bmatrix} 1 & x_1 & \ldots & x_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^{n-1} \end{bmatrix},$$

az ún. Vandermonde-féle mátrix, amely a

$$\det(B) = \prod_{1 \le i < j \le n} (x_j - x_i)$$

összefüggés miatt nemszinguláris. Tehát a Lagrange-féle interpolációs feladatnak egyértelmű megoldása van.

További fontos esetek:

- A trigonometrikus interpolációt a

$$\begin{aligned} \phi_1(x) &= 1, \\ \phi_2(x) &= \sin x,\ \phi_3(x) = \cos x, \ldots, \\ \phi_{2k}(x) &= \sin(kx), \phi_{2k+1}(x) = \cos(kx) \end{aligned}$$

$(k = 1, \ldots, (n-1)/2)$ függvényrendszer, ahol $n = 2k+1$, $[a, b] = [-\pi, \pi]$.

- Az exponenciális interpolációt a

$$\phi_i(x) = e^{\lambda_i x} \quad (i = 1, \ldots, n, \lambda_1 < \lambda_2 < \ldots < \lambda_n)$$

függvényrendszer definiálja.

- Racionális törtfüggvényeket használ a

$$\phi_i(x) = 1/(q_i + x) \quad (i = 1, \ldots, n,\ 0 < q_1 < \ldots < q_n)$$

függvényrendszer. Itt fel kell tennünk, hogy $x + q_1 > 0$. Ez könnyen teljesül, ha $x \in [a, b]$ és $a + q_1 > 0$.

Nem minden $\{\phi_i(x)\}_{i=1}^n$ függvényrendszer és $x_1 < x_2 < \ldots < x_n$ alappontok esetén van megoldása, illetve egyértelmű megoldása a lineáris interpolációs feladatnak.

5.0.8. PÉLDA. Legyen $\phi_1(x) = 1$, $\phi_2(x) = x^2$, $x_1 = -1$, $x_2 = 1$, $f(x_1) = y_1$, $f(x_2) = y_2$. Ekkor

$$B = \begin{bmatrix} 1 & (-1)^2 \\ 1 & 1 \end{bmatrix}, \quad \det(B) = 0,$$

így, ha $y_1 = y_2$, akkor végtelen sok megoldása van a feladatnak, egyébként pedig egyáltalán nincs.

5.1 LAGRANGE INTERPOLÁCIÓ

5.1.1. DEFINÍCIÓ. *Legyenek a bázisfüggvények:*

$$\phi_1(x) = 1, \phi_2(x) = x, \ldots, \phi_n(x) = x^{n-1}$$

és legyenek adottak az $x_1 < x_2 < \ldots < x_n$ *alappontok és az* $y_i = f(x_i)$ $(i = 1, \ldots, n)$ *függvényértékek. Határozzuk meg azt a legfeljebb* $(n-1)$*-ed fokú*

$$p(x) = a_0 + a_1 x + \ldots + a_{n-1} x^{n-1}$$

polinomot, amelyre teljesül a

$$y_i = p(x_i) \quad (i = 1, \ldots, n)$$

interpolációs feltétel.

Geometriailag azt jelenti, hogy illesszünk a sík n darab különböző x koordinátájú pontjára egy legfeljebb $(n-1)$-ed fokú polinomot.

A Lagrange-féle interpolációs polinom létezését és egyértelműségét már beláttuk. A polinom többféle ekvivalens alakban is felírható. Különösen fontos azonban a Lagrange-féle előállítás. Legyen

$$l_i(x) = \prod_{k=1, k\neq i}^{n} \frac{x - x_k}{x_i - x_k} \quad (i = 1, \ldots, n)$$

az i-edik ***Lagrange-féle alappolinom***. Ekkor az interpolációs polinom előáll

$$p(x) = \sum_{i=1}^{n} y_i l_i(x)$$

alakban.

Ennek igazolására vegyük észre, hogy

$$l_i(x_j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

és így teljesül, hogy:

$$p(x_j) = \sum_{i=1}^{n} y_i l_i(x_j) = y_j l_j(x_j) = y_j \quad (j = 1, \ldots, n).$$

Tehát

$$f(x) \approx h(x) = p(x) = \sum_{i=1}^{n} y_i l_i(x).$$

A Lagrange-alappolinommal való előállítás lehetőséget ad a megoldás egyértelmű létezésének közvetlen belátására is, anélkül, hogy a mátrix determinánsát ismernénk. A létezést konkrétan mutatja az előző felírás, ha pedig volna egy másik $q(x)$ legfeljebb $(n-1)$-ed fokú polinom is, mely teljesíti a feltételeket, akkor a $p(x) - q(x)$ polinomnak minden alappont zérushelye lenne. Mivel $p(x) - q(x)$ is legfeljebb $(n-1)$-ed fokú polinom, nem lehet n darab zérushelye, csak ha $p(x) \equiv q(x)$.

A Lagrange-féle interpolációs polinom hibájára vonatkozik a következő tétel:

5.1.2. TÉTEL. *Ha $f \in C^n[a,b]$, $[x_1, x_n] \subseteq [a,b]$ és $x^* \in [a,b]$, akkor*

$$f(x^*) - p(x^*) = \frac{f^{(n)}(\xi)}{n!}(x^* - x_1)(x^* - x_2)\dots(x^* - x_n),$$

ahol $\xi = \xi(x^*)$ az x^* és az x_1, x_n pontok által kifeszített intervallumban van.

B i z o n y í t á s. Ha van olyan i, hogy $x^* = x_i$,akkor állításunk triviális.

Egyébként legyen

$$\omega(x) = (x - x_1)(x - x_2)\dots(x - x_n)$$

és tekintsük a következő segédfüggvényt:

$$W(x) = f(x) - p(x) - [f(x^*) - p(x^*)]\frac{\omega(x)}{\omega(x^*)}.$$

A $W(x) \in C^n[a,b]$ függvénynek van $n+1$ zérushelye: $x^*, x_1, \dots, x_n$.

A Rolle–tétel miatt $W(x)$ bármely két zérushelye között a $W'(x)$ deriváltfüggvénynek is van zérushelye. Ezért $W'(x)$-nek legalább n zérushelye van. Hasonlóképpen okoskodva belátható, hogy $W''(x)$-nek legalább $n-1$, $W^{(3)}(x)$-nek legalább $n-2$ zérushelye van, és így tovább. Végül $W^{(n)}(x)$-nek is van legalább egy zérushelye, amit jelöljön ξ. Minthogy $p^{(n)}(x) \equiv 0$ és $\omega^{(n)}(x) \equiv n!$, ezért

$$W^{(n)}(\xi) = f^{(n)}(\xi) - [f(x^*) - p(x^*)]\frac{n!}{\omega(x^*)} = 0,$$

ahonnan átrendezéssel kapjuk a tétel állítását.

A tételt a következő formában szoktuk alkalmazni, az $x \in [a,b]$-beli hiba becslésére.

5.1.3. TÉTEL. *Legyen M_n az n-edik derivált abszolútértékének egy felső korlátja, azaz*

$$\left|f^{(n)}(x)\right| \le M_n \qquad (x \in [a,b]).$$

Ekkor

$$|f(x)-p(x)| \le \frac{M_n}{n!}\,|(x-x_1)(x-x_2)\dots(x-x_n)| \le \frac{M_n}{n!}(b-a)^n.$$

A második egyenlőtlenség általában sokkal durvább becslést ad, mint az első. Előnye, hogy x-től független, egyenletes korlátot ad a hibára a teljes $[a,b]$-n.

Konkrét n esetén szélsőérték számítással élesebb becslés is levezethető.

5.1.4. PÉLDA. Hány ekvidisztáns alappontban kell megadnunk a $\sin x$ függvény táblázatát a $\left[0, \frac{\pi}{2}\right]$ intervallumon ahhoz, hogy a közbülső pontokban lineáris Lagrange-interpolációt használva az elkövetett hiba legfeljebb $\varepsilon = 10^{-4}$ legyen?

Megoldás. Vezessük be a $h = x_{i+1} - x_i$ jelölést. A

$$|f(x)-p(x)| \le \frac{M_n}{n!}\,|(x-x_1)(x-x_2)\dots(x-x_n)| \le \frac{M_n}{n!}(b-a)^n.$$

alapján olyan h-t keresünk, melyre $(M_2h^2)/2 \le 10^{-4}$. Mivel $(\sin x)'' = -\sin x$, választhatjuk az $M_2 = 1$ értéket. Ezzel $h \le \sqrt{2}/100$, $n \ge \frac{\pi}{2h}$ miatt $n \ge 112$ adódik.

Ha viszont a hibakorlátot az

$$|f(x)-p(x)| \le \frac{M_2}{2}\max|(x-x_i)(x-x_{i+1})|$$

becslésből közvetlenül vezetjük le szélsőérték számítással, akkor az élesebb,

$$|f(x)-p(x)| \le \frac{M_2h^2}{8}$$

eredményt kapjuk. Ez alapján kiderül, hogy $n = 28$ pont is elég.

Az interpolációs eljárásoktól elvárjuk, hogy a pontok számának növelése esetén a közelítés hibája csökken. Ez azonban nem minden esetben van így, amint azt Runge kimutatta az $f(x) = 1/(1+x^2)$ függvénynek az $[a,b] = [-5,5]$ intervallumon, egyre növekvő fokszámú Lagrage interpolációs polinommal való közelítésével.

Példa: Tegyük fel, hogy az $y_i = f(x_i)$ függvény értékeket ε_i hibával ismerjük $(i = 1,\dots,n)$. Ekkor az elméleti

$$p(x) = \sum_{i=1}^{n} f(x_i)l_i(x)$$

Lagrange-interpolációs polinom helyett a perturbált

$$\tilde{p}(x) = \sum_{i=1}^{n} (f(x_i) + \varepsilon_i) l_i(x)$$

polinommal számolunk.

A kettő eltérésére teljesül, hogy

$$\begin{aligned} \delta\left(p\left(x\right)\right) &= |\tilde{p}(x) - p(x)| = \left| \sum_{i=1}^{n} \varepsilon_i l_i(x) \right| \leq \\ &\leq \sum_{i=1}^{n} |\varepsilon_i| \, |l_i(x)| \leq \left(\max_{1 \leq i \leq n} |\varepsilon_i| \right) \sum_{i=1}^{n} |l_i(x)| \, . \end{aligned}$$

Ez a becslés pontos. Igazolható, hogy

$$\sum_{i=1}^{n} |l_i(x)| > \frac{2}{\pi} \log n + c,$$

ahol c konstans. Ha n elég nagy, akkor a $\delta\left(p\left(x\right)\right)$ perturbációs hiba is nagy lesz.

A divergencia és numerikus instabilitás miatt sok esetben – mint már említettük – más típusú interpolációs eljárásokat használunk.

5.1.5. PÉLDA. Közelítsük másodfokú függvénnyel az $f(x) = \cos(\frac{\pi}{2}x)$ függvényt a $[-1, 1]$ intervallumon az $x_1 = -1, x_2 = 0, x_3 = 1$ pontokra támaszkodva.

Megoldás:

Ekkor $f(x) \approx p(x) = A_1 + A_2 x + A_3 x^2$. Az együtthatókra felírható az

$$\begin{aligned} A_1 - A_2 + A_3 &= 0 \\ A_1 &= 1 \\ A_1 + A_2 + A_3 &= 0 \end{aligned}$$

egyenletrendszer. Innen $p(x) = 1 - x^2$.

Természetesen ugyanezt kapjuk az $l_i(x)$ Lagrange-függvényekkel is. Az előállítás szerint most $f(x_1) = f(x_3) = 0$ miatt elég az $l_2(x)$-t meghatározni, ez $1 - x^2$, ami jelen esetben a $p(x)$ polinommal megegyezik.

A közelítés hibáját

$$h \leq \frac{M_3}{3!} \max_{-1 \leq x \leq 1} |(x+1)x(x-1)|$$

becsli, ahol M_3 az $|f'''(x)|$ maximuma, jelen esetben $\pi^3/8$. Szélsőérték-számítással adódik, hogy

$$\max_{-1 \leq x \leq 1} |(x+1)x(x-1)| = \frac{8}{27},$$

azaz $h \leq \pi^3/216 \simeq 0.15$.

5.2 HERMITE INTERPOLÁCIÓ

Tegyük fel, hogy az $x_0, x_1, \ldots, x_k \in [a,b]$ különböző alappontok ($k \leq n$), továbbá $m_0, m_1, \ldots, m_k \in \mathbb{N}$ multiplicitások úgy, hogy

$$\sum_{i=0}^{k} m_i = n+1.$$

Legyenek adottak

$$f^{(j)}(x_i) = y_{ij}, (i = 0, \ldots, k \text{ és } j = 0, \ldots, m_i - 1)$$

értékek. Egy olyan n-edfokú P polinomot keresünk, melyre

$$P^{(j)}(x_i) = y_{ij}, (i = 0, \ldots, k \text{ és } j = 0, \ldots, m_i - 1)$$

5.2.1. TÉTEL. *Egyértelműen létezik egy olyan n-edfokú P polinom, melyre*

$$P^{(j)}(x_i) = y_{ij}, (i = 0, \ldots, k \textit{ és } j = 0, \ldots, m_i - 1)$$

A Hermite interpolációs polinom hibájára vonatkozik a következő tétel:

5.2.2. TÉTEL. *Ha $f \in C^{n+1}[a,b]$, és jelölje $I \subset [a,b]$ az $x_1, x_2, \ldots, x_k$ alappontok által kifeszített intervallumot, ekkor minden $x^* \in [a,b]$ esetén létezik $\xi = \xi(x^*) \in I$, hogy*

$$f(x^*) - P(x^*) = \frac{f^{(n)}(\xi)}{n!}\omega(x^*)$$

ahol

$$\omega(x^*) = \prod_{i=0}^{k}(x^* - x_i)^{m_i}.$$

Ha $\sup_{x\in[a,b]} |f^{(n)}(x)| =: M_{n+1} < \infty$, akkor

$$f(x^*) - P(x^*) \leq \frac{M_{n+1}}{n!}\omega(x^*)$$

5.2.3. MEGJEGYZÉS.

- Ha $m_i = 1$ minden i-re, akkor a Lagrange-féle interpolációt kapjuk.
- Ha $m_i = 2$ minden i-re, akkor a Fejér-Hermite interpolációt kapjuk, melynek ismert az explicit alakja és az a Lagrange-alappolinomok segítségével is felírható.

5.2.4. DEFINÍCIÓ. *(elsőrendű osztott differenciák)*
Legyenek $x_1, x_2, \ldots, x_n$ különböző pontok. Ekkor

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \qquad (i = 1, \ldots, n-1),$$

5.2.5. DEFINÍCIÓ. *(k-adrendű osztott differenciák)*
Legyenek $x_1, x_2, \ldots, x_n$ különböző pontok. Ekkor

$$f[x_i, \ldots, x_{i+k}] = \frac{f[x_{i+1}, \ldots, x_{i+k}] - f[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}$$

$(k = 1, \ldots, n,\ i = 1, \ldots, n-k)$.

A Lagrange-féle interpoláció Newton-alakja

$$P(x) = f(x_1) + \sum_{k=2}^{n} f[x_1, \ldots, x_k] \prod_{j=1}^{k-1} (x - x_j)$$

Azonos alappontok esetén az osztott differencia nem definiálható a szokásos formulával. A forgalom kiterjesztése határátmenettel történik.

5.2.6. DEFINÍCIÓ. *(Osztott differenciák ismétlődő argumentumokkal)*

$$f[x, x] = \lim_{y \to x} \frac{f(x) - f(y)}{x - y} = f'$$

Hasonlóan a $k-1$-edrendű osztott differenciára adódik, hogy

$$f[x, x, \ldots, x] = \frac{f^k(x)}{k!}$$

A többi, azonos alappontokat is tartalmazó osztott differencia a fentiekből a szokásos definícióval számolható.

Hermite-interpolációs polinom előállítása Newton alakkal:
Az osztott differencia táblázat felépítése:

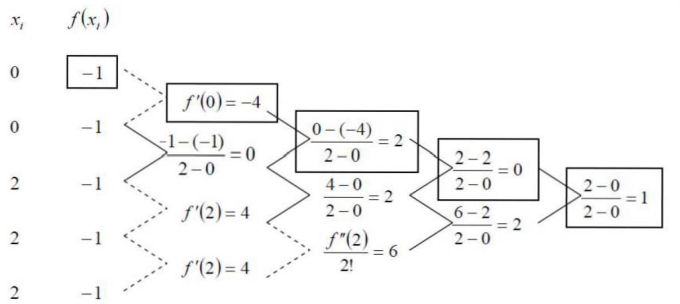
- Minden alappontot annyiszor veszünk fel egymás után, amennyi a multiplicitása, pl. az x_i-t $m_i + 1$-szer egymás után.
- Beírjuk az $f(x_i)$ értékeket és a megfelelő $\frac{f^{(j)}(x_i)}{j!}$ értékeket az x_i alappontra támaszkodó $j-1$-edrendű osztott differenciák helyére ($f^{(j)}(x_i)$ ismeretében).
- A táblázat többi részét az osztott differencia fogalom definíciója alapján töltjük ki.

- Az interpolációs polinom felírása a táblázatból a szokásos módon történik, a táblázat átlójában szereplő elemek, mint együtthatók segítségével.

Hermite-interpolációs polinom előállítása Lagrange alakkal:
Felírjuk a megfelelő alappolinomokat, majd mindegyiket a megfelelő $f^{(j)}(x_i)$-vel szorozva az interpolációs polinomot. Az alappolinomok előállítása bonyolult, az alappontok illetve a multiplicitás értékektől függ az alakjuk.

5.2.7. PÉLDA. Mi lesz az f-et közelítő Hermite-interpolációs polinom, ha $f(0) = -1, f(2) = -1, f'(0) = -4, f'(2) = 4$ és $f''(2) = 12$

Megoldás: Elkészítjük az osztott differencia táblázatot.



$$\begin{aligned} P(x) &= -1 + (-4)(x-0) + 2(x-0)^2 + 0(x-0)^2(x-2) \\ &\quad + 1(x-0)^2(x-2)^2 = -1 - 4x + 2x^2 + x^2(x-2)^2 \\ &= x^4 - 4x^3 + 6x^2 - 4x - 1 \end{aligned}$$

5.3 SPLINE INTERPOLÁCIÓ

A Spline interpoláció is a lineáris interpolációk közé tartozik alkalmasan megválasztott $\{\phi_i\}_{i=1}^n$ bázisfüggvény-rendszerrel.

5.3.1. DEFINÍCIÓ. *A $h(x)$ interpoláló függvényt szakaszonként adjuk meg speciális csatlakozási feltételekkel. Az $a = x_1 < x_2 < \ldots < x_n = b$ alappontokhoz, illetve az $y_i = f(x_i)$ $(i = 1, \ldots, n)$ függvényértékekhez olyan $S(x)$ függvényt keresünk, amely kielégíti a következő feltételeket:*

(1) $S(x) = S_i(x) \qquad (x \in [x_i, x_{i+1}])$

(2) $S(x_i) = y_i \qquad (i = 1, \ldots, n)$,

(3) $S_i(x_{i+1}) = S_{i+1}(x_{i+1}) \qquad (i = 1, \ldots, n-2)$

5.3.2. MEGJEGYZÉS.

- Az (1) feltétel tulajdonképpen csak megfogalmazza, hogy szakaszokból áll az interpolációs függvényünk, azaz részintervallumonként más-más előírással definiáljuk.
- A (2) feltétel azt jelenti, hogy az $S(x)$ valóban interpoláló függvény.
- A (3)-mal a folytonosságot biztosítjuk az $[a, b]$ intervallumon (Ezért nincs ellentmondás (1)-ben: a közbülső osztópontok két szakaszhoz is tartoznak, de ott mindkét szakasz ugyanolyan értékű.)
- Ez a három feltétel tulajdonképpen a Spline általános definíciója. Az egyes S_i függvényeket akár más-más H_i osztályból választhatjuk, azonban rendszerint egyetlen H halmazt írunk elő: $S_i \in H$. A H-tól függően további feltételeket is előírhatunk (gyakran kötelező is, hogy egyértelmű megoldást kapjunk).
- A H lehet például az $y = \alpha + \beta x$ elsőfokú polinomok (az egyenesek) osztálya. Ez esetben nem is kell további feltétel; a megoldást az (1)-(3) feltételek egyértelműen biztosítják.
- Magasabb fokszámú polinomok esetén a közbülső pontokban bizonyos rendig a csatlakozó ágak deriváltjainak megegyezését is megkövetelhetjük. Azt mondjuk, hogy egy **Spline k-ad fokú és m-ed rendű**, ha szakaszonként legfeljebb k-adfokú polinomokból áll, és a közbülső pontokban a jobb- és baloldali szakasz deriváltjai m rendig megegyeznek.

A spline meghatározásakor n db k-adfokú polinomot kell felírnunk, azaz az ismeretlenek száma: $n(k+1)$.

Az (1), (3) feltételekből a feltételek száma : $(k+1)n - (k-1)$, ugyanis $k-1$ db simasági feltétel az $n-1$ belső pontban, azaz $(n-1)(k-1)$ és $2n$ db interpolációs feltétel (n db intervallum két végpontja).

Összesen $(n-1)(k-1)+2n = (k+1)n-(k-1)$. Innen látszik, hogy $k-1$ db feltétel hiányzik a Spline egyértelműségéhez. Ezeket a feltételeket általában a végpontokra adják meg.

Lineáris spline (k=1):

Az (1), (2) és (3) egyértelműen meghatározza. Minden $[x_k, x_{k+1}]$ intervallumon

$$\begin{aligned} S_k(x_k) &= a_k x_k + b_k = y_k \\ S_k(x_{k+1}) &= a_k x_{k+1} + b_k = y_{k+1} \end{aligned}$$

Ebből az két ismeretlenes egyenletrendszerből a_k és b_k meghatározható.

Úgy is meghatározhatjuk a P_k polinomot, hogy az $[x_k, x_{k+1}]$ intervallum végpontjaira felírjuk a lineáris interpolációs polinom Lagrange-féle alakját. Az interpoláció miatt $S \in C[a, b]$.

Kvadratikus spline (k=2):

A $k - 1 = 1$ feltétel hiányzik a spline egyértelmű felírásához. Ezt általában az intervallum elején vagy végén a derivált megadásával szokás teljesíteni. Ebben az esetben az egymás melletti intervallumokra Hermite interpolációt alkalmazva meghatározható a spline.

5.3.1 KÖBÖS MÁSODRENDŰ SPLINE

A gyakorlatban túlnyomórészt harmadfokú, másodrendű Spline interpolációt használunk, és röviden csak **harmadfokú Spline-ról** beszélünk. Ekkor a további feltételek, amelyeket megkövetelünk:

(4) $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}) \qquad (i = 1, \ldots, n-2)$,

(5) $S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}) \qquad (i = 1, \ldots, n-2)$,

(6) $S''(x_1) = A_n$, és $S''(x_n) = B_n$

Ha $A_n = B_n = 0$, akkor az $S(x)$ függvényt **természetes Spline-nak** nevezzük.

Az $S(x)$ Spline-t az $[x_i, x_{i+1}]$ intervallumon a

$$S(x) = S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3$$

alakban keressük ($i = 1, \ldots, n-1$). Leszámlálhatjuk, hogy a $4(n-1)$ darab ismeretlenhez a (2) – (6) feltételek pontosan ugyanannyi egyenletet szolgáltatnak, és az egész együtt egy lineáris egyenletrendszer. Az x^k hatványok helyett az $(x - x_i)^k$ használatának praktikus okai vannak.

Jelöljük h_i-vel az i-edik szakasz hosszát, azaz legyen $h_i = x_{i+1} - x_i$ ($i = 1, \ldots, n-1$). A (2) – (6) feltételek felhasználásával az ismeretlen a_i, b_i, c_iés d_i együtthatókat a következőképpen határozhatjuk meg.

A (2), azaz az $S(x_i) = S_i(x_i) = y_i$ interpolációs feltétel miatt

$$a_i = y_i \qquad (i = 1, \ldots, n-1)$$

(ebből látszik, hogy mi az $(x - x_i)$ eltolás értelme),

valamint:

$$S_{n-1}(x_{n)} = a_{n-1} + b_{n-1}h_{n-1} + c_{n-1}h_{n-1}^2 + d_{n-1}h_{n-1}^3 = y_n.$$

A (3)-as csatlakozási feltétel alakja:

$$S_i(x_{i+1}) = y_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = y_{i+1} \qquad (i = 1, \ldots, n-2),$$

A (4)-es csatlakozási feltétel:

$$S_i'(x_{i+1}) = b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1} = S_{i+1}'(x_{i+1}) \qquad (i = 1, \ldots, n-2).$$

Hasonlóképpen kapjuk, hogy az (5)-ös feltétel alakja:

$$S_i''(x_{i+1}) = 2c_i + 6d_i h_i = 2c_{i+1} = S_{i+1}''(x_{i+1}) \qquad (i = 1, \ldots, n-2).$$

Végül a (6)-os végpont-feltétel:

$$S''(x_1) = 2c_1 = A_n, \quad S''(x_n) = 2c_{n-1} + 6d_{n-1}h_{n-1} = B_n.$$

Így $4n-4$ darab egyenlet kapunk, ahonnan az a_i és a c_1 értékei azonnal adódtak. A többiből a b_i és d_i ismeretleneket $(i = 1, \ldots, n-1)$ fokozatosan ki lehet fejezni és a többibe behelyettesíteni. Végül a következőhöz jutunk:

$$\begin{aligned} b_i &= \frac{y_{i+1} - y_i}{h_i} - \frac{2c_i + c_{i+1}}{3} h_i \qquad (i = 1, \ldots, n-2) \\ b_{n-1} &= \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{4c_{n-1} - B_n}{6} h_{n-1}, \end{aligned}$$

továbbá

$$\begin{aligned} d_i &= \frac{c_{i+1} - c_i}{3h_i} \qquad (i = 1, \ldots, n-2) \\ d_{n-1} &= \frac{B_n - 2c_{n-1}}{6h_{n-1}}. \end{aligned}$$

A c_i együtthatók ismeretében tehát valamennyi ismeretlen közvetlenül számolható. Minthogy $c_1 = A_n$ már ismert, így egy $n-2$ ismeretlenes egyenletrendszert kell megoldani, ami a

$$\Delta_i = (y_{i+1} - y_i)/h_i, \lambda_i = h_{i+1}/(h_i + h_{i+1}), \mu_i = 1 - \lambda_i$$

$(i = 1, \ldots, n-2)$ bevezetésével a következő:

$$2c_2 + \lambda_1 c_3 = \frac{3}{h_1 + h_2}\left(\Delta_2 - \Delta_1 - \frac{h_1}{2} A_n\right), \tag{4}$$

$$\mu_i c_i + 2c_{i+1} + \lambda_i c_{i+2} = \frac{3}{h_i + h_{i+1}} \left(\Delta_{i+1} - \Delta_i\right), \tag{5}$$

$(i = 2, \ldots, n-3)$ és

$$\mu_{n-2} c_{n-2} + 2c_{n-1} = \frac{3}{h_{n-2} + h_{n-1}} \left(\Delta_{n-1} - \Delta_{n-2} + \frac{h_{n-1}}{6} B_n\right). \tag{6}$$

Ez egy $n-2$ ismeretlenes lineáris egyenletrendszer a $c_2, c_3, \ldots, c_{n-1}$ ismeretlenekre. (mivel az a_i $(i = 1, \ldots, n-1)$ és a c_1 együtthatók közvetlenül adódnak).

Az $n = 3$ esetén () és () nincs értelmezve, c_2 ekkor ()-ból számolható.

Az $n = 4$ esetén pedig () nem értelmezett, a $c_2$és c_3 a másik két egyenletből határozható meg. Ekkor az együttható mátrix:

$$A = \begin{bmatrix} 2 & \lambda_1 \\ \mu_2 & 2 \end{bmatrix}.$$

Az egyenletrendszer mátrixa $n > 4$ esetén pedig

$$A = \begin{bmatrix} 2 & \lambda_1 & 0 & & \ldots & 0 \\ \mu_2 & 2 & \lambda_2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & & 0 \\ \vdots & \ddots & \ddots & \ddots & & \\ & & & \mu_{n-3} & 2 & \lambda_{n-3} \\ 0 & & \ldots & 0 & \mu_{n-2} & 2 \end{bmatrix},$$

egy három átlóból álló sávmátrix. Ez a mátrix nemszinguláris (ez bizonyítani például a Gersgorin tétel segítségével lehet).

Minthogy az A mátrix diagonálisan domináns is, ezért az egyenletrendszer főelemkiválasztás nélkül Gauss-eliminációval megoldható $O(n)$ régi flop művelettel.

5.3.3. PÉLDA. Tegyük fel, hogy az f függvényről a következő táblázat áll rendelkezésünkre:

x	-1	0	1
$y = f(x)$	1	0	1

Adjuk meg hozzá a természetes Spline-t!

Megoldás. Természetes Spline esetén $A = B = 0$. Ekkor $a_i = y_i$ miatt $a_1 = 1$ és $a_2 = 0$.

A (6)-os végpont-feltételből pedig $c_1 = 0$ adódik.

Mivel $h_1 = h_2 = 1$ és $\Delta_1 = -1, \Delta_2 = 1$, továbbá μ_1-et a $c_1 = 0$ miatt ki sem kell számolni (egyébként $\mu_1 = 1 - \lambda_1 = 1 - 1/2 = 1/2$).

A () miatt

$$2c_2 = \frac{3}{1+1}\left(1 - (-1) + \frac{1}{6}0\right) = 3 \to c_2 = \frac{3}{2}$$

Továbbá a b-re vonatkozó egyenletekből $b_1 = -3/2$ és $b_2 = 0$ adódik; a d-re vonatkozóakból pedig $d_1 = 1/2$ és $d_2 = -1/2$.

A Spline két tagja tehát

$$\begin{aligned} S_1(x) &= 1 - \frac{3}{2}(x+1) + \frac{1}{2}(x+1)^3 = \frac{3}{2}x^2 + \frac{1}{2}x^3 \\ S_2(x) &= \frac{3}{2}x^2 - \frac{1}{2}x^3 \end{aligned}$$

5.3.4. TÉTEL. *Az (1) – (6) feltételekkel meghatározott Spline létezik és egyértelmű. Ha $f \in C^2[x_1, x_n]$, akkor létezik $K > 0$ konstans, amellyel fennáll a következő:*

$$|f(x) - S(x)| \le K\left(\max_{1\le i\le n-1} h_i\right)^2 \qquad (x \in [x_1, x_n]).$$

Ha pedig $f \in C^3[x_1, x_n]$, $A_n = f''(x_1)$ és $B_n = f''(x_n)$, akkor létezik $\widetilde{K} > 0$ konstans, hogy

$$|f(x) - S(x)| \le \widetilde{K}\left(\max_{1\le i\le n-1} h_i\right)^3 \qquad (x \in [x_1, x_n]).$$

5.3.5. MEGJEGYZÉS. Ha az $f''(x_1)$ és $f''(x_n)$ információk nem állnak rendelkezésre, akkor a természetes Spline-t definiáló $A_n = B_n = 0$ választással élünk. A közelítés hibájára ekkor az első becslés érvényes. A természetes Spline az

$$\int_a^b [s''(x)]^2\, dx \to \min \qquad (s(x_i) = y_i, i = 1, \dots, n)$$

feltételes szélsőértékfeladat megoldása. A természetes elnevezés pedig a mechanikai tartalmára utal. Ugyanis, ha tekintünk egy ideálisan rugalmas rudat (Spline-t), amely átmegy az (x_i, y_i) pontokon (gömbcsúszkákon, surlódás nélkül), akkor a legkisebb alakváltozási energia mechanikai elve miatt a Spline azt az alakot veszi fel, amely az előző kifejezést minimalizálja. A (6) végpont-feltétel helyett más kikötések is lehetségesek. A Spline függvények előnyei: gyors és numerikusan stabil kiszámítás, nagyon jó közelítési tulajdonságok. Hátrányuk a bonyolult megadás, de ez számítógépek használata esetén nem jelent komoly problémát.

6 MÁTRIXOK SAJÁTÉRTÉKEI, SAJÁTVEKTORAI

Szükségünk van a komplex elemű mátrixok és vektorok bevezetésére.

A komplex elemű n-dimenziós oszlopvektorok halmazát $\mathbb{C}^n$-el jelöljük.

Hasonlóképpen az $m \times n$ méretű komplex elemű mátrixok halmazát $\mathbb{C}^{m\times n}$ jelöli.

Nyilvánvalóan tekinthetjük úgy, hogy $\mathbb{R}^n \subset \mathbb{C}^n$ és $\mathbb{R}^{m\times n} \subset \mathbb{C}^{m\times n}$.

A valós elemű vektorokra és mátrixokra bevezetett műveletek és a determináns értelmezése ugyanaz a komplex esetben is, mint valósban. A komplex vektorok és mátrixok normájának definíciója is változatlan marad. (Ezért kell a normák definíciójában négyzetek helyett (valós számok esetén feleslegesen) abszolút érték négyzeteket.

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

$$Ax = \lambda x. \tag{7}$$

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

6.0.7. MEGJEGYZÉS. Felhívjuk a figyelmet az $x \neq 0$ kikötésre, amiről a hallgatók többnyire elfeldkeznek. Pedig enélkül az egésznek semmi értelme, hiszen $A0 = \lambda 0$, ahol λ akármi lehetne.

A sajátérték-feladat tehát olyan λ (valós vagy komplex) szám(ok) keresését jelenti melyekre az () egyenletnek van nemzérus megoldása.

Átrendezés után az eredeti egyenlettel ekvivalens $(A - \lambda I)x = 0$ alakot kapjuk. Ennek a homogén egyenletrendszernek keressük tehát a nemtriviális megoldásait, ami akkor és csak akkor létezik, ha $\det(A - \lambda I) = 0$.

6.0.8. DEFINÍCIÓ. *A*

$$\phi(\lambda) = \det(A - \lambda I) = \det\left(\begin{bmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{bmatrix}\right) = 0 \tag{8}$$

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

$$\phi(\lambda) = (-1)^n\lambda^n + p_{n-1}\lambda^{n-1} + \ldots + p_1\lambda + p_0 \tag{9}$$

karakterisztikus polinomot *kapjuk.*

Az algebra alaptétele szerint a multiplicitásokat is figyelembe véve egy n-ed fokú polinomnak a komplex számok körében pontosan n zérushelye van. (Ezért kellett tehát kilépnünk a valós számok köréből.) Egy λ_i gyök multiplicitásán azt értjük, hogy a polinom gyöktényezős alakjában a $(\lambda - \lambda_i)$ milyen hatványkitevővel szerepel.

6.0.9. TÉTEL. *Egy $A \in \mathbb{C}^{n\times n}$ mátrixnak a multiplicitásokat is figyelembe véve pontosan n sajátértéke van.*

Egy mátrix sajátértékeinek összességét a **mátrix spekrumá**nak nevezzük. A spektrumból a mátrixnak nagyon sok fontos tulajdonsága kiolvasható. Csupán kettőt említve:

- (i) Egy négyzetes mátrix akkor és csak akkor nemszinguláris ha egyik sajátértéke sem zérus.
- (ii) Egy mátrix akkor és csak akkor pozitív definit, ha minden sajátértéke pozitív.

A sajátvektorok darabszámára már nem lehet olyan kijelentést tenni, mint a sajátértékekére. Néhány fontos tulajdonság felsorolásával jellemezhetjük a viszonyokat.

- Ha x a λ sajátértékhez tartozó sajátvektor, akkor bármilyen $t \in \mathbb{C}$ $(t \neq 0)$ mellett tx is az. Ez a definíciós () egyenletbe való behelyettesítésből azonnal kiderül. Tehát az adott λ sajátérték esetén a hozzátartozó lineárisan független sajátvektorokat kell meghatározni, mert azok összes lineáris (de zérustól különböző) kombinációja adja meg a λ-hoz tartozó összes sajátvektort.
- Egy adott λ-hoz tartozó lineárisan független sajátvektorok száma legfeljebb annyi, mint a λ multiplicitása. (Ennek belátása nem egyszerű, itt mellőzzük az igazolását.)
- Különböző sajátértékekhez tartozó sajátvektorok lineárisan függetlenek. Ez a tény azonnal következik az első tulajdonságból.

Ezek után tétel formában megemlítünk a sajátértékek tulajdonságai közül is néhányat, melyek a numerikus értékük meghatározásában, becslésében fontos szerepet játszanak.

6.0.10. TÉTEL. *Ha λ az A mátrix sajátértéke és x egy hozzátartozó sajátvektor, akkor tetszőleges $\sigma \in \mathbb{C}$ esetén az $A-\sigma I$ mátrixnak sajátértéke a $\lambda-\sigma$ és ugyanaz az x egy hozzátartozó sajátvektor.*

Bizonyítás. $(A - \sigma I)\, x = Ax - \sigma Ix = \lambda x - \sigma x = (\lambda - \sigma)\, x.$

A-hoz a σI hozzáadását a spektrum eltolásának is nevezzük .

6.0.11. TÉTEL. *Ha az A mátrix sajátértékei $\lambda_1, \lambda_2, \ldots, \lambda_n$, akkor az A^k mátrix sajátértékei $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$.*

Bizonyítás. Legyen λ akármelyik sajátérték. Ekkor $A^k x = A(...(A(A))...)x = A(...(A)...)(\lambda \cdot \lambda)x = \lambda^k x$.

A tétel kiterjeszthető A tetszőleges polinomjára is.

6.0.12. TÉTEL. *Ha az $A \in \mathbb{C}^{n\times n}$ nemszinguláris mátrixnak λ sajátértéke és x egy hozzátartozó sajátvektor, akkor az A^{-1}-nek sajátértéke az $1/\lambda$ és x itt is hozzátartozó sajátvektor.*

Bizonyítás. $Ax = \lambda x \to A^{-1}Ax = A^{-1}\lambda x \to x = A^{-1}\lambda x$. Osszunk át a nemzérus λ-val.

6.0.13. TÉTEL. *Legyen λ az A mátrix akármelyik sajátértéke. Tetszőleges indukált mátrixnormában fennáll, hogy $|\lambda| \leq \|A\|$.*

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

Bármely λ tehát egy olyan origó közepű körben helyezkedhet el, amelyik sugara egyetlen indukált normánál sem nagyobb.

A következő tétel egy még általánosabb tartományra, n darab körlap egyesítésére szűkíti a sajátértékek lehetséges előfordulását.

6.0.14. TÉTEL. Gersgorin Tétel
Legyen $A \in \mathbb{C}^{n\times n}$, továbbá

$$r_i = \sum_{j=1, j\neq i}^{n} |a_{ij}| \quad (i = 1, \ldots, n)$$

az i-edik kör sugara. Legyen az i-edik körlap:

$$D_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq r_i\} \quad (i = 1, \ldots, n)\,.$$

Ekkor az A mátrix minden λ sajátértékére fennáll, hogy

$$\lambda \in \cup_{i=1}^{n} D_i.$$

B i z o n y í t á s. Legyen $v = [v_1, \ldots, v_n]^T$ az egyik, tetszőleges λ-hoz tartozó sajátvektor, i pedig az az index, amelyre fennáll, hogy $\|v\|_\infty = |v_i|$. ($|v_i|$ értéke $v \neq 0$ miatt nyilván nem zérus.) Az $Av = \lambda v$ egyenletrendszer i-edik egyenlete

$$\lambda v_i = a_{ii} v_i + \sum_{j=1, j\neq i}^{n} a_{ij} v_j,$$

ahonnan átrendezéssel

$$|\lambda v_i - a_{ii} v_i| = |\lambda - a_{ii}|\,|v_i| = \left| \sum_{j=1, j\neq i}^{n} a_{ij} v_j \right| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|\,|v_i|$$

adódik.

Mindkét oldalt $|v_i|$-vel elosztva kapjuk, hogy

$$|\lambda - a_{ii}| \leq r_i.$$

Tehát minden λ sajátérték benne van valamelyik D_i körlemezben, az összes sajátérték benne van az egyesítésükben.

6.0.15. PÉLDA. Legyen

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

- (i) Határozzuk meg a sajátértékeit, azok multiplicitását és a hozzájuk tartozó lineárisan független, vé gtelen normában egységnyi sajátvektorokat.
- (ii) Feltéve, hogy nem tudjuk az előző kérdésre a választ, de ismerjük az A tanult normáit, ábrázoljunk a komplex számsíkon minél szűkebb olyan tartományt, amelybe biztosan beleesik mindegyik sajátérték. ($\|A\|_2 = 7.47$ kerekítve.)

Megoldás: A karakterisztikus polinom, illetve egyenlet

$$\det\left(\begin{bmatrix} 3-\lambda & 1 & 2 \\ 0 & 1-\lambda & 0 \\ -5 & -1 & -4-\lambda \end{bmatrix} \right) = -\lambda^3 + 3\lambda - 2 = 0.$$

A gyöktényezős alak: $-(\lambda+2)(\lambda-1)^2 = 0$. Vagyis a sajátértékek: $\lambda_1 = -2$, egyszeres és $\lambda_2 = 1$ kétszeres. Felírva a két sajátértékhez tartozó egyenletet:

$$\begin{bmatrix} 5 & 1 & 2 \\ 0 & 3 & 0 \\ -5 & -1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0 \quad \text{és} \quad \begin{bmatrix} 2 & 1 & 2 \\ 0 & 0 & 0 \\ -5 & -1 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

Az első egyenletrendszerből független (például) az első kettő, ennek megodása:

$$[-0.4t, 0, t]^T,$$

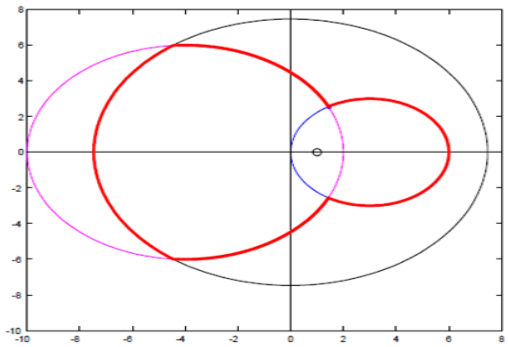
a másodiké:

$$[\tau, 0, -\tau]^T$$

($t, \tau \neq 0$, egyébként tetszőleges komplex számok). A feladat kívánalmának megfelelnek, ha $t, \tau = 1$.

Az A indukált normái: $\|A\|_1 = 8$, $\|A\|_\infty = 10$, $\|A\|_2 = 7.47$. Ezek közül a 7.47 a legkisebb, tehát a tétel szerint minden sajátérték benne van az origó közepű, ilyen sugarú körben. Írjuk fel a Gersgorin tételben szereplő körök egyenletét:

$(x-3)^2 + y^2 = 3^2$; $(x-1)^2 + y^2 = 0^2$; $(x+4)^2 + y^2 = 6^2$



Az ábrán jól láthatóak az egyes körlapok (a második egyetlen pont). A Gersgorin tételben szereplő körlapok uniójának és a tételben említett körlapnak a metszetét vastagabb piros vonallal határoltuk. Látható, hogy azért elég durva lehet ez a becslés.

A sajátérték-feladat megoldása elvileg is nehéz, numerikusan méginkább. A karakterisztikus egyenlet megoldása nagyobb dimenzióban, majd adott esetben az igen nagyméretű szinguláris mátrixú egyenlet megoldása komoly gondot jelenthet. Éppen ezért elméleti jelentősége mellett a gyakorlat számára is fontos kérdés, hogy adott mátrixhoz lehet-e olyan másikat találni, amelynek sajátértékei megegyeznek az eredetivel. Az új mátrix sajátértékei esetleg könnyebben meghatározhatók.

6.0.16. DEFINÍCIÓ. *Legyen $A \in C^{n\times n}$ tetszőleges mátrix és $T \in C^{n\times n}$ nemszinguláris. Ekkor az $A \to T^{-1}AT$ leképezést hasonlósági transzformációnak nevezzük, és ha $B = T^{-1}AT$, akkor azt mondjuk, hogy A és B hasonló.*

6.0.17. TÉTEL. *Ha A és B hasonlóak, azaz $B = T^{-1}AT$ valamely nemszinguláris T mátrixszal, akkor A és B sajátértékei megegyeznek. Továbbá, ha egy λ sajátértékhez az A mátrix x sajátvektora tartozik, akkor az $y = T^{-1}x$ a B sajátvektora.*

B i z o n y í t á s. Definíció szerint

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

ami bizonyítandó volt.

Vannak mátrixok, amelyek spektruma a mátrixból kiolvasható. Azonnal látszik, hogy a diagonálmátrixok sajátértékei a főátló elemei, de hasonlóképp rögtön látható, hogy ugyanez a helyzet a háromszögmátrixoknál is. Azoknál is a főátlóbeli elemek szorzata adja a determinánst és az $A - \lambda I$ olyan háromszögmátrix, melynek főátlóját az $a_{ii} - \lambda$ értékek alkotják. Felmerül tehát az ilyen alakú mátrixokra való hasonlósági transzformáció kérdése.

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

6.0.19. MEGJEGYZÉS. A diagonalizálhatóság pontos feltételét meglehetősen bonyolult ellenőrizni, kivitelezése ugyancsak. Éppen ezért inkább elméleti jelentőségű. Háromszögmátrixra való hasonlósági transzformációt szintén nehéz találni. Viszont ismerünk olyan numerikus eljárást, amelyik hasonlósági transzformációk végtelen sorozatával "kinullázza" a főátló alatti elemeket, miközben a diagonális elemek is konvergálnak. Vagyis elég messze elmenve a sorozattal, a főátlóban lévő elemek az eredeti mátrix sajátértékeinek jó közelítései, függetlenül attól, hogy a főátló fölötti elemek konvergálnak-e vagy sem.

Fenti megjegyzésünkben szereplő eljárást nem ismertetjük. Ismertetjük viszont azt, amiből általánosítással származtatható az említett (a főátló alatt "nullázó") módszer.

6.1 A HATVÁNYMÓDSZER

Sok gyakorlati problémánál a legnagyobb abszolút értékű sajátértéknek döntő szerepe van, ezért azt domináns sajátértéknek nevezzük. A sajátértékek sorszámozása akaratlagos, de rendszerint úgy számozzuk, hogy teljesüljön a következő: $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. Ekkor tehát a λ_1 a domináns. Általában mindegyik sajátértéket felsoroljuk. Valamely sajátérték multiplicitást úgy vesszük figyelembe, hogy annyiszor szerepel, amennyi a multiplicitása. A sajátvektorok közül is megadunk minden felsorolt sajátértékhez egyet, tehát n darabot adunk meg: $x_1, x_2 \ldots, x_n$ (akkor is, ha azok nem feltétlenül lineárisan függetlenek). A domináns sajátérték közelítő meghatározására szolgáló egyik módszer alapgondolata von Mieses-től származik, ezért szokás Mieses-módszernek is nevezni.

Tegyük fel, hogy az $A \in \mathbb{R}^{n\times n}$ valós mátrix sajátértékei is valósak, a domináns pedig egyedüli, azaz

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|\,. \tag{10}$$

Tegyük továbbá fel, hogy a $v^{(0)}$ kezdővektort sikerült úgy megválasztani, hogy előállítható a sajátvektorok olyan lineáris kombinációjaként, amelyben x_1 szerepel, azaz alkalmas α_i konstansokkal: $v^{(0)} = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n$, ahol $\alpha_1 \neq 0$.

Képezzük a $v^{(k)} = Av^{(k-1)} = A^k v^{(0)}$ $(k = 1, 2, \ldots)$ sorozatot. (Itt a zárójel nélküli felső index hatványkitevő; ez is magyarázza a módszer elnevezését.) A kiinduló feltevések miatt

$$\begin{aligned} v^{(1)} &= Av^{(0)} = A(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n) \\ &= \alpha_1 \lambda_1 x_1 + \alpha_2 \lambda_2 x_2 + \ldots + \alpha_n \lambda_n x_n, \end{aligned}$$

illetve

$$\begin{aligned} v^{(k)} = Av^{(k-1)} &= A(\alpha_1 \lambda_1^{k-1} x_1 + \alpha_2 \lambda_2^{k-1} x_2 + \ldots + \alpha_n \lambda_n^{k-1} x_n) \\ &= \alpha_1 \lambda_1^k x_1 + \alpha_2 \lambda_2^k x_2 + \ldots + \alpha_n \lambda_n^k x_n \\ &= \lambda_1^k \left(\alpha_1 x_1 + \alpha_2 \left(\tfrac{\lambda_2}{\lambda_1}\right)^k x_2 + \ldots + \alpha_n \left(\tfrac{\lambda_n}{\lambda_1}\right)^k x_n \right). \end{aligned}$$

Legyen $y \in \mathbb{R}^n$ tetszőleges olyan vektor, amelyre $y^T v^{(k)} \neq 0$. Ekkor

$$\frac{y^T A v^{(k)}}{y^T v^{(k)}} = \frac{y^T v^{(k+1)}}{y^T v^{(k)}} = \frac{\lambda_1^{k+1} \left(\alpha_1 y^T x_1 + \sum_{i=2}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^{k+1} y^T x_i \right)}{\lambda_1^k \left(\alpha_1 y^T x_1 + \sum_{i=2}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^{k} y^T x_i \right)} \to \lambda_1,$$

ha $k \to \infty$, mert () miatt $|\lambda_k/\lambda_1| \leq |\lambda_2/\lambda_1| < 1$ $(k \geq 2)$. Ugyanezért

$$\frac{v^{(k)}}{\lambda_1^k} \to \alpha_1 x_1.$$

A $v^{(k)}$ tehát egyre "párhuzamosabb" lesz x_1-gyel, azaz maga is egyre jobb közelítése egy λ_1-hez tartozó sajátvektornak. Gondot okoz viszont, hogy komponensei a λ_1 abszolút értékétől függően egyre nagyobb abszolút értékűek lehetnek, vagy gyorsan tarthatnak zérushoz. Mindkét eset numerikusan előbb utóbb kezelhetetlenné válik. Megoldást jelent viszont, ha időnként, leginkább minden lépésben osztjuk egy számmal (hisz a hossz változásával egy sajátvektor sajátvektor marad). Legkényelmesebb, ha normálunk, azaz a saját ∞ jelű normájával osztunk. Összefoglalva, az algoritmus a következő:

6.1.1 A HATVÁNYMÓDSZER ALGORITMUSA

Input $v^{(0)} \in \mathbb{R}^n, \varepsilon \quad (\varepsilon > 0)$

for $k = 1, 2, \ldots \quad z^{(k)} = Av^{(k-1)}$

$\gamma_k = y^T z^{(k)} / y^T v^{(k-1)}$, ($y^T \in \mathbb{R}^n$, lépésenként változhat, de $y^T v^{(k-1)} \neq 0$)

$v^{(k)} = z^{(k)} / \left\| z^{(k)} \right\|_\infty$

end

A fentiek alapján fennáll, hogy

$$v^{(k)} \to x_1, \quad \gamma_k \to \lambda_1.$$

A $v^{(k)} \to x_1$ konvergencián itt azt értjük, hogy $v^{(k)}$ hatásvonala tart x_1 hatásvonalához. A $v^{(k)}$ sorozat elemeinek komponensei váltakozó előjelűek, ha λ_1 negatív. Az y vektort általában egységvektornak választjuk úgy, hogy ha $\left| v_i^{(k-1)} \right| = \left\| v^{(k-1)} \right\|_\infty$, akkor legyen $y = e_i$. Ekkor tehát csupán két komponens hányodosát kell kiszámítani: $\gamma_k = z_i^{(k)} / v_i^{(k-1)}$; azt a kettőt, ahol a nevezőben szereplő $v^{(k-1)}$ legnagyobb abszolút értékeű komponense áll (amiről ráadásul azt is tudjuk, hogy 1 vagy -1).

Az eljárás a $|\lambda_2/\lambda_1|$ nagyságrendjétől függő konvergencia sebességgel rendelkezik. A módszer erősen érzékeny a $v^{(0)}$ kezdővektor megválasztására is.

Ha $\alpha_1 = 0$, akkor az eljárás nem konvergál a λ_1 domináns sajátértékhez. Bizonyos mátrixosztályok esetén igazolták, hogy véletlenül választott $v^{(0)}$ kezdővektorok mellett 1 valószínűséggel konvergál az eljárás. Komplex sajátértékek, illetve többszörös λ_1 esetén az eljárást módosítani kell. Az eljárás konvergenciáját gyorsítani lehet, ha az $A - \sigma I$ ún. eltolt mátrixra hajtjuk végre, ahol σ alkalmasan megválasztott szám. Az $A - \sigma I$ mátrix sajátértékei ui. (mint azt a tételben láttuk): $\lambda_1 - \sigma, \lambda_2 - \sigma, \ldots, \lambda_n - \sigma$. A konvergenciát befolyásoló $|\lambda_2 - \sigma| \,/\, |\lambda_1 - \sigma|$ pedig a σ ügyes megválasztásával kisebbé tehető mint $|\lambda_2/\lambda_1|$.

A hatványmódszert az

$$\frac{\|r_k\|_2}{\|v^{(k)}\|_2} = \frac{\left\|Av^{(k)} - \gamma_k v^{(k)}\right\|_2}{\|v^{(k)}\|_2} \leq \varepsilon \tag{11}$$

kilépési feltétellel szokás leállítani, ahol r_k a γ_k és $v^{(k)}$ ()-hez tartozó reziduális hibája.

6.1.1. PÉLDA. Keressük meg hatványmódszerrel az előző példában szereplő mátrix domináns sajátértékét és a hozzátartozó sajátvektort.

Megoldás Jó kezdővektort nem mindig sikerül találni, különösen ilyenkor, amikor nincs n darab független sajátvektor. Induljunk ki a $v^{(0)} = [2, 2, 2]^T$-ből. A számítási erdményeket táblázatba foglaltuk.

z^T	γ	v^T	$\|r\|_\infty$
$[12, 2, -20]$	6	$[0.6000, 0.1000, -1]$	6.9
$[-0.1000, 0.1000, 0.9000]$	-0.9000	$[-0.1111, 0.1111, 1]$	2.6556
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[-0.6754, 0.0088, 1.7982]$	-1.7982	$[-0.3756, 0.0049, 1]$	0.3286
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[0.8006, 0.0000, -2.0009]$	-2.0009	$[0.4001, 0.0000, -1]$	0.0014

A $k = 1, 2, \ldots, 6, \ldots, 15$ lépéseket tüntettük fel. A konvergencia meglehetősen lassúnak bizonyult, mint láthatjuk. Most a $\sigma = 0.8$ eltolással, az $A - \sigma I$-re végrehajtva az algoritmust, már $k = 4$-nél elértük azt, amit az előbb a 15-ik lépésben. $k = 15$-nél pedig megkaptuk az eredményeket a szabványos duplapontos aritmetikában elérhető maximális, azaz 16 decimális jegy pontossággal.

A hatványmódszert, amely igen előnyös lehet nagyméretű ritka mátrixok esetén, leginkább a legnagyobb, ill. a legkisebb abszolút értékű sajátértékek meghatározására használjuk. Ez utóbbi a következőképpen történhet. Az A^{-1} sajátértékei: $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}$. Ezek közül a legnagyobb abszolút értékű sajátérték $\frac{1}{\lambda_n}$ lesz. Itt sem szokás invertálni, hanem a $z^{(k)} = A^{-1}v^{(k-1)}$ helyett az $Az^{(k)} = v^{(k-1)}$ egyenletrendszert megoldani; legkényelmesebben LU-módszerrel.

A ()-ben szereplő r_k reziduális hiba (vagy a relatív hibája) kicsiségétől azt reméljük, hogy nagyságrendben valóban jól adja vissza a k-adik közelítések hibáját.

6.1.2. MEGJEGYZÉS. Sajnos ez nem mindig így van. Tekintsük az

$$A(\varepsilon) = \begin{bmatrix} 1 & 1 \\ \varepsilon & 1 \end{bmatrix}$$

mátrixot, ahol $\varepsilon \approx 0$ kicsi. Az $A(\varepsilon)$ mátrix sajátértékei $1 \pm \sqrt{\varepsilon}$. Legyen $\gamma_k = 1$ és $x_k = [1, 0]^T$. Ekkor a közelítő sajátérték tényleges hibája $\pm\sqrt{\varepsilon}$ de

$$\|r\|_2 = \left\| \begin{bmatrix} 0 \\ \varepsilon \end{bmatrix} \right\|_2 = \varepsilon.$$

Ha most például $\varepsilon = 10^{-10}$, akkor a reziduális hiba öt nagyságrenddel pontosabbat mutat, mint a tényleges hiba. Óvatosan kell tehát a () eredményét kezelni.

6.1.3. MEGJEGYZÉS. Végül megjegyezzük, hogy rangszámcsökkentő eljárásokkal és egyéb módosításokkal a Mieses eljárás alkalmassá tehető az összes sajátérték-sajátvektor meghatározására is.

7 NUMERIKUS DIFFERENCIÁLÁS (DERIVÁLÁS)

Alapfeladat: A numerikus deriválás alapfeladata az analitikusan ismeretlen vagy

nehezen számolható, esetleg csak diszkrét pontokban ismert $f : D(\subseteq \mathbb{R}) \to \mathbb{R}$ függvény deriváltjának kiszámítása egy vagy több adott pontban.

A numerikus deriválást akkor alkalmazzák, ha az $a < x_0 < x_1 < \ldots < x_n < b$ alappontokban adottak az $y = f(x)$ függvény értékei

$$y_k = f_k = f(x_k), \qquad k = 0, 1, 2, \ldots, n.$$

Ekvidisztáns alappontok esetén:

$$x_k = x_0 + k \cdot h, \qquad k = 0, 1, 2, \ldots, n.$$

A numerikus deriválást úgy lehet végrehajtani, hogy az adott függvényt interpolációs polinommal közelítjük:

$$f(x) = p(x) + R(x), \qquad \text{azaz } f(x) \approx p(x)$$

és az $f(x)$ függvény deriváltját a $p(x)$ polinom deriváltjával közelítjük.

$$f^{(k)}(x) = p^{(k)}(x) + R^{(k)}(x), \text{ azaz } f^{(k)}(x) \approx p^{(k)}(x)$$

Ha a maradék-tag deriváltjainak értéke $\left|R^{(k)}(x)\right| < \varepsilon$, akkor az $f(x)$ függvény deriváltjai az interpolációs polinom alapján kiszámíthatóak:

$$f'(x) \approx p'(x), \ldots, f^{(n)}(x) \approx p^{(n)}(x)$$

7.1 NUMERIKUS DIFFERENCIÁLÁS DIFFERENCIA HÁNYADOSOKKAL

Ha $f \in C^2[a, b]$ akkor megszerkeszthetjük a másodfokú Taylor-polinomot:

$$f(x + h) = f(x) + hf'(x) + h^2 f''(\xi)/2, \text{ ahol } \xi \in [x, x + h].$$

Innen

$$f(x + h) - f(x) = h \cdot f'(x) + h^2 f''(\xi)/2,$$

amiből adódik, hogy

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

A képlet hibájának nagyságrendje $O(h)$.

A derivált definíciója:

$$f'(x) = \lim_{x_0 \to x} \frac{f(x_0) - f(x)}{x_0 - x}.$$

Legyen $x_0 = x + h$. Ekkor $x_0 - x = h$, és $x \to x_0$ miatt $h \to 0$. Így az előző definíció

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$$

alakban írható.

Ha $f \in C^3[a,b]$ akkor a harmadfokú Taylor-polinom segítségével pontosabb közelítést kaphatunk.

$$f(x+h) = f(x) + hf'(x) + h^2 f''(x)/2 + h^3 f'''(\xi_1)/6, \text{ ahol } \xi_1 \in [x, x+h].$$

Hasonlóan:

$$f(x-h) = f(x) - hf'(x) + h^2 f''(x)/2 - h^3 f'''(\xi_2)/6, \text{ ahol } \xi_2 \in [x-h, x].$$

Ha az első képletből kivonjuk a másodikat, akkor a következőt kapjuk:

$$f(x+h) - f(x-h) = 2hf'(x) + h^3/6[f'''(\xi_1) + f'''(\xi_2)],$$

amelyből adódik, hogy

$$f(x+h) - f(x-h) = 2hf'(x) - h^3 f'''(\xi_3)/3, \text{ ahol } \xi_3 \in [x-h, x+h].$$

Innen kapjuk, hogy

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

Ebben az esetben a hiba $O(h^2)$.

A második derivált leggyakrabban alkalmazott képletei:

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

A második képletet a centrális differencia formulának nevezzük. A képletek hibájának nagyságrendje $O(h^2)$.

7.2 NUMERIKUS DIFFERENCIÁLÁS LAGRANGE INTERPOLÁCIÓVAL

Amennyiben $f(x)$-et az alappontokon átmenő

$$p(x) = \sum_{i=1}^{n} y_i \cdot l_i(x)$$

Lagrange-féle interpolációs polinommal közelítjük, akkor az $f(x)$ függény x-pontbeli j-edik deriváltjának közelítését az

$$f^{(j)}(x) \approx p^{(j)}(x) = \sum_{i=1}^{n} y_i \cdot l_i^{(j)}(x) \tag{12}$$

összefüggés adja meg.

Hibakorlátot is adhatunk amennyiben f elég sokszor folytonosan differenciálható $[a, b]$-n.

$$\left|f^{(j)}(x) - p^{(j)}(x)\right| \leq \sum_{i=0}^{j} \frac{j!}{(j-i)!(n+i)!} \max_{x\in[a,b]} \left|f^{(n+i)}(x)\right| \left|\omega^{(j-i)}(x)\right|, \tag{13}$$

ahol $x, x_1, x_n \in [a, b]$ és $\omega(x) = (x - x_1)(x - x_2)\dots(x - x_n)$.

Az első derivált közelítő értéke $n = 2$ és $x_1 = x$, $x_2 = x + h$ esetén

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}.$$

Ennek a hibája $O(h)$ nagyságrendű, ha $f \in C^2[a, b]$.

Ha pedig $n = 3$ és $x_1 = x - h$, $x_2 = x$, $x_3 = x + h$, akkor

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

Ezen közelítés hibájának nagyságrendje $O\left(h^2\right)$, ha $f \in C^3[a, b]$.

Az

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

képlet segítségével másodrendű derivált is származtatható, mégpedig szintén $O\left(h^2\right)$ hibával:

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

7.3 NUMERIKUS DIFFERENCIÁLÁS NEWTON-FÉLE INTERPOLÁCIÓVAL

$$N_n(x) = f_0 + \frac{x - x_0}{h}\Delta y_0 + \frac{(x - x_0)(x - x_1)}{2!h^2}\Delta^2 y_0 + \ldots + \frac{(x - x_0)\dots(x - x_n)}{n!h^n}\Delta^n y_0.$$

Ekvidisztáns pontok esetén, bevezetve a $t = \frac{x - x_0}{h}$ jelölést kapjuk, hogy $x - x_n = (t - n)h$ és

$$N_n(x) = f_0 + t\Delta y_0 + \frac{t(t-1)}{2!}\Delta^2 y_0 + \ldots + \frac{t(t-1)\dots(t-(n-1))}{n!}\Delta^n y_0.$$

Az x_0, x_1, x_2, x_3, x_4 alappontok alapján az első Newton-féle interpolációs polinom differenciálásával a következő képletet kapjuk:

$$f'(x) \approx [(\Delta y_0) + (2t-1)(\Delta^2 y_0)/2 + (3t^2 - 6t + 2)(\Delta^3 y_0)/6 + (2t^3 - 9t^2 + 11t - 3)(\Delta^4 y_0)/12]/h$$

Mivel az x_0 kezdőpontot tetszőlegesen lehet kiválasztani, ezért feltehetjük, hogy az x_0 az a pont, amelyben a derivált értékét akarjuk kiszámítani.

Figyelembe vesszük, hogy ebben a pontban $t = 0$, akkor az első Newton-féle interpolációs polinom alapján:

$$f'(x_0) \approx [(\Delta y_0) - (\Delta^2 y_0)/2 + (\Delta^3 y_0)/3 - (\Delta^4 y_0)/4 + \ldots + (-1)^{n-1}(\Delta^n y_0)/n]/h$$

Ezeket a képleteket átalakíthatjuk úgy, hogy a differenciák helyett a képletek csak függvény értékeit tartalmazzák:

$$f'(x_0) \approx [(\Delta y_0) - (\Delta^2 y_0)/2]/h = (-3y_0 + 4y_1 - y_2)/2h.$$

$$f'(x_0) \approx [(\Delta y_0) - (\Delta^2 y_0)/2 + (\Delta^3 y_0)/3]/h = (-11y_0 + 18y_1 - 9y_2 + 2y_3)/6h.$$

$$f'(x_0) \approx [(\Delta y_0) - (\Delta^2 y_0)/2 + (\Delta^3 y_0)/3 - (\Delta^4 y_0)/4]/h = (-25y_0 + 48y_1 - 36y_2 + 16y_3 - 3y_4)/12h.$$

A második derivált közelítésére a Newton-féle interpolációs polinom segítségével a következő képletet kapjuk:

$$f''(x_0) \approx [(\Delta^2 y_0) - (\Delta^3 y_0) + (\Delta^4 y_0) \cdot 11/12 - (\Delta^5 y_0) \cdot 5/6 + \ldots]/h^2$$

7.3.1. PÉLDA. *Példa a Newton-féle interpolációs képletek alkalmazására:*
Számítsuk ki az $y = \mathrm{sh}(2x)$ függvény $y'(x)$ és $y''(x)$ deriváltjait az $x = 0$ pontban:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0,00	0,00000	0,10017	0,00100	0,00101	0,00003
0,05	0,10017	0,10117	0,00201	0,00104	0,00003
0,10	0,20134	0,10318	0,00305	0,00107	
0,15	0,30452	0,10623	0,00412		
0,20	0,41075	0,11035			
0,25	0,52110				

Megoldás:

$$\begin{aligned} y'(0) &\approx [(\Delta y_0) - (\Delta^2 y_0)/2 + (\Delta^3 y_0)/3 - (\Delta^4 y_0)/4]/h \\ &= 20 \cdot (0.10017 - 0.00050 + 0.00034 - 0.00001) = 2.0000. \\ y''(0) &\approx [(\Delta^2 y_0) - (\Delta^3 y_0) + (\Delta^4 y_0) \cdot 11/12]/h^2 \\ &= 400 \cdot (0.00100 - 0.00101 + 0.00003) = 0.008 \end{aligned}$$

Az $y = \text{sh}(2x)$ függvény analitikus differenciálásával kapjuk, hogy $y'(x) = 2\,\text{ch}(2x)$ és $y''(x) = 4\,\text{sh}(2x)$. Így a pontos derivált értékek: $y'(0) = 2$ és $y''(0) = 0$.

7.4 NUMERIKUS DIFFERENCIÁLÁS SPLINE INTERPOLÁCIÓVAL

Spline közelítést is használhatunk deriváltak numerikus közelítésére. Elsőrendű deriváltak előnyösen közelíthetők a harmadfokú természetes Spline-nal.

A képlet levezetése könnyű, hiszen szakaszonként egy-egy harmadfokú polinomot kell deriválni.) A közelítés hibájára $f \in C^2\,[a, b]$ esetén fennáll, hogy

$$|f'(x) - S'(x)| \leq K_1 \left(\max_{1 \leq i \leq n-1} h_i \right),$$

ahol $K_1 > 0$ konstans. A közelítés hibája tehát a legnagyobb részintervallum hosszával arányos. A Spline közelítéseknek van egy simító jellege is, amely mérsékeli a kerekítési, ill. adathibák negatív hatását. A Spline használata akkor előnyös, ha a derivált sok pontban szükséges.

8 NUMERIKUS INTEGRÁLÁS

Newton-Leibnitz formula: Ha f Riemann-integrálható $[a, b]$-n és itt létezik F primitív függvénye, akkor

$$I = \int_a^b f(x)dx = F(b) - F(a).$$

Ezt a képletet csak akkor lehet alkalmazni, ha megtudjuk határozni az F primitív függvényt, különben kénytelenek vagyunk numerikus integrálást alkalmazni.

A numerikus integrálásnak úgy lehet végrehajtani, hogy az f függvényt egy p interpolációs polinommal közelítjük:

$$f(x) \approx p(x)$$

és ez alapján a határozott integrál értékét általános formában így közelítjük:

$$I = \int_a^b f(x)w(x)dx \approx \int_a^b p(x)w(x)dx,$$

ahol $w \geq 0$ egy tetszőleges súlyfüggvény.

8.1 LAGRANGRE INTERPOLÁCIÓ ALKALMAZÁSA

Legyenek adottak az $x_1, x_2, \ldots, x_n$ alappontok és ezekben az $y_1, y_2, \ldots, y_n$ függvényértékek. Ha az interpolációs polinomként a Lagrange féle interpolációs polinomot választjuk:

$$p(x) = \sum_{j=1}^{n} y_j \cdot l_j(x),$$

akkor

$$I = \int_a^b f(x)w(x)dx \approx \int_a^b p(x)w(x)dx = \int_a^b \left(\sum_{j=1}^{n} y_j \cdot l_j(x) \right) w(x)dx$$
$$= \sum_{j=1}^{n} y_j \cdot \underbrace{\left(\int_a^b l_j(x)w(x)dx \right)}_{H_j} = \sum_{j=1}^{n} y_j \cdot H_j$$

Hibabecslés: A numerikus integrálás hibabecslése $w \equiv 1$ esetén a következő képlet alapján állapítható meg:

$$R_n(f) = \int_a^b f(x)dx - \int_a^b p(x)dx = \frac{f^{(n+1)}(\xi)}{(n+1)!} \int_a^b \omega(x)dx$$

ahol $\xi \in [a, b]$ és $\omega(x) = (x - x_1)(x - x_2) \ldots (x - x_n)$. Innen:

$$|R_n(f)| \leq \frac{M_{n+1}(b-a)^{n+1}}{(n+1)!}$$

ahol

$$M_{n+1} = \max_{x \in [a,b]} \left| f^{(n+1)}(x) \right|$$

8.2 NEWTON-COTES FORMULÁK

Legyen $w \equiv 1$. A Newton-Cotes formulákat úgy kaphatjuk, hogy az f függvényt a Lagrange-féle interpolációs polinommal közelítjük, úgy hogy az alappontok ekvidisztánsak legyenek:

$$x_k = a + k \cdot h, \qquad k = 0, 1, 2, \ldots, n$$

ahol a lépésköz

$$h = (b-a)/n$$

A Newton-Cotes formulákat két változatban alkalmazhatjuk.

Zárt Newton-Cotes formula:

A zárt Newton-Cotes formula az alappontokat (az a és b pontokat) is tartalmazza. Ekkor az integrált az

$$\int_a^b f(x)dx = \sum_{k=0}^{n} A_k^{(n)} f(a+k\cdot h) = \sum_{k=0}^{n} A_k^{(n)} y_k$$

alakban írhatjuk fel. Az $A_k^{(n)}$ együtthatókat a $B_k^{(n)}$ együtthatók alapján lehet kiszámítani:

$$A_k^{(n)} = (b-a)B_k^{(n)},$$

ahol

$$B_k^{(n)} = \int_a^b l_k(x)dx.$$

Mivel ekvidisztáns beosztásunk van, ezért bevezetve a $t = (x-x_0)/h$ jelölést a Lagrange együtthatókra igaz, hogy

$$l_k(x) = (-1)^{n-k}\binom{n}{k}\frac{1}{t-k}\cdot\frac{t(t-1)\cdots(t-n)}{n!}$$

Ezekre a $B_k^{(n+1)} = \int_a^b l_k(x)dx$ együtthatókra igaz, hogy nem függnek az $[a,b]$ intervallumtól,

$$B_k^{(n)} = B_{n-k}^{(n)}$$

és

$$\sum_{k=0}^{n} B_k^{(n)} = 1.$$

A következő táblázat tartalmazza a $B_k^{(n)}$ értékeket, amelyek a zárt Newton-Cotes formulákhoz tartoznak ($n \le 4$).

$$\begin{aligned}
n=1: &\quad B_0^{(1)} = 1/2, && B_1^{(1)} = 1/2, \\
n=2: &\quad B_0^{(2)} = 1/6, && B_1^{(2)} = 4/6, && B_2^{(2)} = 1/6, \\
n=3: &\quad B_0^{(3)} = 1/8, && B_1^{(3)} = 3/8, && B_2^{(3)} = 3/8, && B_3^{(3)} = 1/8 \\
n=4: &\quad B_0^{(4)} = 7/90, && B_1^{(4)} = 32/90, && B_2^{(4)} = 12/90, && B_3^{(4)} = 32/90 \\
&\quad B_4^{(4)} = 7/90
\end{aligned}$$

Nyílt Newton-Cotes formula:

A nyílt Newton-Cotes formula az alappontokat (az a és b pontokat) nem tartalmazza. Ekkor az integrált az

$$\int_a^b f(x)dx = \sum_{k=1}^{n-1} A_k^{(n)} f(a + k \cdot h) = \sum_{k=1}^{n-1} A_k^{(n)} y_k$$

alakban írhatjuk fel. Az $A_k^{(n)}$ együtthatókat az előzőhöz hasonlóan a $B_k^{(n)}$ együtthatók alapján lehet kiszámítani.

A következő táblázat tartalmazza a $B_k^{(n)}$ értékeket, amelyek a nyílt Newton-Cotes formulákhoz tartoznak ($n \leq 5$).

$$\begin{aligned}
n = 3: &\quad B_1^{(3)} = 1/2, && B_2^{(3)} = 1/2, \\
n = 4: &\quad B_1^{(4)} = 2/3, && B_2^{(4)} = -1/3, && B_3^{(4)} = 2/3, \\
n = 5: &\quad B_1^{(5)} = 11/24, && B_2^{(5)} = 1/24, && B_3^{(5)} = 1/24, && B_4^{(5)} = 11/24
\end{aligned}$$

8.2.1. PÉLDA. Határozzuk meg az

$$I = \int_0^1 \frac{1}{1+x^2} dx$$

integrál értékét zárt Newton-Cotes formulával $n = 4$ esetén!

Megoldás:

$n = 4$ esetén az alappontok: $x_0 = 0, x_1 = 0.25, x_2 = 0.5, x_3 = 0.75, x_4 = 1$.

A függvényértékek: $y_0 = 1, y_1 = 16/17, y_2 = 4/5, y_3 = 16/25, y_4 = 1/2$.

A hozzátartozó együtthatók, $b-a = 1$ miatt: $A_0^{(4)} = 7/90, A_1^{(4)} = 32/90, A_2^{(4)} = 12/90, A_3^{(4)} = 32/90, A_4^{(4)} = 7/90$.

Így

$$\begin{aligned}
I = \int_0^1 \frac{1}{1+x^2} dx &= (7/90) \cdot 1 + (32/90) \cdot (16/17) \\
&+ (12/90) \cdot (4/5) + (32/90) \cdot (16/25) + (7/90) \cdot (1/2) \\
&= 6677/8500 = 0.7855294118.
\end{aligned}$$

A pontos érték:

$$I = \int_0^1 \frac{1}{1+x^2} dx = \pi/4 = 0.7853981635.$$

8.3 TÉGLALAP-FORMULÁK

A lépésköz $h = (b-a)/n$. **Első téglalap-formula:**

$$I^{(1)} = h \cdot \sum_{j=0}^{n-1} y_j$$

Második téglalap-formula:

$$I^{(2)} = h \cdot \sum_{j=1}^{n} y_j$$

Ha az f függvény monoton, akkor

$$\left| I^{(.)} - \int_a^b f(x)dx \right| \leq \left| I^{(1)} - I^{(2)} \right| = h \cdot |y_n - y_0| = h \cdot |f(b) - f(a)|$$

Ebből látszik, hogy a lépésköztől lineárisan függ a hiba, azaz $O(h)$.

Harmadik téglalap-formula:

$$I^{(3)} = h \cdot \sum_{j=1}^{n} y_{j-1/2},$$

ahol

$$y_{j-1/2} = f(x_{j-1/2}), \qquad x_{j-1/2} = (x_{j-1} + x_j)/2, \qquad j = 1, 2, \ldots, n$$

8.4 TRAPÉZ-FORMULÁK

8.4.1 EGYSZERŰ TRAPÉZ MÓDSZER (N=1):

Ekkor $a = x_0, b = x_1$.

$$I = \int_a^b f(x)dx = \frac{y_0 + y_1}{2} \cdot h + R, \qquad \int_a^b f(x)dx \approx \frac{y_0 + y_1}{2} \cdot h.$$

8.4.1. TÉTEL. *Ha f kétszer folytonosan differenciálható és $|f''| \leq M_2$ az $[a, b]$ intervallumon, akkor*

$$|R| \leq \frac{M_2 \cdot h^3}{12}.$$

B i z o n y í t á s. A Lagrange interpoláció hibáját felhasználva:

$$f(x) = p(x) + \frac{f''(\xi)}{2}(x - x_0)(x - x_1),$$

ahol $p(x)$ lineáris (egyenes) és $\xi \in [a, b]$. Innen

$$\begin{aligned} I = \int_a^b f(x)dx &= \int_a^b p(x)dx + \int_a^b \frac{f''(\xi)}{2}(x - x_0)(x - x_1)dx \\ &= \frac{y_0 + y_1}{2} \cdot h + \int_a^b \frac{f''(\xi)}{2}(x - x_0)(x - x_1)dx. \end{aligned}$$

Ebből adódik, hogy

$$R = \int_a^b f(x)dx - \frac{y_0 + y_1}{2} \cdot h = \int_a^b \frac{f''(\xi)}{2}(x - x_0)(x - x_1)dx.$$

$$\begin{aligned} |R| &= \left| (1/2) \cdot \int_a^b f''(\xi)(x - x_0)(x - x_1)dx \right| \\ &\leq (1/2) \cdot \int_a^b |f''(\xi)| \, |(x - x_0)(x - x_1)| \, dx \\ &\leq (M_2/2) \cdot \int_a^b |(x - x_0)(x - x_1)| \, dx, \end{aligned}$$

Mivel

$$\begin{aligned} \int_a^b |(x - x_0)(x - x_1)|dx &= \int_a^b \left(-x^2 + (x_0 + x_1)x - x_0x_1\right) dx \\ &= \left[-x^3/3 + (x_0 + x_1)x^2/2 - (x_0x_1)x\right]_a^b \\ &= -(b^3 - a^3)/3 + (a + b)(b^2 - a^2)/2 - ab(b - a) = h^3/6. \end{aligned}$$

Így

$$|R| \leq \frac{M_2}{2} \cdot \frac{h^3}{6} = \frac{M_2 \cdot h^3}{12}.$$

8.4.2 Összetett trapéz formula

Kihasználva a határozott integrál additív tulajdonságát, az $[a, b]$ intervallumot felbonthatjuk az

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b$$

pontokkal n részintervallumra, minden intervallumon kiszámoljuk az egyszerű trapéz formulával az integrált, és ezeket összegezzük. Így jutunk el az összetett trapézformulához:

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n-1} \frac{x_{i+1}-x_i}{2}\left[y_i + y_{i+1}\right].$$

Hibabecslés: Ha $f \in C^2[a,b]$ és $|f''| \leq M_2$, akkor

$$\left|\int_a^b f(x)dx - \sum_{i=0}^{n-1} \frac{x_{i+1}-x_i}{2}\left[f(x_i)+f(x_{i+1})\right]\right| \leq \frac{M_2}{12}\sum_{i=0}^{n-1}\left(x_{i+1}-x_i\right)^3.$$

Ha az alappontok ekvidisztánsak, azaz $x_k = a + k \cdot h$ $(k = 0, \ldots, n)$, akkor a képlet alakja egyszerűsödik:

$$\int_a^b f(x)dx \approx T_n(f) := \frac{h}{2}\left[y_0 + 2\sum_{i=1}^{n-1} y_i + y_n\right].$$

A képlet hibája ($nh = b - a$ miatt):

$$\left|\int_a^b f(x)dx - T_n(f)\right| \leq \frac{M_2(b-a)h^2}{12} = \frac{M_2(b-a)^3}{12n^2}.$$

A hiba h^2-tel arányos, vagyis n^2-tel fordítottan arányos. A gyakorlatban csak összetett formulát használunk.

8.5 AZ ÉRINTŐFORMULA

Az érintőformula egy nyílt Newton-Cotes formula, amelyre:

$$\int_a^b f(x)dx \approx (b-a)f\left(\frac{a+b}{2}\right).$$

Az érintőformula úgy is értelmezhető, hogy a függvényt az $[a,b]$ intervallum középpontjához húzott érintőegyenessel közelítjük, és az egyenes alatti területet vesszük. Ez mutatja, hogy legfeljebb elsőfokú polinomig pontos.

Tegyük fel, hogy $f \in C^2([a,b])$. Ekkor az érintő formulával

$$\int_a^b f(x)dx = (b-a)f\left(\frac{a+b}{2}\right) + \frac{(b-a)^3 \cdot f''(\xi)}{24}.$$

Így a kapott hibabecslés:

$$\left|\int_a^b f(x)dx - (b-a)f\left(\frac{a+b}{2}\right)\right| \leq \frac{(b-a)^3 M_2}{24}.$$

ahol $M_2 \geq |f''(x)|$ az $[a,b]$ intervallumon.

A gyakorlatban ezt a formulát nem az egész $[a,b]$ intervallumra alkalmazzuk, hanem azt n részre osztjuk ($h=(b-a)/n$), és az egyes részintervallumokban az érintőformulával integrálunk.

$$\int_a^b f(x)dx \approx \frac{(b-a)}{n}\sum_{i=1}^{n} f\left(a-\frac{h}{2}+ih\right).$$

8.6 A SIMPSON FORMULA

8.6.1 EGYSZERŰ SIMPSON FORMULA

Legyen most $x_1 = a, x_2 = \dfrac{a+b}{2}$ és $x_3 = b$. Alkalmazzuk a három pontra támaszkodó másodfokú Lagrange-féle interpolációs polinomot.

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right].$$

Vegyük észre, hogy az egyszerű Simpson formula egy zárt Newton-Cotes formula, amelyre $n=2$ és $B_0^{(2)}=1/6, B_1^{(2)}=4/6, B_2^{(2)}=1/6$.

Hibabecslés:

Az egyszerű Simpson formula hibáját $f \in C^4[a,b]$ esetén a Lagrange-féle interpolációs polinom hibáját felhasználva kapjuk. Ekkor

$$\left|\int_a^b f(x)dx - \frac{b-a}{6}\left[f(a)+4\left(\frac{a+b}{2}\right)+f(b)\right]\right| \leq M_4\frac{(b-a)^5}{90}$$

formulával becsülhetjük, ahol $M_4 \geq \left|f^{(4)}(x)\right|$ az $[a,b]$-n.

8.6.2 ÖSSZETETT SIMPSON FORMULA, 1. VÁLTOZAT

Tegyük fel, hogy n páros és az alappontok ekvidisztánsak, azaz $x_i = x_0 + i\cdot h$ (ahol $h=\frac{b-a}{n}$, és $i=0,\ldots,n$). Ekkor a képlet alakja

$$S_n(f) = \frac{2h}{6}\big[f(x_0)+4(f(x_1)+f(x_3)+\cdots+f(x_{n-1})) \\ +2(f(x_2)+f(x_4)+\cdots+f(x_{n-2}))+f(x_n)\big].$$

A képlet hibája:

$$\left|\int_a^b f(x)dx - S_n(f)\right| \leq \frac{M_4(b-a)}{90}h^4 = \frac{M_4(b-a)^5}{90n^4}.$$

8.6.3 ÖSSZETETT SIMPSON FORMULA, 2. VÁLTOZAT

Ha az $[a, b]$ intervallumot itt is felosztjuk az

$$a = x_0 < x_1 < \ldots < x_n = b$$

pontokkal n részintervallumra, akkor az összetett Simpson formula:

$$\int_a^b f(x)dx \approx S_n := \sum_{i=0}^{n-1} \frac{x_{i+1} - x_i}{6} \left[f(x_i) + 4f\left(\frac{x_i + x_{i+1}}{2}\right) + f(x_{i+1}) \right].$$

Ha az alappontok ekvidisztánsak, azaz $x_i = x_0 + i \cdot h$ (ahol $h = \frac{b-a}{n}$ és $i = 0, \ldots, n$), akkor a képlet alakja

$$S_n(f) = \frac{h}{6} \left[f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + 4\sum_{i=0}^{n-1} f\left(x_i + \frac{h}{2}\right) + f(x_n) \right].$$

A képlet hibája pedig

$$\left| \int_a^b f(x)dx - S_n(f) \right| \leq \frac{M_4(b-a)}{32 \cdot 90} h^4 = \frac{M_4(b-a)^5}{2880n^4}.$$

8.7 GAUSS-KVADRATÚRÁK

Az eddigi, interpolációból származtatott kvadratúra-formulák legfeljebb annyiad-fokú polinomra pontosak, ahányad fokú polinomból származtattuk őket. A Gauss-kvadratúrák abból az észrevételből származnak, hogy az alappontok speciális megválasztásával a kavadratúra-formula rendje növelhető.

Ehhez szükségünk lesz az ortogonális polinomokra. A $\{(p_k(x))\}$ egy ortogonális polinom rendszer, ha

$$< p_i, p_j >= \int_a^b p_i(x)p_j(x)w(x)dx = 0 \qquad (i \neq j)$$

8.7.1. TÉTEL. *Legyen $\{(p_k(x))\}$ egy ortogonális polinom rendszer. Ekkor bármely n-re a $p_{n+1}(x)$ polinom gyökei valósak, egyszeresek és az $[a, b]$ intervallumban vannak, ahol $[a, b]$ a skalárszorzat integrálási tartománya.*

Az $n + 1$-pontos Gauss-kvadratúrát úgy kapjuk, hogy a $p_{n+1}(x)$ ortogonális polinom gyök-helyein készítjük az interpolációból származtatott kvadratúra-formulát.

A séma a következő:

Legyen $f(x) = p(x) + R(x)$, ahol p a Lagrange interpolációs polinom, R a hiba. Ekkor

$$\int_a^b f(x)w(x)dx = \int_a^b p(x)w(x)dx + \int_a^b R(x)w(x)dx$$
$$= \sum_{i=0}^{n} a_i f(x_i) + R_n,$$

ahol $a_i = \int_a^b l_i(x)w(x)dx$.

8.7.2. TÉTEL. *Legyenek a $p_{n+1}(x)$ ortogonális polinom gyökei $x_0, x_1, \ldots, x_n$ és legyen $a_i = \int_a^b l_i(x)w(x)dx$ ahol l_i az i -edik Lagrange alappolinom a fenti alappontokon. Ekkor a Gauss-kvadratúra*

$$G_n(f) = \sum_{i=0}^{n} a_i f(x_i)$$

pontos minden legfeljebb $2n+1$-edfokú polinomra, azaz, ha f egy legfeljebb $2n+1$-edfokú polinom, akkor

$$G_n(f) = \int_a^b f(x)w(x)dx.$$

8.7.3. TÉTEL. *(A Gauss kvadratúra hibaformulája)*
Legyen $f \in C^{2n+2}[a,b]$ és $G_n(f) = \sum_{i=1}^{n} a_i f(x_i)$, ahol az alappontok a $p_{n+1}(x)$ ortogoális polinom gyökei. Ekkor

$$\int_a^b f(x)w(x)dx - G_n(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \cdot < p_{n+1}, p_{n+1} >,$$

ahol $p_{n+1}(x)$ egy 1-főegyütthatós ortogonális polinom.

8.7.4. MEGJEGYZÉS. Az a_i együtthatók pozitívak.
Az $f(x) = 1$ függvény integálásával kapjuk a $\mu_0, \mu_1 \ldots$ momentumokat:

$$\sum_{i=1}^{n} a_i = \int_a^b w(x)dx =: \mu_0$$

$$\mu_i := \int_a^b x^i w(x)dx.$$

Legendre-polinom:

$$P_n(x) = \frac{(2n)!}{2^n n!^2}\big[x^n - \frac{n(n-1)}{2(2n-1)}x^{n-2} + \frac{n(n-1)(n-2)(n-3)}{2(2n-1)4(2n-3)}x^{n-4} \pm \cdots\big]$$

$$[a,b] = [-1,1]$$

$$w(x) = 1$$

$$\mu_0 = 2$$

$$P_0(x) = 1$$

$$P_1(x) = x$$

$$P_2(x) = 3/2(x^2 - 1/3)$$

Csebisev-polinom: Az n-edfokú Csebisev polinom a következő összefüggéssel adható meg:

$$T_n(x) = \cos(n \arccos(x))$$

$$[a,b] = [-1,1]$$

$$w(x) = \frac{1}{\sqrt{1-x^2}}$$

$$\mu_0 = \pi$$

$$T_0(x) = 1$$

$$T_1(x) = x$$

$$T_2(x) = 2(x^2 - 1/2)$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

Laugerre-polinom:

$$L_n(x) = \frac{2n-1-x}{n}L_{n-1}(x) - \frac{n-1}{n}L_{n-2}(x) \qquad n > 1.$$

$$[a,b] = [0,\infty[$$

$$w(x) = e^{-x}$$

$$\mu_0 = 1$$

$$L_0(x) = 1$$

$$L_1(x) = -(x-1)$$

$$L_2(x) = x^2 - 4x + 2$$

Hermite-polinom:

$$H_n(x) = 2xH_{n-1}(x) - 2(n-1)H_{n-2}(x) \qquad n > 1.$$

$$[a, b] =]-\infty, \infty[$$

$$w(x) = 1$$

$$\mu_0 = e^{-x^2}$$

$$H_0(x) = 1$$

$$H_1(x) = x$$

$$H_2(x) = 4(x^2 - 1/2)$$

8.8 KVADRATURAFORMULÁK HIBÁINAK UTÓLAGOS BECSLÉSE

Elvégezzük a numerikus integrálást n és $2n$ részintervallum esetén. Ha fennáll, hogy

$$|T_n(f) - T_{2n}(f)| \leq \varepsilon,$$

akkor a $T_{2n}(f)$ közelítést ε pontosságúnak fogadjuk el. Ha

$$|T_n(f) - T_{2n}(f)| > \varepsilon,$$

akkor az n értékét tovább kell növelni. A hibabecslést az un. Runge-elv alapján lehet végrehajtani.

Hiba utólagos becslése trapéz-formula esetén:

Ha $f''(x)$ előjele állandó, akkor az összetett trapézformula hibájára fennáll, hogy

$$\left| \int_a^b f(x)dx - T_{2n}(f) \right| \leq |T_n(f) - T_{2n}(f)| .$$

Hiba utólagos becslése Simpson-formula esetén:

Ha $f^{(4)}(x)$ előjele állandó, akkor az összetett Simpson-formula hibakorlátja a következő:

$$\left| \int_a^b f(x)dx - S_{2n}(f) \right| \leq |S_n(f) - S_{2n}(f)| .$$

9 NEMLINEÁRIS EGYENLETEK

Az

$$f(x) = 0 \quad (\ f : \mathbb{R} \to \mathbb{R}),$$

alakú egyváltozós egyenletek, illetve az

$$F(x) = 0 \quad (\ F : \mathbb{R}^n \to \mathbb{R}^n,\ n \geq 2),$$

alakú többváltozós egyenletrendszerek közelítő megoldási módszereit vizsgáljuk. Az egyenlet (egyenletrendszer) pontos megoldását következetesen x^*-gal jelöljük; $x^* \in \mathbb{R}$ (illetve $x^* \in \mathbb{R}^n$).

Direkt módszerek, tehát amelyek véges sok lépés után legalább elméletileg adnák x^*-ot, itt még egyváltozós esetben sem igen konstruálhatók.

A gyakorlatban mindig elegendő tudni egy elméleti eredmény többé-kevésbé pontos közelítését, így az iterációs eljárások a legtöbb esetben szóba jöhetnek.

Egy-egy módszer csak speciális esetben alkalmazható. Különösen így van ez a nemlineáris egyenleteknél, egyenletrendszereknél. Nincs univerzális eljárás, ami minden egyenletnél megfelelő pontosságú közelítést garantálna és különösen nincs olyan, ami mindezt elég gyorsan is tenné.

A továbbiakban minden esetben feltételezzük, hogy az f függvény folytonos. A konvergenciát garantáló feltételek között azonban többnyire még szigorúbb, differenciálhatósági megkötések is szerepelnek.

Az $f\,(x) = 0$ ($f : \mathbb{R} \to \mathbb{R}$) alakú valós egyenletek megoldási módszerei egy, az x^* megoldáshoz konvergáló $\{x_i\}_{i=0}^{\infty}$ sorozatot képeznek.

9.1 INTERVALLUMFELEZŐ MÓDSZER

Tegyük fel, hogy $f : \mathbb{R} \to \mathbb{R}$ folytonos az $[a, b]$ intervallumon és a végpontokban ellentétes a függyvény előjele, azaz fennáll, hogy

$$f(a)f(b) < 0. \tag{14}$$

Ismert tétel, hogy zárt intervallumban folytonos függvénynek van az intervallumon legnagyobb és legkisebb értéke, továbbá a függvény e két szám közötti valamennyi értéket felvesz az intervallumban. Ebből azonnal következik, hogy az $f\,(x) = 0$ egyenletnek van legalább egy $x^* \in (a, b)$ gyöke.

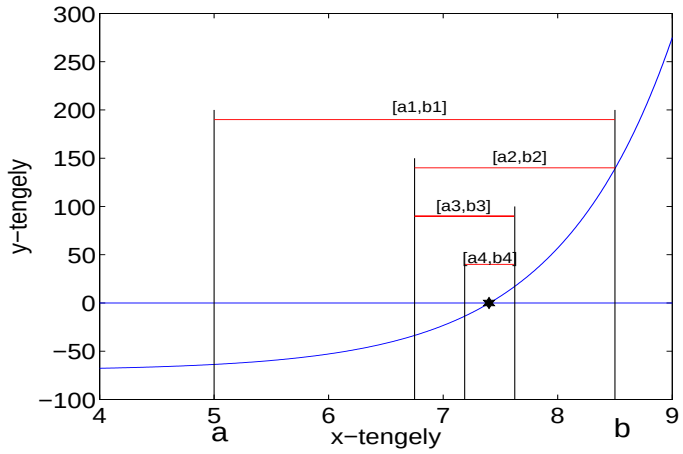
Az intervallumfelező eljárás (amit szoktunk röviden csak felező eljárásnak is nevezni) lényege a nevében benne van. Az intervallum középpontja két részre, két félintervallumra osztja az eredeti intervallumot. Megnézzük, hogy közülük melyikben teljesül () és ott folytatjuk. Kicsit pontosabban megfogalmazva: legyen $c = (a + b)\,/2$ és vizsgáljuk az $f\,(c)$ értékét. Ha annak előjele az a-beli előjellel

ellentétes, azaz $f(a)\,f(c) < 0$, akkor az $[a, c]$ intervallumban van gyök. Egyébként a $[c, b]$ intervallum tartalmaz gyököt. Az új intervallumot újra megfelezzük és így tovább.

Az egymásba skatulyázott

$$[a, b] = [a_1, b_1] \supset [a_2, b_2] \supset \ldots [a_k, b_k] \supset \ldots, \tag{15}$$

zárt intervallumok ráhúzódnak az egyenlet egy x^* gyökére. Más formában is megadva: $[a_1, b_1] = [a, b]$, $c_i = (a_i + b_i)/2$, $[a_{i+1}, b_{i+1}] = \begin{cases} [a_i, c_i], & \text{ha } f(a_i)f(c_i) < 0 \\ [c_i, b_i], & \text{egyébként} \end{cases}$, $(i = 1$



Az is nyilvánvaló, hogy az i-edik intervallum hossza: $(b_i - a_i) = (b_{i-1} - a_{i-1})/2 = (b - a)/2^{i-1}$ szigorúan monoton csökkenő mértani sorozat és az x^* gyököt az $[a_i, b_i]$ intervallum tetszőleges y pontjával közelíthetjük. Az y közelítés hibája legfeljebb a két végponttól mért nagyobbik távolság lehet, azaz fennáll, hogy

$$|x^* - y| \leq \max\{y - a_i, b_i - y\}. \tag{16}$$

A $\max\{y - a_i, b_i - y\}$ korlát akkor a legkisebb, ha $y = \frac{a_i + b_i}{2}$. Ezért az x^* gyök i-edik közelítéseként általában az $x_i = (a_i + b_i)/2$ felezőpontot használjuk. Az intervallumok hosszára tett megállapításból így egzakt hibabecslés is adódik:

$$|x^* - x_i| \leq \frac{b_i - a_i}{2} = \frac{b - a}{2^i} \quad (i = 1, 2, \ldots). \tag{17}$$

Az algoritmust tehát akkor állítjuk le, ha a közelítés hibája kisebb, mint egy előre megadott $\varepsilon > 0$ hibakorlát. Ezért az intervallumfelező eljárás gyakorlati formája a következő.

Az intervallumfelező eljárás algoritmusa:
input $[a,b]$, $\varepsilon > 0$.
$fa = f(a)$
while $b - a > 2\varepsilon$
  $x = (a+b)/2$
  if $fa * f(x) < 0$
    $b = x$
  else
    $a = x$
  end
end
$x = (a+b)/2$

9.1.1. TÉTEL. *Ha $f \in C[a,b]$ és $f(a) f(b) < 0$, akkor az az intervallumfelező eljárás során kapott $\{x_i\}$ sorozat konvergál az f valamely $[a,b]$-beli gyökéhez és érvényes az () hibabecslés.*

Ha az $[a,b]$ intervallum több gyököt is tartalmaz, az $\{x_i\}$ sorozat közülük egyikhez konvergál.

9.1.2. MEGJEGYZÉS. Az intervallumfelező módszer legnagyobb előnye, hogy nagyon enyhe (mondhatni, hogy csak a gyakorlatban szinte kötelezően elvárt) feltételek mellett biztosítja a megoldáshoz való konvergenciát. Előnye még az egyszerűsége is. Ugyanakkor hátránya a lassú konvergencia. A hibát a () szerint egy konvergens mértani sorozattal tudjuk becsülni. Az x_i hibakorlátja x_{i-1} hibájának a fele, így a konvergenciája lineáris vagy másképpen elsőrendű.

9.1.3. PÉLDA. Oldjuk meg intervallumfelező eljárással az $f(x) = 4 - 4x^2 - e^x = 0$ egyenletet az $[a,b] = [0,1]$ intervallumon $\delta x_i = \varepsilon = 10^{-6}$ hibakorláttal.

Megoldás. Az f függvény nyilvánvalóan folytonos, továbbá $f(a) = fa = f(0) = 3$ és $f(b) = f(1) = -e$ ellentétes előjelű, ezért alkalmazható a felező módszer. A $b - a = 1$ hosszú intervallumból indulva () alapján az $1/2^i \leq 10^{-6}$ egyenlőtlenségből $i \geq -\log\varepsilon/\log 2 \approx 19.93$, azaz $i = 20$ lépés szükséges. A lépéseket táblázatba foglaltuk, bekeretezve a végeredményt:

(18)

i	$[a,b]$	x	δx	$f(x) \approx$	
1	$[0,1]$	0.5	0.5	1.35	$\to a = x$
2	$[0.5,1]$	0.75	0.25	-0.4	$\to b = x$
3	$[0.5,0.75]$	0.625	0.125	0.57	$\to a = x$
$\vdots$		$\vdots$	$\vdots$		
20		$\boxed{0.703439}$	0.95×10^{-6}		

9.2 A FIXPONT ITERÁCIÓS ELJÁRÁS

A módszert az $f(x) = x - g(x) = 0$ alakú vagy ilyen alakra hozott egyenletek esetén alkalmazzuk. Az $f(x) = 0$ egyenlet ekvivalens az

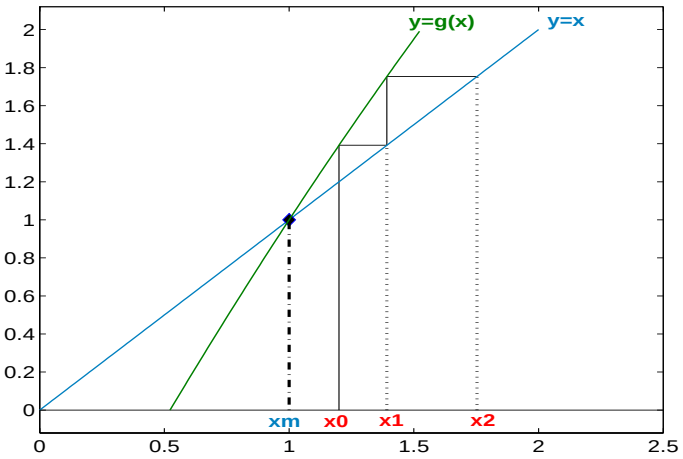
$$x = g(x) \tag{19}$$

egyenlettel. A módszer elnevezésének az az alapja, hogy a megoldás egyben az $x \to g(x)$ leképezés fixpontja: a leképezés során helyben maradó pont.

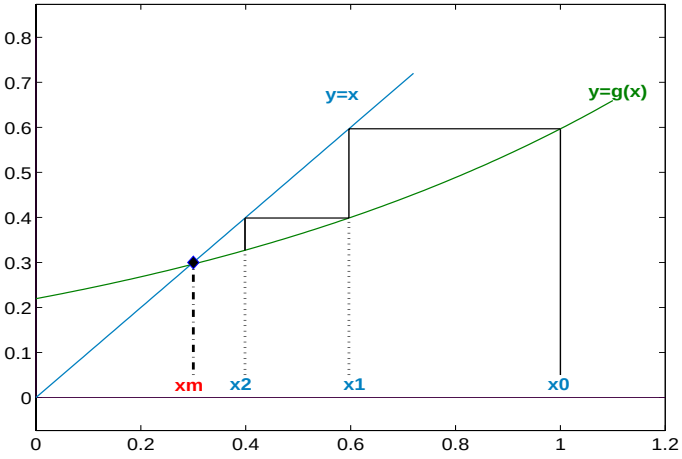
Választunk tehát egy x_0 kezdeti értéket, majd képezzük az

$$x_1 = g(x_0),\ x_2 = g(x_1), \ldots, x_{k+1} = g(x_k), \ldots \tag{20}$$

sorozatot. Azt reméljük, hogy a sorozat az egyenlet x^* megoldásához (a leképezés fixpontjához) konvergál. További lényeges kérdés, hogy az iterációt az i-edik lépésben megállítva, legfeljebb mekkora távolságra vagyunk a megoldástól. A következő két ábrán geometriai szemléltetését adjuk az eljárásnak, ami önmagában is igazolja, hogy a fixpont akár egyértelmű létezése sem ad garanciát a konvergenciára.



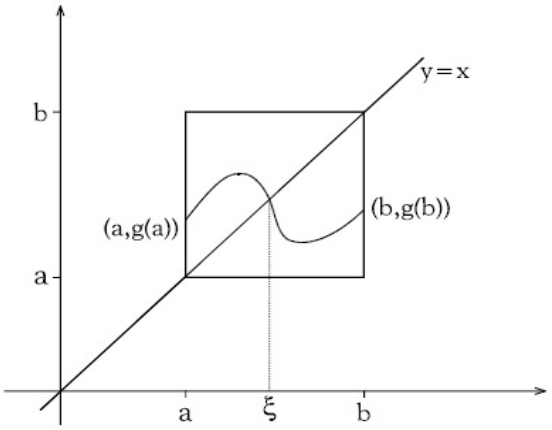
A fixpont iterációs eljárás vonzó gyök:



9.2.1. TÉTEL. *Ha* $g \in C[a,b]$ *és* $a \leq g(x) \leq b$ *minden* $x \in [a,b]$ *esetén, akkor a* $g(x)$ *függvénynek az* $[a,b]$ *intervallumon van fixpontja.*

Bizonyítás. Feltehetjük, hogy $g(a) > a$ és $g(b) < b$. (Egyenlőség esetén maga az a, illetve b fixpont lenne.) Legyen $h(x) = g(x) - x$. Ekkor $h(x)$ folytonos $[a,b]$-n, $h(a) > 0$ és $h(b) < 0$. Ezért a $h(x)$ függvénynek van $\xi \in (a,b)$ gyöke, azaz $h(\xi) = g(\xi) - \xi = 0$. ■

A fixpont iterációs eljárás vonzó gyök:



9.2.2. DEFINÍCIÓ. *A* $g \in C[a,b]$ *függvény kontrakció az* $[a,b]$ *intervallumon, ha létezik* $0 \leq q < 1$ *úgy, hogy*

$$|g(x) - g(y)| \leq q|x - y|, \quad x, y \in [a,b]. \tag{21}$$

9.2.3. PÉLDA. A

$g(x) = x^2$ függvény kontrakció a $\left[0, \frac{1}{4}\right]$ intervallumon, de a $[0,1]$ intervallumon már nem az.

$$\left|x^2 - y^2\right| = \underbrace{|x+y|}_{\leq 1/2} |x-y| \leq \frac{1}{2}|x-y|, \quad x, y \in \left[0, \frac{1}{4}\right].$$

legyen például $x, y \in \left[\frac{3}{4}, 1\right]$, de $x \neq y$ Két ílymódon választott szám ellenpédát ad, ugyanis

$$\left|x^2 - y^2\right| = \underbrace{|x+y|}_{\geq 3/2} |x-y| \geq \frac{3}{2}|x-y| > |x-y| \quad (x \neq y)$$

9.2.4. TÉTEL. Banach-féle fixponttétel
Ha $g \in C[a,b]$, $a \leq g(x) \leq b$ $(x \in [a,b])$ és $g(x)$ kontrakció $[a,b]$-n, akkor pontosan egy fixpont létezik $[a,b]$-ben.

9.2.5. TÉTEL. *Legyen $g \in C[a,b]$ Ha teljesülnek a következők:*
(i) $a \leq g(x) \leq b$ minden $x \in [a,b]$ esetén,
(ii) az $x \to g(x)$ kontrakció $[a,b]$-n q kontrakciós tényezővel,
akkor az $()$ egyenletnek pontosan egy megoldása van $[a,b]$-n és az $()$ sorozat minden $x_0 \in [a,b]$ esetén az egyenlet egyetlen x^* megoldásához konvergál. Érvényes továbbá az alábbi hibabecslés:

$$|x^* - x_k| \leq \frac{q}{1-q}|x_k - x_{k-1}| \leq \frac{q^k}{1-q}|x_1 - x_0|. \tag{22}$$

9.3 FIXPONT ITERÁCIÓ

9.3.1. TÉTEL. *Ha a g függvényre teljesülnek a következők:*
(i) $g \in C^1[a,b]$ (azaz folytonosan differenciálható $[a,b]$-n),
(ii) $\max_{x\in[a,b]} |g'(x)| = q < 1$,
akkor az $x \to g(x)$ kontrakció $[a,b]$-n q kontrakciós tényezővel.

Bizonyítás. Ha $x, y \in [a,b]$, akkor a differenciálszámítás Lagrange-féle középértéktétele alapján létezik $\xi \in [a,b]$, hogy

$$|g(x) - g(y)| = |g'(\xi)(x-y)| = |g'(\xi)|\,|x-y| \leq q|x-y|.$$

A fixpont iterációs eljárás algoritmusa:
Input x_0, ε $(\varepsilon > 0)$.

$k = 1, \quad x_1 = g(x_0)$
while kilépési feltétel=hamis
$x_{k+1} = g(x_k)$
$k = k + 1$
end

Az algoritmus költsége a felező eljáráshoz hasonló. Lépésenként itt is egy függvénykiértékelést kell végezni és ez határozza meg a költséget. Tárolni nem kell a sorozat minden elemét, elegendő mindig csak az utolsó kettőt. A kilépési feltételhez rendszerint ezekre szükség van, ezért is számítjuk ki még a ciklus előtt az x_1-et.

Kilépési feltételként a következőket szokás használni:

(A) $\frac{q}{1-q}|x_k - x_{k-1}| \leq \varepsilon$ (egzakt hibabecslés)
(B) $|x_k - x_{k-1}| \leq \varepsilon$
(C) $|x_k - g(x_k)| \leq \varepsilon$

Ha nem alkalmazhatjuk az egzakt kilépési feltételt, akkor nagyon körültekintően kell eljárni, mert a (B) és (C) egyidejű teljesülése sem garantálja, hogy valóban a gyök ε környezetében járunk, sőt, még a konvergenciát sem.

9.3.2. PÉLDA. Oldjuk meg az $f(x) = 4 - 4x^2 - e^x = 0$ egyenletet a $[0, 1]$ intervallumon, $\delta x_i = \varepsilon = 10^{-6}$ hibakorláttal, fixpont iterációs módszert alkalmazva. A számítások előtt igazoljuk a konvergenciát is.

Megoldás: Itt kísérletezhetünk az x^2 egyik oldalra rendezésével és utána négyzetgyökvonással, hogy az alkalmas $x = g(x)$ alakhoz jussunk. Kapjuk tehát az

$$x = g(x) = \frac{1}{2}\sqrt{4 - e^x}$$

egyenletet. (A négyzetgyökvonást pozitív előjellel végezzük mert pozitív gyököt kesesünk.)

A kapott g függvény folytonos, de az is azonnal látszik, hogy szigorúan monoton csökken a $[0, 1]$-en. (Ezt a derivált negatív volta mutatja.) Tehát a minimumát az intervallum végpontjában, a maximumát pedig a kezdőpontjában veszi fel. Igaz tehát a következő egyenlőtlenséglánc:

$$a = 0 \leq g(1) = 0.56606 \leq g(x) \leq g(0) = 0.86603 \leq 1 = b.$$

Ezzel a konvergencia tétel első feltételét beláttuk. (Vele együtt azt is, hogy az egyenletnek van gyöke a keresett helyen.) A kijelölt intervallumon a deriváltfüggvény, illetve az abszolút értéke:

$$|g'(x)| = \left|\frac{-e^x}{4\sqrt{4 - e^x}}\right| = \frac{e^x}{4\sqrt{4 - e^x}}.$$

A $[0,1]$ intervallumon ez folytonos. Ugyanakkor $|g'(x)|$ szigorúan monoton növekvő, ezért maximumát az intervallum végpontjában veszi fel:

$$|g'(x)| \leq |g'(1)| = 0.60026 = q < 1.$$

Tehát fennáll a konvergeciatételünk (ii) feltétele is. Alkalmazhatjuk az (A) egzakt kilépési feltételt és mivel a hibabecsléshez szükséges $\alpha = q/(1-q) = 1.502$ a példára nézve állandó, ezt előre kiszámoltuk. Legyen az iteráció kezdő adata $x_0 = 1 \in [a,b]$ A számítások részleteit itt is táblázatban foglaltuk össze, bekeretezve a végeredményt:

k	x	$\delta x = \alpha \lvert x_{k+1} - x_k \rvert$
0	1	
1	0.566065	0.65
2	0.748111	0.27
$\vdots$	$\vdots$	$\vdots$
14	$\boxed{0.703439}$	0.43×10^{-6}

9.4 HÚRMÓDSZER

Legyen $f : [a,b] \to \mathbb{R}$ folytonos, $f(a) \cdot f(b) < 0$. Ekkor (a,b)-ben van gyöke f -nek. Legyen $x_0 = a, x_1 = b$. Az x_k és x_s közelítések ismeretében (ahol $f(x_k) \cdot f(x_s) < 0$) az f függvényt közelítsük az $(x_k, f(x_k))$ és $(x_s, f(x_s))$ pontokon átmenő egyenessel. Az egyenes x tengellyel vett metszéspontja legyen x_{k+1}. A következő közelítésben szereplő s index legyen:

$$s = \begin{cases} k, & \text{ha } f(x_k) \cdot f(x_{k+1}) < 0 \\ s, & \text{ha } f(x_k) \cdot f(x_{k+1}) > 0 \end{cases}$$

Képlettel: legyen adott x_0, x_1 úgy, hogy $f(x_0) \cdot f(x_1) < 0$. Ekkor

$$x_{k+1} = x_k - \frac{f(x_k) \cdot (x_k - x_s)}{f(x_k) - f(x_s)} \qquad (k \in \mathbb{N}),$$

ahol s a legnagyobb index, melyre $f(x_s) \cdot f(x_k) < 0$.

9.4.1. TÉTEL. *Ha*

$(i) f \in C^2[a,b]$,
$(ii) f(a) \cdot f(b) < 0$,
$(iii) f'$ *és* f'' *állandó előjelű az* $[a,b]$ *intervallumon*,
$(iv) 0 < m_1 \leq |f'|$,
$(v) |f''| \leq M_2 < \infty$

akkor a húrmódszer konvergens és a hibabecslése:

$$|x_{k+1} - x^*| \leq \frac{M_2}{2m_1} |x_k - x^*| \cdot |x_s - x^*| \qquad (k \in \mathbb{N}).$$

9.5 SZELŐMÓDSZER

Legyen $f : [a, b] \to \mathbb{R}$ folytonos, és tegyük fel, hogy (a, b)-ben van gyöke f -nek. Legyen $x_0 = a, x_1 = b$. Az x_{k-1} és x_k közelítések ismeretében az f függvényt közelítsük az $(x_{k-1}, f(x_{k-1}))$ és $(x_k, f(x_k))$ pontokon átmenő egyenessel. Az egyenes x tengellyel vett metszéspontja legyen x_{k+1}. Képlettel: legyen adott x_0, x_1 úgy, hogy $f(x_0) \cdot f(x_1) < 0$. Ekkor

$$x_{k+1} = x_k - \frac{f(x_k) \cdot (x_k - x_{k-1})}{f(x_k) - f(x_{k-1})} \qquad (k \in \mathbb{N}).$$

9.5.1. TÉTEL. *Ha*

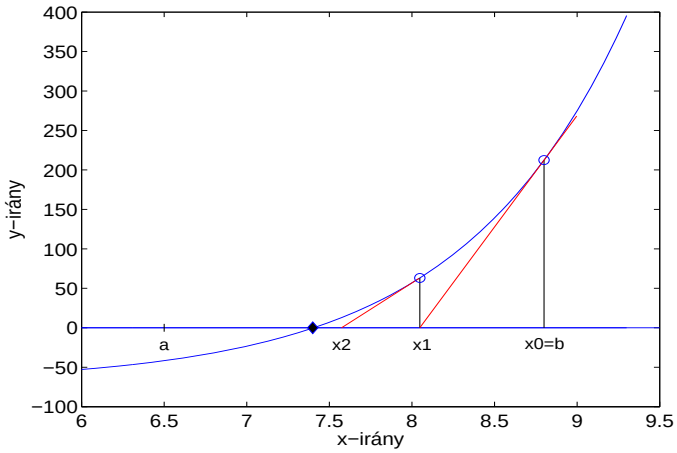
$(i) f \in C^2[a, b]$,
$(ii) f$*-nek létezik* $x^* \in (a, b)$ *gyöke*,
$(iii) f'$ *állandó előjelű az* $[a, b]$ *intervallumon*,
$(iv) 0 < m_1 \leq |f'|$,
$(v) |f''| \leq M_2 < \infty$
(vi)*legyen* $x_0 \in [a, b]$ *olyan, hogy* $|x_0 - x^*| < r = \min \left\{ \frac{2m_1}{M_2}, |a - x^*|, |b - x^*| \right\}$

akkor a megfelelő környezetből indított szelőmódszer konvergens és a konvergencia rendje $p = \frac{1+\sqrt{5}}{2}$. *A hibabecslése:*

$$|x_{k+1} - x^*| \leq \frac{M_2}{2m_1} |x_k - x^*| \cdot |x_{k-1} - x^*| \qquad (k \in \mathbb{N}).$$

9.6 A NEWTON-MÓDSZER

A módszert szokás érintőmódszernek is nevezni. Tegyük fel, hogy $f : \mathbb{R} \to \mathbb{R}$ folytonosan differenciálható. A módszer lényege, hogy az x_k pontban a függvényhez érintőt húzunk és ennek az érintőnek a zérushelye adja meg a keresett gyök $(k + 1)$-edik közelítését, azaz x_{k+1}-et.



Az ábrán az előre felvett kezdeti x_0 (ami gyakran a szóbanforgó intervallum valamelyik alkalmas végpontja) közelítést és további kettőt tüntettünk fel. Az érintő iránytangense a k-adik pontban $f'(x_k)$ és egyenlete

$$y - f(x_k) = f'(x_k)(x - x_k).$$

Az $y = 0$ helyen metszi az x-tengelyt; ezt behelyettesítve:

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

feltéve, hogy $f'(x_k) \neq 0$.

Ehhez képlethez eljuthatunk egy kissé más érveléssel is. Nevezetesen $f(x)$-et linearizáljuk az x_k pontban, azaz közelítjük az elsőfokú Taylor-polinomjával:

$$f(x) \approx f(x_k) + f'(x_k)(x - x_k).$$

Ezután az $f(x) = 0$ egyenletet helyettesítjük az

$$f(x_k) + f'(x_k)(x - x_k) = 0$$

egyenlettel, amelynek gyöke közelíti az $f(x) = 0$ egyenlet gyökét.

A Newton-módszer tehát a következő. Adott egy $x_0 \in \mathbb{R}$ kezdeti közelítés és képezzük az

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad (k = 0, 1, \ldots)$$

sorozatot. A Newton-módszer konvergenciájára az alábbi tételt ismertetjük:

9.6.1. TÉTEL. *Ha az f függvényre teljesülnek a következők:*
(i) $f \in C^2[a, b]$ (kétszer folytonosan differenciálható $[a, b]$-n),

(ii) $f'(x)$, $f''(x) \neq 0$, *ha* $x \in [a,b]$ (*állandó előjelűek*),
(iii) f*-nek van zérushelye* (a,b)*-n,*
(iv) $x_0 \in [a,b]$, *valamint teljesül, hogy* $f'(x_0)$ *és* $f''(x_0)$ *azonos előjelű,*
akkor az $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ *sorozat konvergál az egyetlen* $[a,b]$*-beli* x^* *megoldáshoz. Érvényes továbbá az alábbi hibabecslés:*

$$|x^* - x_k| \leq \frac{M}{2m} \left(x_k - x_{k-1}\right)^2,$$

ahol $0 < m \leq |f'(x)|$ *és* $M \geq |f''(x)|$, *azaz a deriváltak abszolút értékeinek alsó, illetve felső korlátjai* $[a,b]$*-n.*

B i z o n y í t á s. A zérushely egyedüli volta az f szigorú monotonitásából következik ($f' \neq 0$). Az is világos, hogy ekkor az intervallum két végpontjában ellentétes az f előjele, következésképpen az egyik végpont mindig teljesíti a (iv) feltételt. Négy eset különböztethető meg, az $f'(x)$, és $f''(x)$ előjeleitől függően. Mind a négyre hasonló a bizonyítás, ezért csak az ábrán is látható $f'(x) > 0$, $f''(x) > 0$ esetet mutatjuk meg. Ekkor az (iv) feltétel miatt $f(x_0)$ is pozitív, az $f'(x_0)$ pozitívitása és korlátossága miatt (zárt intervallumon folytonos függvény ott korlátos is) pedig $x_1 < x_0$.

Felírva a másodfokú, maradéktagos Taylor-polinomot

$$f\left(x_1\right) = f\left(x_0\right) + f'\left(x_0\right)\left(x_1 - x_0\right) + \frac{f''\left(\mu\right)}{2}\left(x_1 - x_0\right)^2 \quad \left(\mu \in [x_1, x_0]\right),$$

látható, hogy f az érintő fölött halad, így $f(x_1) > 0$ is fennáll.

A második lépésben az x_1 játsza az x_0 szerepét, tehát $x_2 < x_1$, és így tovább. Az $\{x_k\}_{k=1}^{\infty}$ sorozat monoton csökkenő tehát, de korlátos is, mert minden érintő az f alatt marad az intervallumban. Következésképp az x^* egy alsó korlát. Monoton korlátos sorozat konvergens is, legyen a határérték t. Vegyük az $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ sorozat mindkét oldalán a $k \to \infty$ határátmenetet, a folytonosság miatt mindkét oldal határértéke ugyanaz a t, ami csak az $f(t) = 0$ mellett lehetséges, azaz $t = x^*$. A megoldáshoz való konvergenciát ezzel beláttuk.

A hibabecsléshez írjuk fel az elsőfokú maradéktagos Taylor-polinomot az x^* pontban, a másodfokút pedig az x_k-ben:

$$f\left(x_{k+1}\right) = f\left(x^*\right) + f'\left(\xi\right)\left(x_{k+1} - x^*\right),$$

$$f\left(x_{k+1}\right) = f\left(x_k\right) + f'\left(x_k\right)\left(x_{k+1} - x_k\right) + \frac{f''\left(\zeta\right)}{2}\left(x_{k+1} - x_k\right)^2,$$

ahol $\xi,\ \zeta \in [a,b]$. A két egyenlet baloldala egyenlő, így a jobboldalukon álló kifejezések is.

A elsőben $f(x^*) = 0$, a másodikban $f(x_k) + f'(x_k)(x_{k+1} - x_k) = 0$, így

$$f'(\xi)(x_{k+1} - x^*) = \frac{f''(\zeta)}{2}(x_{k+1} - x_k)^2.$$

Vegyük mindkét oldal abszolút értékét, k helyett írjunk $k-1$-et. Rendezzük át az egyenletet és vegyük figyelembe, hogy m az $|f'(\xi)|$ alsó, M pedig az $|f''(\zeta)|$ felső korlátja. Ezzel megkapjuk a hibabecslést.

Ez a hibabecslés mutatja, hogy a konvergencia sebesség másodrendű (négyzetes). Az is egyszerűen belátható, hogy a $k+1$-edik közelítés relatív hibája a k-adikénak a négyzete.

A Newton-módszer algoritmusa:

Input x_0, ε ($\varepsilon > 0$).
$k = 1, \quad x_1 = x_0 - f(x_0)/f'(x_0)$
while kilépési feltétel=hamis
  $x_{k+1} = x_k - f(x_k)/f'(x_k)$
  $k = k + 1$

A kilépési feltételek hasonlóak lehetnek, mint a fixpont iterációnál. Ha igazolni tudtuk a konvergenciát és becsülni az m, M értékeket, akkor az (A) egzakt hibabecslést végezhetünk, egyébként a (B) vagy/és a (C)-t (ez utóbbinál $|f(x_k)| < \varepsilon$).

9.6.2. PÉLDA. Oldjuk meg az $f(x) = 4 - 4x^2 - e^x = 0$ egyenletet a $[0, 1]$ intervallumon, $\delta x_i = \varepsilon = 10^{-6}$ hibakorláttal, Newton-módszerrel. A számítások előtt igazoljuk a konvergenciát is.

Megoldás: Könnyen kiszámolható, hogy

$$f'(x) = -8x - e^x < 0 \quad \text{és } f''(x) = -8 - e^x < 0.$$

Ezekből a deriváltakból azonnal látszik azok állandó előjele is az adott intervallumon; ezzel a konvergenciatétel első két feltételét igazoltuk.

Az $f(0) = 3$ és $f(1) = -e$ értékek ellentétes előjele egyben az (iii) feltétel teljesülését jelenti. Az $x_0 = 1$ választással pedig az utolsó feltétel is fennáll. Van tehát egyedüli x^* megoldás a $[0, 1]$-n és az a Newton-módszerrel megközelíthető. Az egzakt hibabecsléshez szükséges konstansok is könnyen adódnak: mindkét derivált abszolút értéke szigorúan monoton nő, így az első a minimumát az a helyen, a második pedig a maximumát a b helyen veszi fel. Tehát $m = |f'(0)| = 1$ és $M = |f''(1)| = 10.72$. Itt is előre kiszámolhatjuk a példára nézve állandó $\beta = M/(2m) = 5.36$ értéket.

Az iterációk eredményeit itt is táblázatba foglaltuk.

i	x	$\delta x = \beta\,(x - x_{elozo})^2$
0	1	
1	0.74638828573	0.35
2	0.70459003270	0.94×10^{-2}
3	0.70344043705	0.71×10^{-5}
4	0.70343957116	0.41×10^{-11}

9.7 AZ ÉRINTŐ PARABOLA-MÓDSZER

Ennél a módszernél az érintő egyenesek helyett érintő parabolákat használunk.

Legyen adott egy $[a, b]$ intervallum, amelyben az $f(x) = 0$ egyenletnek van gyöke. Tegyük fel, hogy f differenciálható $[a, b]$-n és $f(b) > 0$. (Ha nem teljesül, akkor a $-f(x) = 0$ egyenletet oldjuk meg). A $(b, f(b))$ pontba végtelen sok olyan érintőparabolát illeszthetünk, amely metszi az x tengelyt.

Ugyanis a

$$p(x) = f(b) + f'(b)(x - b) - M(x - b)^2, \qquad M > 0$$

függvények bármelyikének a képe parabola, és $f(b) = p(b), f'(b) = p'(b)$. Az $M > 0$ miatt az x tengelyt két pontban metszi mindegyik parabola.

Mivel az $[a, b]$-ben keresünk gyököt, ezért a baloldali metszéspontot vesszük figyelembe. A $p(x) = 0$ egyenletből:

$$x = x_1 = b + \frac{f'(b)}{2M} - \frac{1}{2M}\sqrt{(f'(b))^2 + 4Mf(b)}.$$

Innen a módszer formulája (b helyére x_k-t, x_1 helyére x_{k+1}-et írva):

$$x_{k+1} = x_k + \frac{f'(x_k)}{2M} - \frac{1}{2M}\sqrt{(f'(x_k))^2 + 4Mf(x_k)} \qquad k = 0, 1, 2 \ldots$$

A következő tétel megadja, hogy milyen M értékeket érdemes használni.

9.7.1. TÉTEL. *Az $f(x) = 0$ egyenlet esetén legyen $f \in C^2[a, b]$, legyen $f(b) > 0$ és tegyük fel, hogy*

$$M \geq \frac{1}{2} \max_{x \in [a,b]} |f''(x)|\,.$$

Ekkor az $x_0 = b$ választással a módszer konvergál a b ponthoz legközelebbi $[a, b]$-beli gyökhöz (ha az $[a, b]$ intervallumban van gyök), ellenkező esetben véges sok lépés után $x_k < a$.

10 NEMLINEÁRIS EGYENLETRENDSZEREK MEGOLDÁSA

Az $x = G(x)$ ($G : \mathbb{R}^n \to \mathbb{R}^n$) és az $F(x) = 0$ ($F : \mathbb{R}^n \to \mathbb{R}^n$) alakú egyenletrendszerek közelítő megoldási módszerei közül említünk kettőt. Mindkettő a már egydimenzióban megismert valamelyik módszer kiterjesztése többdimenzióra.

10.1 FIXPONT ITERÁCIÓS ELJÁRÁS

Az $x = G(x)$ egyenletrendszer fixpont iterációját tulajdonképpen már elmondtuk (a lineáris egyenletrendszereknél is és az egydimenziós eljárásoknál is). Ha a $D \subseteq \mathbb{R}^n$ zárt tartományra teljesül a

$$G(x) \in D, \quad x \in D$$

és valamilyen indukált normában a

$$\|G(x) - G(y)\| \leq q \|x - y\|, \quad x, y \in D \quad (0 \leq q < 1)$$

kontrakciós feltétel, akkor tetszőleges $x^{(0)} \in D$ esetén az alábbi algoritmus lineáris sebességgel konvergál az $x = G(x)$ egyenlet egyetlen megoldásához.

A fixpont iteráció algoritmusa:

1 Input $x^{(0)}$, $\varepsilon > 0$.
2 $k = 1, \quad x^{(1)} = G\left(x^{(0)}\right)$
3 **WHILE** kilépési feltétel=hamis **DO**
4 $\quad x^{(k+1)} = G\left(x^{(k)}\right)$
5 $\quad k = k + 1$

10.1.1. MEGJEGYZÉS. A kilépési feltételek is azonosak az egydimenziós esettel, annyi módosítással, hogy abszolút érték helyett normát veszünk és felső indexet használunk.

10.2 A NEWTON-MÓDSZER

Itt is valamely $x^{(0)} \in \mathbb{R}^n$ kezdeti vektorból kiindulva állítunk elő egy sorozatot, amely reményeink szerint (bizonyos feltételek megléte esetén garantáltan) az egyenletrendszer megoldásához konvergál. Az $F(x) = 0$ ($x \in \mathbb{R}^n$) egyenletrendszer skalár alakja:

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ &\vdots \quad \vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0. \end{aligned}$$

Az egyes skalár egyenleteket linearizáljuk az $x^{(k)} = [x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}]^T \in \mathbb{R}^n$ pontban, majd az így kapott lineáris egyenletrendszer megoldása szolgáltatja az $x^{(k+1)}$ közelítést. Az i-edik skalár függvény elsőfokú Taylor-polinomja az $x^{(k)}$ pontban:

$$\begin{aligned} f_i(x_1, x_2, \ldots, x_n) &\approx f_i(x_1^{(ki)}, x_2^{(k)}, \ldots, x_n^{(k)}) \\ &\quad + \sum_{j=1}^{n} \frac{\partial}{\partial x_j} f_i(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}) \left(x_j - x_j^{(k)}\right). \end{aligned}$$

Tömörebb formában ugyanez

$$\begin{aligned} f_1(x) &\approx f_1(x^{(k)}) + \nabla f_1(x^{(k)})^T (x - x^{(k)}) \\ &\vdots \\ f_n(x) &\approx f_n(x^{(k)}) + \nabla f_n(x^{(k)})^T (x - x^{(k)}), \end{aligned}$$

ahol $\nabla f_i(x^{(k)})$ az i-edik skalár függvény gradiense az $x^{(k)}$ pontban, azaz az F Jacobi-mátrixának i-edik sorvektora az $x^{(k)}$ pontban. Az $F(x) = 0$ egyenletrendszer megoldása helyett keressük a linearizált egyenletrendszer

$$\begin{aligned} f_1(x^{(k)}) + \nabla f_1(x^{(k)})^T (x - x^{(k)}) &= 0 \\ &\vdots \\ f_n(x^{(k)}) + \nabla f_n(x^{(k)})^T (x - x^{(k)}) &= 0, \end{aligned}$$

megoldását. Vegyük észre, hogy az $y = f_i(x^{(k)}) + \nabla f_i(x^{(k)})^T (x - x^{(k)})$ egyenlet az $y = f_i(x)$ függvény érintő (hiper)síkja az $x^{(k)}$ pontban. Az $x^{(k+1)}$ közelítés az érintősíkoknak az $y = 0$ síkon lévő közös pontja.

Az F függvény

$$J(x) = \left[\frac{\partial f_i(x)}{\partial x_j}\right]_{i,j=1}^{n} = \begin{bmatrix} \nabla f_1(x)^T \\ \vdots \\ \nabla f_n(x)^T \end{bmatrix}$$

Jacobi-mátrixával tömörebben is megfogalmazható a linearizált egyenletrendszer:

$$F(x^{(k)}) + J(x^{(k)})(x - x^{(k)}) = 0. \tag{23}$$

Ennek megoldása:

$$x^{(k+1)} = x^{(k)} - \left[J\left(x^{(k)}\right)\right]^{-1} F(x^{(k)}) \qquad (k = 0, 1, \ldots). \tag{24}$$

Az $\left\{x^{(k)}\right\}$ sorozat fenti előállítását nevezzük Newton-módszernek. Amennyiben a skalár számmal való osztást úgy tekintjük, hogy az az inverzével (reciprokával)

balról történő szorzás, akkor tökéletes analógiát látunk az () és az egydimenziós Newton-módszer között. Ugyanígy az algoritmusok között is, annyi eltéréssel, hogy a gyakorlatban soha nem invertáljuk a $J(x^{(k)})$ Jacobi-mátrixot. Helyette az () lineáris egyenletrendszert oldjuk meg valamely alkalmas módszerrel, a $\Delta^{(k)} = x - x^{(k)}$ új változó bevezetésével.

A Newton-módszer algoritmusa egyenletrendszerekre:

1 Adott $x^{(0)} \in \mathbb{R}^n$, $\varepsilon > 0$.
2 **FOR** $i = 0, 1, 2, \ldots$
3 Oldjuk meg a $J(x^{(k)})\Delta^{(k)}$
4 $x^{(k)} = x^{(k)} + \Delta^{(k)}$

Az algoritmust itt taxatív ciklussal írtuk le; természetesen gondoskodni kell a kilépési feltételről. Az alkalmazható kilépési feltétel itt is a korábban megadott (B) vagy/és (C) lehet, értelemszerűen valamilyen normában alkalmazva. (Természetesen az iterációszám korlátozásáról se feledkezzünk meg.) Egzakt (A) hibabecslésre igen ritkán van módunk. Ismerünk ugyan konvergenciatételeket, amelyek becslést is adnak a hibára, ám azok feltételeit a legritkább esetben sikerül belátni, illeve az azokban szereplő konstansokat megbecsülni (a költségekről nem is beszélve). Mindenesetre ezen tételek igazolják, hogy az eljárás konvergenciája alkalmas feltételek esetén lokális és másodrendű.

10.2.1. MEGJEGYZÉS. A lokális konvergencián azt értjük, hogy csak bizonyos tartományból, rendszerint az x^* megoldás szűk környezetéből választott $x^{(0)}$ kezdeti vektor esetén konvergens. Ebben az értelemben az egy- vagy többdimenziós fixpontiteráció és az egydimenziós Newton-módszer is lokálisan konvergens.

10.2.2. PÉLDA. Oldjuk meg az $F(x) = [f_1(x), f_2(x)]^T = 0, \quad x = [x_1, x_2]^T$ egyenletrendszert Newton-módszerrel, ha

$$\begin{aligned} f_1(x) &= x_1 + x_2^2 - 5 \\ f_2(x) &= x_1^3 - x_2\sqrt{x_1} - 3. \end{aligned}$$

Megoldás: A Jacobi-mátrix:

$$J(x) = \begin{bmatrix} 1 & 2x_2 \\ 3x_1^2 - \frac{x_2}{(2\sqrt{x_1})} & -\sqrt{x_1} \end{bmatrix}.$$

Legyen a kezdővektor $x^{(0)} = [3, -5]^T$. A

$$J(x^{(0)})\Delta^{(0)} = -F(x^{(0)})$$

lineáris egyenletrendszer:

$$\begin{bmatrix} 1 & -10.000000 \\ 28.443376 & -1.732051 \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} = - \begin{bmatrix} 23.000000 \\ 32.660254 \end{bmatrix}.$$

Az egyszerűség kevéért itt elhagytuk a Δ vektor felső indexeit. A megoldás:

$$\begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} = \begin{bmatrix} -1.014374 \\ 2.198563 \end{bmatrix}, \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \end{bmatrix} = x^{(0)} + \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} = \begin{bmatrix} 1.985626 \\ -2.801437 \end{bmatrix}.$$

A kilépési feltételhez szükséges normák:

$$\|\Delta\|_\infty = 2.198563, \quad \left\|F(x^{(1)})\right\|_\infty = 8.776312.$$

A továbbiakban pedig

$$\begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \end{bmatrix} = \begin{bmatrix} 1.384174 \\ -2.046071 \end{bmatrix},$$

$$\|\Delta\|_\infty = 0.755366, \quad \left\|F(x^{(2)})\right\|_\infty = 2.059214.$$

$$\vdots$$

$$\begin{bmatrix} x_1^{(6)} \\ x_2^{(6)} \end{bmatrix} = \begin{bmatrix} 1.000000 \\ -2.000000 \end{bmatrix},$$

$$\|\Delta\|_\infty = 0.16 \times 10^{-4}, \quad \left\|F(x^{(2)})\right\|_\infty \approx 10^{-9}.$$

Megjegyezzük, hogy a gyök szoros közelében az eljárás ugyanúgy begyorsul, mint egydimenzióban; míg az $x^{(5)}$öt tizedesjegyre volt pontos, az $x^{(6)}$ már valójában 10 jegyre, a 7-lépésben pedig már elértük a lehetséges 15 tizedesjegyű maximális pontosságot és a kilépési normák is 10^{-9}, illetve 10^{-15} nagyságrendűek.