Hegedűs Csaba

NUMERIKUS ANALÍZIS

Jegyzet

ELTE, Informatikai Kar

TARTALOM

1. Gépi szám, hibák

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

1.1. A gépi számok

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

$$\underset{\text{előjel, } t \text{ db bináris jegy}}{\pm .101\ldots01} \cdot 2^{\overset{\nwarrow \text{kitevő}}{k}} = \pm m \cdot 2^k \tag{1.1}$$

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

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

1	8	23

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

1	11	52

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

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

1.2. Nevezetes gépi számok

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

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

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

1.3. Valós számok konverziója gépi számmá

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

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

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

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

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

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

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

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

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

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

ahol $\varepsilon_M = \varepsilon_1/2 = 2^{-t}$ a *gépi epszilon*, ez adja az ε_0 és M_∞ közé eső szám ábrázolásának relatív hibáját. Itt az első sornak csak jelzés értéke van. A második sor önmagáért beszél, egyedül a harmadik

sor kíván némi magyarázatot. Azt fejezi ki, hogy az ábrázolt szám hibája nem nagyobb, mint a t-edik bináris jegyben elkövetett hiba. A harmadik sor átrendezése a relatív hiba korlátját adja:

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

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

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

1.4. A gépi aritmetika

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

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

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

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

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

$$\begin{array}{ll} 0.3460\ +02 & 0.3460\ +02 \\ \underline{0.00004524\ +02} & \underline{0.00003872\ +02} \\ 0.3460\ +02 & 0.3460\ +02 \end{array}$$

de először a két kicsi számot összeadva

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

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

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

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

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

- Összevonás után az eredmény relatív hibája jelentősen megnőhet. Például

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

ami egyenlő $0.1100\ +00$-val. Látjuk, itt már csak az első két jegy pontos. Ezt jelenséget *kivonási jegyveszteségnek* nevezzük. Néha adhatók fogások a kivonási jegyveszteség elkerülésére vagy csökkentésére, pl. ha $\sqrt{3472}-\sqrt{3471}$-et így számítjuk, kihasználva, hogy a gyök alatt egész számok vannak:

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

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

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

- Előfordulhat olyan eset, amikor a közbülső eredmény túlcsordul (nagyobb mint M_∞), emiatt rossz a program futása, pedig a végeredmény az ábrázolható számok közt van. Például legyen $a=0.3265+60$, $b=0.5671\ +02$ és számítandó $\sqrt{a^2+b^2}$. Az első szám kitevője négyzetre emeléskor 120, így túlcsordult számot kapunk. Ha viszont $s\sqrt{(a/s)^2+(b/s)^2}$-et számítjuk, ahol $s=\max(|a|,|b|)$, akkor ez nem következik be.

- Néha arra is számítani kell, hogy egy függvény nem adja olyan pontossággal vissza a helyettesítési értéket, mint amilyen pontossággal indultunk. Például tekintsük a sin függvényt. Ha az argumentum kicsi, akkor nincs semmi baj. Ha azonban x értéke nagy, például $x=2356$, akkor $\sin(2356)$ számításakor $2356\ \pi$-vel vett osztási maradékát kell vennünk. A maradékban már csak 1 jegy lesz pontos ha a fenti aritmetikát használjuk, így az eredménynél sem remélhetünk nagyobb pontosságot.

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

1.5. Hibák

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

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

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

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

A valóságban a Δx hibát nem ismerjük, csak annak felső korlátját. Emiatt kiindulásul annyit tudunk, hogy x^* az x érték valamely Δ_x-sugarú környezetében van.

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

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

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

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

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

$$f(x) = f(x^*) + f'(x^*)\Delta x + f''(\xi)(\Delta x)^2/2.$$

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

$$f(x) - f(x^*) = \Delta f \approx f'(x^*)\Delta x. \tag{1.7}$$

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

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

Az abszolút értékekre áttérve:

$$|\delta f| \approx c(f,x)|\delta x|, \tag{1.8}$$

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

Az (1.7) és (1.8) összefüggések sugallják a következő stabilitás fogalmat: egy algoritmus *stabil,* ha két bemenő érték: x_1, x_2 és a hozzájuk tartozó kimenő értékek, f_1, f_2 között fennáll egy

$$|f_2 - f_1| \le C|x_1 - x_2|, \quad x_1, x_2 \in X \tag{1.9}$$

típusú összefüggés, ahol C az algoritmus adataitól független nem túlságosan nagy állandó. Vegyük észre, itt x_i, f_i gépi számok, egy véges halmaz elemei.

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

2. Normák, egyenlőtlenségek

Ebben a szakaszban vektorok és mátrixok között távolságfüggvényeket fogunk bevezetni.

1.1. Metrikus tér

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

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

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

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

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

2.1. A vektorok hatványnormája

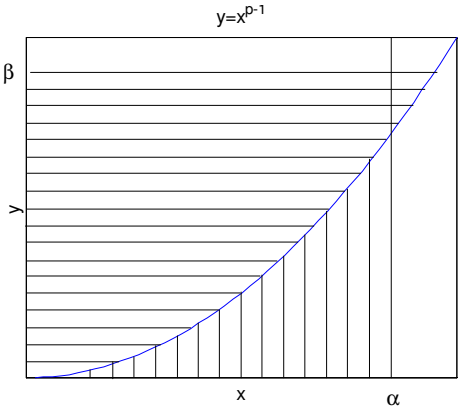
A vektor normája $\|x\|:\ \mathbb{R}^n\to\mathbb{R}$ a következő tulajdonságokkal rendelkezik:

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

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

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

a harmadikat később fogjuk belátni.



2.2. A Hölder-egyenlőtlenség

A hatványnormákra fennáll a Hölder-egyenlőtlenség:

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

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

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

Ezután az

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

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

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

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

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

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

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

2.3. A hatványnormák néhány tulajdonsága

A hatványnormákra teljesül:

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

hiszen ez az összefüggés átrendezhető a

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

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

A fontosabb hatványnormák a következők:

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

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

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

az x vektor euklidészi, kettes vagy gömbnormája.

A $p \to \infty$ határesetben adódik

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

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

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

2.4. Konvergencia normában. A normák ekvivalenciája

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

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

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

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

2.5. Mátrixnormák

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

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

Az utolsó két tulajdonságot akkor követeljük meg, ha a két mátrix összeadható vagy összeszorozható. Mivel a vektorok speciális mátrixoknak tekinthetők, minden mátrixnorma meghatároz egy vektor-normát, amelyet a mátrixnormával *kompatibilis* vektornormának nevezünk. Ez az út fordítva is bejárható, ugyanis minden vektornorma *indukál egy mátrixnormát* a következőképpen:

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

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

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

2.6. Tétel

Az indukált mátrixnorma eleget tesz a (2.6) feltételeknek.

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

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

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

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

2.7. Az indukált mátrixnormák meghatározása

$p = 1$, *oszlopnorma:*

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

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

$p = \infty$, *sornorma*:

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

Legyen $\|x\|_\infty = 1$, ekkor $\|Ax\|_\infty = \max_{(i)} \left| \sum_{j=1}^n a_{ij} x_j \right| \le \max_{(i)} \sum_{j=1}^n |a_{ij}| |x_j| \le \max_{(i)} \sum_{j=1}^n |a_{ij}|$. Tegyük fel, a maximum a k -adik sorra következett be. Ekkor $\|x\|_\infty = 1$ és $\|Ax\|_\infty$ éppen a megállapított felső korlát az $x = [x_j] = [\overline{a}_{kj} / |a_{kj}|]$ vektorral, ahol a felülvonás komplex konjugáltat jelöl.

$p = 2$, *spektrál norma:*

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

Ekkor a következő maximumot keressük:

$$\|A\|_2^2 = \max \frac{\|Ax\|_2^2}{\|x\|_2^2} = \max \frac{x^T A^T A x}{x^T x}.$$

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

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

ami mutatja, hogy a maximumot találtuk meg.

2.8. A spektrálsugár és az indukált normák összefüggése

Egy mátrix *spektrál sugara* alatt a következőt értjük:

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

ahol $\lambda_k(A)$ az A mátrix sajátértéke. Az $\|A\|$ mátrixnorma és az $\|x\|$ vektornorma *illeszkedő,* ha bármely x-re eleget tesznek a (2.8) összefüggésnek. Ez utóbbi definíció arra az esetre szól, amikor a vektornorma nem kompatibilis, vagy a mátrixnorma nem a vektornormából indukált, mert különben az illeszkedés triviális. Igaz az összefüggés:

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

ahol $\|A\|$ tetszőleges norma, mert $Au_k = \lambda_k u_k$, $\| u_k \| = 1$ mellett a vektorokra is ugyanazt a mátrixnormát alkalmazva

$$|\lambda_k| \|u_k\| = |\lambda_k| = \|Au_k\| \le \|A\| \|u_k\| = \|A\|, \quad \forall k\text{-ra.}$$

A (2.13) reláció akkor is igaz, ha olyan mátrixnormánk van, ami csak négyzetes mátrixokra van definiálva. Ekkor az $Au_k u_k^T = \lambda_k u_k u_k^T$ kifejezést képezve lehet a bizonyítást megismételni.

2.9. A lineáris egyenletrendszer megoldásának perturbációi

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

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

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

Az eredeti és a perturbált értékekre vonatkozó egyenletekből

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

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

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

2.9.1 Lemma.

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

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

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

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

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

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

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

az utolsó lépésben felhasználtuk az előbbi lemmát.

2.10. A mátrix kondiciószáma

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

2.11. A relatív maradék

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

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

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

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

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

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

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

Ekkor 2-es normában $\left\|r\tilde{x}^T\right\|_2=\|r\|_2\left\|\tilde{x}^T\right\|_2$ (l. 2.5 gyakorlat), s ezzel $\dfrac{\|\delta A\|_2}{\|A\|_2}=\dfrac{\|r\|_2}{\|A\|_2\|\tilde{x}\|_2}$.

2.12. Gyakorlatok

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

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

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

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

2.5. $\left\|ab^T\right\|_2 = \|a\|_2 \|b\|_2$. $\left\|ab^T\right\|_1 = \|a\|_1 \|b\|_\infty$. $\left\|ab^T\right\|_\infty = \|a\|_\infty \|b\|_1$.

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

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

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

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

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

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

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

2.13. $\|AB\|_2 = \|BA\|_2$, ha $A = A^T$ és $B = B^T$.

2.14. $\mathrm{cond}_2(A^T A) = \mathrm{cond}_2^2(A)$.

3. A numerikus lineáris algebra egyszerű transzformációi

3.1. Jelölések

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

3.2. A mátrixok szorzása

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

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

A továbbiakban rátérünk speciális mátrixok ismertetésére.

3.3. Permutáció-mátrix

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

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

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

3.4. Diáddal módosított egységmátrix

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

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

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

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

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

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

3.5. Példa

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

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

Az inverze:

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

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

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

3.6. Példa

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

Megoldás. $A + \alpha A e_i e_k^T = A(I + \alpha e_i e_k^T)$.

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

3.7. Példa

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

Megoldás. Feltesszük, az a és b vektorok egyike sem zérus, mert különben a feladat triviális volna. Legyen az a vektor i-edik eleme $e_i^T a = a_i \neq 0$, és tekintsük az $I - (a/a_i - e_i)e_i^T$ mátrixot. Ennek minden átlóeleme 1 és az i-edik oszlopában vannak még nemzérus elemek. De ezeket a nemzérus elemeket az i-edik sor valamely számszorosának hozzáadásával ki lehet nullázni, ebből következik, hogy a determinánsa 1. Most szorozzuk az $I + ab^T$ mátrixot balról $I - (a/a_i - e_i)e_i^T$-vel. Ez az a vektort az $a_i e_i$ vektorba viszi, így az eredmény: $I - (a/a_i - e_i)e_i^T + a_i e_i b^T$, amely már csak az i-edik sorában és oszlopában különbözik az egységmátrixtól. Most szorozzuk a kapott mátrix k-adik oszlopát a_k/a_i-vel és adjuk hozzá a i-edik $(i \neq k)$ oszlophoz (ld. 3.6 Példa):

$$\left(I - (\frac{a}{a_i} - e_i)e_i^T + a_i e_i b^T\right)\left(I + \frac{a_k}{a_i}e_k e_i^T\right) = I - \left(\frac{a - a_k e_k}{a_i} - e_i\right)e_i^T + a_i e_i b^T + a_k b_k e_i e_i^T.$$

Mint látjuk, az a vektor k-adik eleme kinullázódott, és az i-edik átlóelem $1 + a_i b_i + a_k b_k$ lett. Ezt a műveletet minden $k \neq i$-re végrehajtva az a/a_i vektor minden átlón kívüli eleme kinullázódik, az i-

edik átlóelem $1+b^T a$, a többi pedig 1-gyel egyenlő. A következő fázisban az e_k^T, $k \neq i$ sorvektorokkal az $a_i e_i b^T$ sorvektor nemdiagonális elemeit a determináns megváltozása nélkül kinullázhatjuk.

3.8. Diádösszegek, kifejtések

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

$$AB = \sum_{i=1}^n A e_i e_i^T B \ ,$$

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

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

Tudjuk, az n-edrendű x vektor előállítása az egységvektorok segítségével $x = \sum_{i=1}^n e_i (e_i^T x)$. Az előállítás hasonló a $\{q_i\}_{i=1}^n$ ortonormált vektorrendszerrel. Ugyanis vezessük be a $Q = [q_1 q_2 \dots q_n]$ mátrixot. Ekkor $Q^T Q = I = Q Q^T$ az ortonormáltság miatt, tehát írható $x = Q Q^T x = \sum_{i=1}^n q_i (q_i^T x)$. Az ilyen tulajdonságú Q mátrixokat *ortogonális* (komplex megfelelője: *unitér*) mátrixoknak nevezzük.

3.9. Definíció

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

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

3.10. Tétel, mátrix egyszerű szorzatokra bontása

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

Bizonyítás. Megadjuk az eljárást. Az első lépésben vizsgáljuk meg az A mátrix első oszlopát. Ha az első elem $a_{11} = e_1^T A e_1 \neq 0$, akkor sorcserére nincs szükség. Ha az első elem zérus, akkor az oszlopban keresünk egy nemzérus elemet, majd ennek a sorát felcseréljük az első sorral. Ha az oszlop minden eleme zérus volna, akkor nem lenne invertálható a mátrix. Az első permutáció mátrixot jelüljük Π_1-gyel és legyen $A_1 = \Pi_1 A$.

Most szorozzuk A_1-et a $T_1 = I - (A_1 e_1 - e_1) e_1^T / e_1^T A_1 e_1$ mátrixszal. Tudjuk, ennek eredményeként az első oszlop e_1-be megy át és $T_1^{-1} = I + (A_1 e_1 - e_1) e_1^T$. A második lépésben hasonlóan járunk el $T_1 A_1$ második oszlopával: $A_2 = \Pi_2 T_1 A_1$ olyan mátrix lesz, ahol a 22-es pozición nemzérus elem van. Így a $T_2 = I - (A_2 e_2 - e_2) e_2^T / e_2^T A_2 e_2$ mátrixszal szorozva a második oszlopot az e_2 vektorba visszük. Vegyük észre, T_2 az e_1 vektort helyben hagyja.

Hasonlóan folytatva, az i-edik lépésben $A_i = \Pi_i T_{i-1} A_{i-1}$ olyan mátrix, ahol az ii pozicióban nemzérus áll. (Ha az i-edik oszlop zérus volna, ismét ellentmondásba kerülnénk azzal a feltevéssel, hogy a mátrix nemszinguláris.) Ekkor a $T_i = I - (A_i e_i - e_i) e_i^T / e_i^T A_i e_i$ mátrixszal szorozva kapunk e_i-t az i-

edik oszlopban és az eddig elkészült kisebb indexű egységvektorok sem romlottak el. A n-edik lépés után egységmátrixot kapunk, tehát végeredményben a mátrix inverzével szoroztunk. A szorzatokat összegyűjtve:

$$\Pi_1^T T_1^{-1} \Pi_2^T T_2^{-1} \ldots T_n^{-1} = A\,.$$

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

3.11. Háromszögmátrixok szorzatokra bontása

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

$$L = \left(I + (L-I)e_1 e_1^T\right)\left(I + (L-I)e_2 e_2^T\right)\ldots\left(I + (L-I)e_n e_n^T\right),$$

ami tömören így is írható

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

ha megjegyezzük, hogy a tényezők növekvő indexek szerint mindig balról jobbra haladva írandók. A kifejezést közvetlenül is igazolhatjuk a j-edik oszlop meghatározásával. Jobbról az e_j vektorral szorozva az első e_j vektortól különböző eredményű szorzat $e_j + Le_j - e_j = Le_j$ az L mátrix j-edik oszlopa. A többi szorzatban lévő e_k, $k < j$ vektorral ennek a skaláris szorzata zérus, emiatt a végeredmény Le_j. Felírható a sorvektorokkal is a szorzatokra bontás:

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

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

A U felső háromszögmátrixra vonatkozó hasonló összefüggések:

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

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

3.12. Vetítőmátrixok

Tekintsük a

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

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

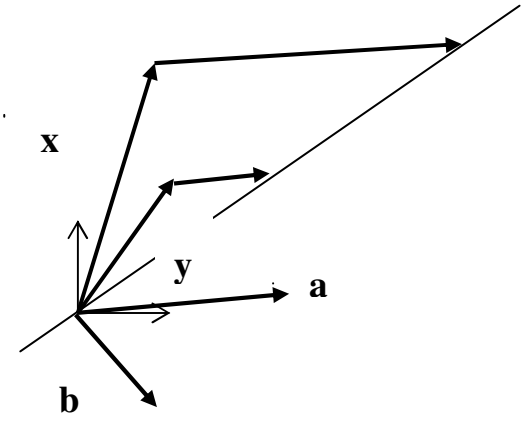
$$\left(I - ab^T\right)\left(I - ab^T\right) = I - 2ab^T + ab^T ab^T = I - ab^T.$$

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

Ha $a = b$, akkor a mátrix szimmetrikus. A szimmetrikus vetítőmátrixok *ortogonális* vetítők, mert egy altérre merőleges vetítést valósítanak meg. Ha a és b nem egyirányú, akkor *ferde* vetítésről

beszélünk. Szokás még a projektorokat *idempotens* mátrixoknak nevezni arra a tulajdonságukra utalva, hogy a mátrix minden hatványa önmaga. Vegyük észre, (3.3)-ból: $Pa = 0$ és $b^T P = 0$.

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



1. Ábra

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

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

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

3.13. Involutórius mátrixok

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

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

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

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

Innen látható, 1-nél nagyobb méretű egységmátrixból végtelen sokféleképp lehet gyököt vonni.

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

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

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

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

3.14. Blokk mátrixok

A mátrixokat nemcsak skalár elemekből rakhatjuk össze, hanem kisebb méretű mátrixokból is. Az ilyen mátrix elemeit *blokkoknak* nevezzük, ha pedig egy mátrixot kisebb mátrixokra bontunk, akkor a mátrixot *blokkosítjuk*. A blokkosítás történhet a következőképp: Az egységmátrixot az oszlopok

mentén felszeleteljük k részre: $I=[E_1,E_2,\ldots,E_k]$. Ha a sorokat ugyanilyen módon osztjuk fel blokkokra, akkor az ij-edik blokk $A_{ij}=E_i^TAE_j$ és a mátrix:

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

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

4. Mátrixok LU-felbontása, Gauss-Jordan algoritmus

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

4.1. Tétel, LU-felbontás.

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

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

$$\left(I - (Ae_1 / a_{11} - e_1)e_1^T\right) Ae_1 = Ae_1 - Ae_1 + a_{11}e_1 = a_{11}e_1 \tag{3.4}$$

-be megy át, tehát

$$L_1^{-1}A = \begin{bmatrix} a_{11} & * & \dots & * \\ 0 & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \dots & * \end{bmatrix}, \tag{3.5}$$

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

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

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

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

$$L = L_1 L_2 \dots L_{n-1}, \quad U = L_{n-1}^{-1} L_{n-2}^{-1} \dots L_1^{-1} PA = \begin{pmatrix} * & * & \dots & * \\ & * & \dots & * \\ & & \ddots & \vdots \\ & & & * \end{pmatrix}. \tag{3.6}$$

■

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

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

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

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

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

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

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

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

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

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

4.2. Az LU-felbontás műveletigénye.

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

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

4.3. Blokk LU-felbontás

Néha célszerű a felbontást – vagy a mátrix invertálását – blokkosított formában végezni. Tipikusan ez a helyzet akkor, amikor az egyik elkülönített blokk egyszerűen invertálható, például azért mert egységmátrix, vagy alsó ill. felső háromszögmátrix. Mi most a blokk *LU*-felbontást a 2×2-es esetben fogjuk megnézni. A főelem ilyenkor blokk, amelyről fel kell tételeznünk, hogy létezik az inverze.

Legyen az egységmátrix egy felosztása $I=[E_1,E_2]$, $A_{ij}=E_i^T A E_j$, ekkor az L mátrix a (3.4)-ben látható L_1 mátrix blokkos megfelelője (ld. még *3.13 Gyakorlat*)

$$L=I-\left(AE_1A_{11}^{-1}-E_1\right)E_1^T \tag{3.7}$$

és a mátrix blokkos felbontása a következő:

$$\begin{bmatrix} \boxed{A_{11}} & A_{12} \\ A_{21}A_{11}^{-1} & A_{22}-A_{21}A_{11}^{-1}A_{12} \end{bmatrix},\ \text{ahol}\ L=\begin{bmatrix} I_1 & 0 \\ A_{21}A_{11}^{-1} & I_2 \end{bmatrix},\ U=\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22}-A_{21}A_{11}^{-1}A_{12} \end{bmatrix}. \tag{3.8}$$

4.4. Schur-komplemens.

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

4.5. Particionált inverz

A (3.8) felbontás alapján írhatjuk:

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

ahonnan

$$A^{-1}=\begin{bmatrix} I_1 & -A_{11}^{-1}A_{12} \\ 0 & I_2 \end{bmatrix}\begin{bmatrix} A_{11}^{-1} & \\ & \left(A\middle|A_{11}\right)^{-1} \end{bmatrix}\begin{bmatrix} I_1 & 0 \\ -A_{21}A_{11}^{-1} & I_2 \end{bmatrix}. \tag{3.9}$$

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

$$A^{-1}=\begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}+\begin{bmatrix} -A_{11}^{-1}A_{12} \\ I_2 \end{bmatrix}\left(A\middle|A_{11}\right)^{-1}\begin{bmatrix} -A_{21}A_{11}^{-1} & I_2 \end{bmatrix} \tag{3.10}$$

alakban.

4.5.1 Feladatok

1. A *3.13 Gyakorlat* alapján mutassuk meg, hogy a (3.7) mátrix inverze úgy készíthető, hogy a 21-es blokk negatívját vesszük. A felső háromszögmátrixra vonatkozó eredmény innen transzponálással származtatható.

2. L_{11} alsó háromszögmátrix, amelyet alul kiegészítünk egy blokk-sorral $\begin{bmatrix} L_{21} & L_{22} \end{bmatrix}$ nagyobb méretű alsó háromszögmátrixra. Mutasssuk meg, hogyha a diagonálblokkok invertálhatók, akkor particionált inverz formulával a kiegészített mátrix inverze

$$\begin{bmatrix} L_{11} & \\ L_{21} & L_{22} \end{bmatrix}^{-1}=\begin{bmatrix} L_{11}^{-1} & \\ -L_{22}^{-1}L_{21}L_{11}^{-1} & L_{22}^{-1} \end{bmatrix}$$

4.6. A Gauss-Jordan módszer az inverz mátrix kiszámítására

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

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

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

$$\left(I - \frac{A_i e_i - e_i}{e_i^T A_i e_i} e_i^T \right) A_i = A_i - \frac{A_i e_i e_i^T A_i}{e_i^T A_i e_i} + \frac{e_i e_i^T A_i}{e_i^T A_i e_i}.$$

Itt jobb oldalon az első két tag azt a diád-levonást jelenti, amit már megismertünk. Az LU-felbontáshoz képest azonban eltérés, hogy az i-edik sor és oszlop kivételével minden területre kell alkalmaznunk. Az harmadik tag azt mutatja, hogy az i-edik sort a főelemmel kell osztani, az első két tagból származó i-edik sor ugyanis zérus.

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

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

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

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

Az utolsó lépésben az induló egységmátrix helyén megjelent az inverz.

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

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

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

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

$$\rightarrow \begin{bmatrix} 3/2 & -3/2 & -1/2 \\ 1/2 & 3/2 & -1/2 \\ -1/2 & -1/2 & 1/2 \end{bmatrix} \overset{\substack{1\leftrightarrow 2\\ \text{oszlopcsere}}}{\rightarrow} \begin{bmatrix} -3/2 & 3/2 & -1/2 \\ 3/2 & 1/2 & -1/2 \\ -1/2 & -1/2 & 1/2 \end{bmatrix}.$$

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

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

5. Az LU-felbontás tulajdonságai, speciális inverzek

5.1. Szimetrikus pozitív definit mátrixok

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

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

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

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

5.1.1 Tétel, elegendő feltétel pozitív definitségre.

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

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

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

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

Bizonyítás. Legyen A blokkos alakja

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

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

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

5.1.3 Tétel, pozitív szemidefinit mátrix felbonthatósága.

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

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

$$A_2 = A - \frac{Ae_1 e_1^T A}{e_1^T A e_1}, \tag{4.1}$$

mátrixot, amelyről tudjuk, hogy az első sora és oszlopa zérus. Válasszuk L első oszlopának $Le_1 = Ae_1 / \sqrt{a_{11}}$ -et, ezzel $A_2 = A - Le_1 e_1^T L$.

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

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

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

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

5.1.4 Példa Choleski-felbontásra

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

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

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

5.1.5 Feladatok.

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

5.2. Főátló-dominancia

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

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

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

5.2.1 Tétel, a főátló-dominancia megmaradása.

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

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

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

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

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

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

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

5.2.2 Feladatok. Mutassuk meg:

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

5.3. Két- és háromátlójú mátrixok

5.3.1 Speciális mátrixok

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

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

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

$$T=\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} \tag{4.2}$$

mátrix inverze:

$$T^{-1}=\left(K+K^T\right)^{-1}=\left[K\left(S+S^T\right)K^T\right]^{-1}=K^{-T}\left(I+ee^T\right)^{-1}K^{-1}=K^{-T}\left(I-\frac{ee^T}{1+n}\right)K^{-1}, \tag{4.3}$$

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

5.3.2 Főátló-domináns háromátlós mátrix

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

$$Hx=\begin{bmatrix} d_1 & c_1 & & \\ a_1 & d_2 & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ & & a_{n-1} & d_n \end{bmatrix}x=\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}. \tag{4.4}$$

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

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

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

$$\begin{aligned} &d_1'=d_1;\ d_i'=d_i-a_{i-1}c_{i-1}/d_{i-1}', \quad i=2,3,\dots,n, \\ &b_1'=b_1;\ b_i'=b_i-a_{i-1}b_{i-1}'/d_{i-1}', \quad i=2,3,\dots,n. \end{aligned} \tag{4.5}$$

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

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

Látjuk: az *L* mátrix nem is kell a megoldáshoz, másrészt (4.5) mindkét sorában szerepel a_{i-1}/d_{i-1}', amit elegendő egyszer előállítani. Ezzel a megoldási algoritmus:

Kezdés: $d'_1 = d_1, \quad b'_1 = b_1.$
$i = 2,3,\ldots,n$-re
$s := a_{i-1} / d'_{i-1}; \quad d'_i := d_i - c_{i-1} * s; \qquad b'_i := b_i - b'_{i-1} * s.$
$x_n := b'_n / d'_n;$
$i = n-1, n-2,\ldots,1$-re
$x_i := (b'_i - c_i * x_{i+1}) / d'_i.$

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

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

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

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

innen

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

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

Kezdés: $f_1 = b_1 / d_1, \quad g_1 := c_1 / d_1.$
$i = 2,3,\ldots,n$-re
$s := d_i - a_{i-1} g_{i-1}; \quad f_i := (b_i - a_{i-1} f_{i-1}) / s; \qquad g_i := c_i / s.$
$x_n := f_n;$
$i = n-1, n-2,\ldots,1$-re
$x_i := f_i - g_i * x_{i+1}.$

5.3.3 Feladat

- Ha új jobboldal vektort kapunk, milyen részletszámításokat őrizzünk meg és mit számítsunk újra mindkét algoritmusban?
- Igazoljuk, hogy az (4.2)-ben szereplő háromátlós mátrix pozitív definit, mert van LL^T-felbontása.

6. Gram-Schmidt ortogonalizáció, *QR*-felbontás

Az egyszerű lineáris algebrai transzformációk között a harmadik fejezetben megismerkedtünk a vetítő mátrixokkal. E mátrixok alkalmasak arra, hogy a vektorok egy adott készletéből ortogonális vektorokat állítsunk elő. Ha ezen vektorokat egy mátrix oszlopaiba rendezzük, akkor a kapott módszer a mátrix egy újabb felbontását szolgáltatja, ezt hívjuk a QR-felbontásnak.

6.1. A Gram-Schmidt ortogonalizáció

Tegyük fel, van egy lineárisan független vektorokból álló halmaz: $\{a_i\}_{i=1}^k$, $a_i \in \mathbb{R}^m$. Szeretnénk e vektorok felhasználásával olyan ortogonális rendszert készíteni, amellyel a halmaz vektorai előállíthatók. Ekkor eljárhatunk a következőképpen. A készülő ortogonális vektorokat jelöljük q-val.

Az első lépésben válasszuk $q_1 = a_1$-et. A következő vektort készítsük úgy, hogy az a_2 vektort szimmetrikus - vagy más szóhasználattal - ortogonális vetítéssel ortogonalizáljuk q_1-re:

$$\left(I - \frac{q_1 q_1^T}{q_1^T q_1} \right) a_2 = q_2 . \tag{5.1}$$

Beszorzással $q_1^T q_2 = 0$. A következő vektort úgy készítjük, hogy az a_3 vektort q_1 és q_2-re ortogonalizáljuk:

$$\left(I - \frac{q_2 q_2^T}{q_2^T q_2} \right) \left(I - \frac{q_1 q_1^T}{q_1^T q_1} \right) a_3 = q_3 . \tag{5.2}$$

Ismét beszorzással ellenőrízve kapjuk, hogy $q_1 \perp q_3$ és $q_2 \perp q_3$. A továbbiakban vezessük be az i-edik ortogonális vektorhoz a

$$P_i = I - \frac{q_i q_i^T}{q_i^T q_i} \tag{5.3}$$

vetítőmátrixot. Látjuk, ha ezzel a mátrixszal bármely vektorra szorzunk, eredményül a q_i vektorra ortogonális vektorhoz jutunk.

Az $i+1$-edik ortogonális vektort a következő vetítések sorozatával kapjuk az a_{i+1} vektorból:

$$P_i P_{i-1} \ldots P_1 a_{i+1} = q_{i+1} . \tag{5.4}$$

Vegyük észre, a vetítőmátrixok a szorzatban tetszőleges sorrendben írhatók a bennük szereplő vektorok ortogonalitása miatt. Fennáll az összefüggés:

$$P_i P_{i-1} \ldots P_1 = \prod_{j=1}^{i} \left(I - \frac{q_j q_j^T}{q_j^T q_j} \right) = I - \sum_{j=1}^{i} \frac{q_j q_j^T}{q_j^T q_j} , \tag{5.5}$$

aminek az igazolását egy feladatra hagyjuk. Ez mutatja, hogy numerikusan kétféle lehetőség van az ortogonalizálásra. Az egyik, amikor a fenti összefüggésben a szummás alakot használjuk. Ekkor minden q_j vektor az a_{i+1} vektorral szorzódik és (5.4), (5.5) összevetéséből kapjuk:

$$q_{i+1} = a_{i+1} - \sum_{j=1}^{i} \frac{q_j q_j^T a_{i+1}}{q_j^T q_j}, \quad \rightarrow \quad a_{i+1} = q_{i+1} + \sum_{j=1}^{i} \frac{q_j q_j^T a_{i+1}}{q_j^T q_j} , \tag{5.6}$$

azaz minden a_{i+1} vektor az ortogonális vektorok segítségével előállítható, ahol a kifejtési együtthatók

$$r_{j,i+1} = \frac{q_j^T a_{i+1}}{q_j^T q_j}. \tag{5.7}$$

Ha (5.4)-ben a mátrixszorzatot alkalmazzuk, akkor a következő vektor-sorozatot készítjük:

$$z_1 = a_{i+1},\ \ z_2 = P_1 z_1,\ \ \dots\ ,\ z_{j+1} = P_j z_j,\ \ q_{i+1} = z_{i+1}.$$

A vetítőmátrixok kiírásával ekkor az

$$r_{j,i+1} = \frac{q_j^T z_j}{q_j^T q_j} \tag{5.8}$$

előállításra jutunk.

A szummás alakot nevezzük a *klasszikus Gram-Schmidt (G-S) ortogonalizációnak,* a szorzatmátrixos változatot pedig *módosított Gram-Schmidt ortogonalizációnak.* Åke Björck a numerikus tulajdonságok vizsgálata során kimutatta, hogy a módosított G-S módszer jobb hibatulajdonságokkal rendelkezik. Újabb eredmények szerint mindkét módszer egyformán jó, ha minden ortogonalizációs lépést kétszer hatjunk végre. Ekkor a kapott normált vektorok ortogonalitása közel gépi pontossággal teljesül.

6.1.1 Feladatok

- Igazoljuk a (5.5) formulát!
- Mutassuk meg, hogy a (5.7) és (5.8) formulával adott $r_{j,i+1}$ megegyezik!
- Gyűjtsük az ortogonális vektorokat a $Q = [q_1 q_2 \dots q_i]$ mátrixba. Vezessük le, hogy $P_i P_{i-1} \dots P_1 = I - Q\left(Q^T Q\right)^{-1} Q^T$.
- Legyen $A \in \mathbb{R}^{m \times n}$, ahol A oszlopai lineárisan függetlenek. Ellenőrízzük, hogy $I - A\left(A^T A\right)^{-1} A^T$ szintén vetítőmátrix és egy vektorra alkalmazva az eredmény olyan vektor lesz, amely A összes oszlopára ortogonális.

6.2. Tétel, QR-felbontás

Legyen $A \in \mathbb{R}^{m \times n}$, ahol A oszlopai lineárisan függetlenek. Ekkor A mindig felírható

$$A = QR \tag{5.9}$$

alakban, ahol Q oszlopai egymásra ortogonális vektorok és R felső háromszög mátrix. Q és R oszlopai az elsővel kezdve rekurzívan felépíthetők.

Bizonyítás. Szükséges $n \le m$, különben nem lehetnének A oszlopai lineárisan függetlenek. Fogjuk fel az A mátrixot úgy, mint ami az $a_1, a_2, \dots, a_n$ oszlopvektorokból van összeállítva és alkalmazzuk az előző szakaszban megismert G-S ortogonalizációt! Ekkor (5.6) és (5.7) összevetéséből kapjuk:

$$a_{i+1} = \sum_{j=1}^{i+1} q_j r_{j,i+1},\ \ \text{ahol}\ \ r_{i+1,i+1} = 1.$$

Az $A = [a_1, a_2, \dots, a_n],\ \ Q = [q_1, q_2, \dots, q_n]$ mátrixokkal ez pedig nem más, mint a (5.9) előállítás, ahol $R = [r_{ij}]$. A G-S ortogonalizációban az $r_{ij},\ i > j$ elemek nem voltak definiálva. Nincs is rájuk szükség, így ezeket az elemeket zérusnak választva (5.9) pontosan teljesül. ■

6.2.1 Feladatok

- A G-S ortogonalizációnál kidolgozható az a változat, amikor a q_j vektorok normáltak, $\|q_j\|_2 = 1$. Írjuk át a formulákat erre az esetre!
- Legyen $D = Q^T Q$, ezzel $\tilde{Q} = QD^{-1/2}$ oszlopvektorai normáltak. (5.9)-ben legyen $A = \tilde{Q}\tilde{R}$. Adjuk meg $\tilde{R}$-et mátrixos alakban és a normált ortogonális vektorok segítségével fejezzük ki $\tilde{r}_{ij}$-t!
- A 3.12 gyakorlatban láttuk, hogy minden x vektor egy σe_1 vektorba vihető tükrözéssel, ahol $|\sigma| = \|x\|_2$. Ha egy ilyen tükrözést alkalmazunk A első oszlopára, akkor a QR-felbontás az első oszlopra megvalósult: $R_1 = I - 2v_1 v_1^T / v_1^T v_1$ ortogonális mátrix, ahol $v = x - \sigma e_1$ és $A = R_1(R_1 A)$. Hogyan folytassuk a tükrözéseket, hogy egy QR-felbontáshoz jussunk?
- Ha rendelkezésünkre áll A egy QR-felbontása, hogyan oldjunk meg egy $Ax = b$ egyenletrendszert?

6.3. Példa QR-felbontásra

Elkészítjük az

$$A = \begin{bmatrix} 1 & 1 & -1 \\ 2 & 1 & 3 \\ 1 & 1 & -2 \\ 2 & 1 & 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}$$

mátrixra a QR-felbontás nemnormált változatát. Induláskor $q_1 = a_1$ és $q_1^T q_1 = 10$. A következő vektorhoz $q_1^T a_2 = 6$ és

$$q_2 = a_2 - q_1 \frac{q_1^T a_2}{q_1^T q_1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - \frac{6}{10}\begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 2 \\ -1 \\ 2 \\ -1 \end{bmatrix}.$$

$q_2^T q_2 = \frac{10}{25} = \frac{2}{5}$, ezt az eredményt előállíthatjuk a kapott q_2 vektorból, de úgy is számíthatjuk, hogy észrevesszük: $q_2^T q_2 = \left(a_2^T - \frac{q_1^T a_2}{q_1^T q_1} q_1^T\right)\left(a_2 - q_1 \frac{q_1^T a_2}{q_1^T q_1}\right) = a_2^T a_2 - \frac{\left(q_1^T a_2\right)^2}{q_1^T q_1} = 4 - \frac{36}{10} = \frac{2}{5}$. A harmadik vektor előállításához $q_1^T a_3 = 5$ és $q_2^T a_3 = -2$, ezekkel a harmadik vektor és a QR-felbontás:

$$q_3 = a_3 - q_1 \frac{q_1^T a_3}{q_1^T q_1} - q_2 \frac{q_2^T a_3}{q_2^T q_2} = \begin{bmatrix} -1 \\ 3 \\ -2 \\ 1 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \end{bmatrix} + \frac{2 \cdot 5}{2 \cdot 5}\begin{bmatrix} 2 \\ -1 \\ 2 \\ -1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 \\ 2 \\ -1 \\ -2 \end{bmatrix},$$

$$A = \begin{pmatrix} 1 & 2/5 & 1/2 \\ 2 & -1/5 & 1 \\ 1 & 2/5 & -1/2 \\ 2 & -1/5 & -1 \end{pmatrix}\begin{pmatrix} 1 & 3/5 & 1/2 \\ 0 & 1 & -5 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & 2 & -1 \\ 2 & -1 & -2 \end{pmatrix}\begin{pmatrix} 1 & 3/5 & 1/2 \\ 0 & 1/5 & -1 \\ 0 & 0 & 1/2 \end{pmatrix}.$$

6.3.1 Feladat

Készítsük el a következő mátrix QR-felbontását: $\begin{bmatrix} 2 & 6 & 5 \\ -1 & -4 & 1 \\ -1 & -2 & -3 \end{bmatrix}$.

6.4. Az Arnoldi-módszer

Ez a *Krilov-bázis* vektorainak G-S ortogonalizációja. A *Krilov-bázis* vektorai $x, Ax, A^2x, \ldots$, ahol $x \neq 0$, egyébként tetszőleges induló vektor. Az Arnoldi-módszernél ennek alapján $q_1 = x / \|x\|_2$, a következő vektor Aq_1 és ezt q_1-re q_1-re ortogonalizálva kapjuk q_2-t. Általában q_{i+1}-et úgy kapjuk, hogy az Aq_i vektort ortogonalizáljuk a meglévő q_j vektorokra és az eredményt normáljuk. Belátható, hogy az így nyert q_j vektorok ugyanazt az alteret feszítik ki, mint a Krilov-bázis vektorai. A módszerrel a következő QR-felbontáshoz jutunk:

$$\begin{bmatrix} q_1 & Aq_1 & Aq_2 & \ldots & Aq_i \end{bmatrix} = \begin{bmatrix} q_1 & q_2 & \ldots & q_{i+1} \end{bmatrix} R, \tag{5.10}$$

ahol R felső háromszögmátrix. Ebből a sémából szokás elhagyni bal oldalon az első vektort, q_1-et. Ez azt jelenti, hogy a jobb oldalon R első oszlopát is elhagyjuk. Sőt, hogy R helyén négyzetes mátrix maradjon, az utolsó sorát is elhagyjuk. Jelöljük a maradék mátrixot H-val. Ekkor a q_j vektorokból mátrixot képezve a

$$Q = \begin{bmatrix} q_1 & q_2 & \ldots & q_i \end{bmatrix}, \quad AQ = QH + h_{i+1,i} q_{i+1} e_i^T \tag{5.11}$$

összefüggésre jutunk, ahol H un. *felső Hessenberg-féle* mátrix. E mátrixok közeliek a felső háromszögmátrixokhoz, azzal a különbséggel, hogy a főátló alatti elemek sem zérusok. A rendszert jobbról az e_i vektorral szorozva kapunk rekurziót q_{i+1} számítására:

$$Aq_i = \sum_{j=1}^{i} h_{ji} q_j + h_{i+1,i} q_{i+1}, \quad h_{ji} = q_j^T A q_i, \tag{5.12}$$

a $h_{i+1,i}$ elemet abból a feltételből határozhatjuk meg, hogy q_{i+1} normált. Ha $h_{i+1,i} = 0$, akkor a rekurzió megáll és $i < n$ esetén Q oszlopai A egy invariáns alterét feszítik ki.

6.4.1 Feladat

1. Legyen az x kezdővektor A három sajátvektorának az összege. Hány lépés után áll le az Arnoldi-módszer?

7. Az algebrai sajátértékfeladat

Eszerint keresendő egy (λ, y, x) hármas, amelyre teljesül

$$Ax = \lambda x, \qquad y^T A = \lambda y^T, \tag{6.1}$$

ahol λ az $A \in \mathbb{R}^{n\times n}$ mátrix *sajátértéke*, x a *jobboldali* és y a *baloldali sajátvektor*. A sajátértékek a $|\lambda I - A|$ *karakterisztikus polinom* gyökei és a sajátértékhelyeken $\lambda I - A$ szinguláris. A determináns alakból látható, hogy a mátrix hasonlósági transzformáltjának karaketerisztikus polinomja ugyanaz: $|\lambda I - S^{-1}AS| = |S^{-1}(\lambda I - A)S| = |S^{-1}||\lambda I - A||S| = |\lambda I - A|$, tehát a hasonlósági transzformáció a sajátértékeket helyben hagyja.

A mátrixot általában valósnak tekintjük. De mivel valós mátrixnak is lehetnek komplex sajátértékei és sajátvektorai, emiatt sokszor a komplex esetre is gondolni kell.

7.1. Néhány tulajdonság

Az alábbiakban felidézünk néhány a sajátértékfeladattal kapcsolatos ismeretet.

7.1.1 Legalább 1 sajátpár létezése

Minden λ_i sajátértékhez tartozik legalább egy jobb- és baloldali sajátvektor.

Mert $\lambda_i I - A$ és $\lambda_i I - A^T$ magtere legalább 1-dimenziós (ui. nemcsak a null-vektorból áll). ■

7.1.2 Lineáris függetlenség

Különböző sajátértékekhez tartozó sajátvektorok lineárisan függetlenek.

Ennek a bizonyítása indirekt módon történhet. Feltesszük két különböző sajátértékhez tartozó sajátvektorról, hogy lineárisan összefüggők. Ekkor ellentmondásra jutunk, mert két vektor úgy lehet linerárisan összefüggő, hogy egyirányúak, ekkor viszont nem lehetnek a sajátértékek különbözők. ■

7.1.3 Különböző sajáértékhez tartozó bal és jobb sajátvektorok ortogonalitása

Legyen v_i a λ_i sajátértékhez tartozó bal sajátvektor, u_j pedig a λ_j sajátértékhez tartozó jobb sajátvektor, $i \neq j$. Ekkor $v_i^T u_j = 0$.

Bizonyítás. Tekintsük a következő kifejezést: $v_i^T A u_j = \lambda_i v_i^T u_j = \lambda_j v_i^T u_j$, ahol az egyik esetben a bal, a másik esetben pedig a jobb sajátvektor tulajdonságot alkalmaztuk. Kapjuk, hogy $(\lambda_i - \lambda_j) v_i^T u_j = 0$, s ebből következik az állítás, mert $\lambda_i \neq \lambda_j$. ■

7.1.4 Következmény

Ha minden sajátérték különböző, akkor a sajátvektorokat egy $X = [x_1 x_2 \dots x_n]$ és $Y = [y_1 y_2 \dots y_n]$ mátrixba rendezve kapjuk:

$$AX = X\Lambda, \quad Y^T A = \Lambda Y^T. \tag{6.2}$$

A lineáris függetlenség miatt X és Y invertálhatók, így $X^{-1}AX = \Lambda = Y^T A Y^{-T}$, ahol a transzponált inverzét jelöltük $-T$-vel. Írhatjuk: $Y^T = D^{-1}X^{-1}$, ahol D egy nemszinguláris diagonálmátrix és

szerepe csupán annyi, hogy az y_i vektorok hosszát skálázza. Tehát az általánosság megszorítása nélkül vehetjük: $Y^T = X^{-1}$, a sajátvektorok saját-altereket adnak, a vektorok hossza tetszőleges. A kapott alak mutatja, hogy ekkor a mátrix *hasonlósági transzformációval diagonalizálható*.

7.1.5 Schur tétele

Minden négyzetes mátrix unitér hasonlósági transzformációval felső háromszög alakra hozható.

Bizonyítás. Jelölje $R(u) = I - 2uu^H / u^H u$ a Householder tükröző mátrixot. Legyen $x \neq e_1$ egy normált sajátvektor, $\|x\|_2 = 1$, amelyet skálázzunk úgy, hogy az első eleme valós, nempozitív szám legyen. (Ezt mindig elérhetjük, ha a vektort a nemzérus $-\overline{x}_1 / |x_1|$ számmal beszorozzuk.) Ekkor $R(x - e_1)e_1 = x$ és $R(x - e_1)x = e_1$. Feltéve, hogy $Ax = \lambda x$ teljesül,

$$R(x - e_1)AR(x - e_1)e_1 = R(x - e_1)Ax = R(x - e_1)\lambda x = \lambda e_1.$$

Mivel $R(x - e_1)$ involutórius (azaz megegyezik az inverzével), hasonlósági transzformációt végeztünk, ahol az első oszlopvektor az e_1 vektor λ-szorosába ment át, amivel a felső háromszög alak az első sorban és oszlopban előállt. Az eljárást folytatva az eggyel kisebb méretű jobb alsó blokkokra, végül a kívánt alakra jutunk. ■

Megjegyzés. Ha $x = e_1$, akkor A első oszlopa már λe_1 alakú. A felső háromszög-alakra hozás egy lépése egyben *deflációs módszer*, mert olyan 1-gyel kisebb méretű mátrixot kapunk a jobb alsó sarokban, amelynek a sajátértékei a megtalált λ-t kivéve megegyeznek a kiinduló mátrixéval.

A Schur-tétel segítségével további fontos tulajdonságok láthatók be.

7.1.6 Tétel, diagonalizálhatóság unitér hasonlósági transzformációval

Az A mátrix normális $\Leftrightarrow$ A unitér hasonlósági transzformációval diagonalizálható.

Bizonyítás. $A \in \mathbb{C}^{n\times n}$ normális, ha teljesül $AA^H = A^H A$, ^{H} a transzponált konjugáltat jelöli.

$\Leftarrow$: Tfh Legyen $A = U\Lambda U^H$, innen $AA^H = U\Lambda U^H U\overline{\Lambda}U^H = U\Lambda\overline{\Lambda}U^H = U\overline{\Lambda}U^H U\Lambda U^H = A^H A$.

$\Rightarrow$: Ha A normális, akkor bármely unitér hasonlósági transzformáltja is az. Legyen a Schur-tétel alapján $B = U^H AU$ felső háromszögmátrix, ekkor $BB^H = B^H B$. Ennek az 1,1-indexű eleme:

$$e_1^T BB^H e_1 = \left\|B^H e_1\right\|_2^2 = \sum_{j=1}^{n} \left|b_{1j}\right|^2 = e_1^T B^H Be_1 = \left\|Be_1\right\|_2^2 = \left|b_{11}\right|^2,$$

azaz B első sorának kettes normája megegyezik az első oszlopéval. Ez csak úgy lehetséges, ha $b_{1j} = 0,\ j = 2,\ldots,n$. Az eljárást folytatva az eggyel kisebb méretű jobb alsó blokkal, minden sorra azt kapjuk, hogy csak főátlóbeli elem lehet nemzérus. ■

A tétel következménye, hogy a valós szimmetrikus mátrixok ortogonális, az hermitikus mátrixok pedig unitér hasonlósági transzformációval diagonalizálhatók. E mátrixok sajátértékei mindig valósak.

7.1.7 Tétel

Az egyszeres sajátértékhez tartozó y bal- és x jobboldali sajátvektorok skaláris szorzata nemzérus: $y^H x \neq 0$.

Hozzuk a mátrixot unitér hasonlósági transzformációval felső háromszög alakra: $B = U^H AU$. Ekkor a sajátvektorok átmennek az $y \to U^H y$ és $x \to U^H x$ vektorokba, ahonnan látható, hogy skaláris szorzatuk nem változik meg. Az általánosság megszorítása nélkül feltehetjük, hogy

$$B = \begin{bmatrix} \lambda & b^T \\ 0 & C \end{bmatrix}$$

alakú, ahol λ az x és y vektorhoz tartozó sajátérték. A transzformálás után x átment e_1-be, a bal oldali vektor pedig legyen $y^H U = \begin{bmatrix} \overline{\eta} & y_2^H \end{bmatrix}$. Ha ezzel balról szorozzuk B-t, akkor

$$\begin{bmatrix} \overline{\eta} & y_2^H \end{bmatrix} \begin{bmatrix} \lambda & b^T \\ 0 & C \end{bmatrix} = \begin{bmatrix} \overline{\eta}\lambda & \overline{\eta} b^T + y_2^H C \end{bmatrix} = \lambda \begin{bmatrix} \overline{\eta} & y_2^H \end{bmatrix},$$

ahonnan $\overline{\eta} b^T + y_2^H C = \lambda y_2^H$. Ha itt $\overline{\eta} = y^H x$ zérus volna, akkor a jobb alsó C blokknak λ sajátértéke volna. Ez ellentmondana annak, hogy λ egyszeres sajátérték. ■

7.1.8 Jordan-blokkok

A

$$J(\mu) = \begin{bmatrix} \mu & 1 & & \\ & \mu & \ddots & \\ & & \ddots & 1 \\ & & & \mu \end{bmatrix} \in \mathbb{C}^{k \times k} \tag{6.3}$$

alakú mátrixot *Jordan-blokknak* nevezzük. Ez a hasonlósági transzformációval nem diagonalizálható mátrixok prototípusa. Ránézésre látható, hogy a karakterisztikus polinomja $|\lambda I - J(\mu)| = (\lambda - \mu)^k$, tehát $\lambda = \mu$ k-szoros gyök. A sajátértékhelyen a karakterisztikus mátrix rangvesztése csak 1, mert a bal alsó sarokelemhez tartozó aldeterminánsban éppen az átló feletti -1-esek szorzata lesz, így a rang csak eggyel csökken. Következésképp 1 jobb és baloldali sajátvektor van, e_1 és e_k^T, amelyek skaláris szorzata 0, ha $k > 1$.

A sajátérték karakterisztikus polinombeli multiplicitását *algebrai multiplicitásnak,* m_A nevezzük. A sajátértékhez tartozó sajátvektorok által kifeszített altér dimenziója pedig a sajátérték *geometriai multiplicitása,* m_G. A fenti Jordan-blokknál $m_A = k$ és $m_G = 1$.

Bizonyítás nélkül megemlítjük, hogy általában a mátrixok hasonlósági transzformációval *Jordan-féle kanonikus alakra* hozhatók, ahol minden sajátértékhez egy vagy több Jordan-blokk tartozik, amelyek a főátló mentén helyezkednek el. Lehetséges 1×1-es Jordan-blokk, az egyszeres sajátértékeknek például ez van. Könnyen belátható, hogy a sajátértékhelyen a karakterisztikus mátrix rangvesztése egyenlő m_G-vel, ami a sajátértékhez tartozó Jordan-blokkok száma, a sajátérték algebrai multiplicitása m_A viszont egyenlő a hozzátartozó Jordan-blokkokban lévő átlóelemek számával.

7.1.9 Feladatok

1. Legyen A olyan felső Hessenberg mátrix, amelynek minden átló alatti eleme nemzérus. Mutassuk meg, hogy ennek a mátrixnak minden sajátértékéhez csak 1 Jordan-blokk tartozhat.
2. Mutassuk meg, ha A sajátértékei a λ_i-k, akkor A^{-1} sajátértékei $1/\lambda_i$-k.

7.2. A sajátértékek lokalizációja

Mivel még valós mátrixoknak is lehetnek komplex sajátértékei, emiatt a komplex síkon kell megadni olyan tartományokat, ahol a sajátértékek lehetnek. Egy ilyen becsléssel már megismerkedtünk a normákkal foglalkozó 2.8 szakaszban, mely szerint a spektrál sugár nem nagyobb, mint a mátrix valamely indukált normája. Így egyik sajátérték sem lehet nagyobb abszolút értékben, mint például $\|A\|_1$ vagy $\|A\|_\infty$. Ennél pontosabb becslést tesz lehetővé

7.2.1 Gersgorin tétele

Legyen az i-edik K_i Gersgorin-kör középpontja a_{ii}, sugara pedig $r_i = \left\| e_i^T (A - a_{ii} I) \right\|_\infty$, ami nem más, mint a mátrixban az i-edik sorelemek abszolút összege a diagonálelem kivételével. A tétel szerint az A mátrix sajátértékei a Gersgorin-körök egyesített halmazában vannak.

Bizonyítás. Tekintsük az $Ax = \lambda x$ egyenlet i-edik sorát, ahol x sajátvektor, λ a hozzátartozó sajátérték, és $|x_i| = \|x\|_\infty$. Kissé átrendezve: $\lambda - a_{ii} = \sum_{j=1, j\neq i}^{n} \frac{a_{ij} x_j}{x_i}$, ahonnan $|\lambda - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} \left| \frac{a_{ij} x_j}{x_i} \right| \leq$

$\leq \sum_{j=1, j\neq i}^{n} |a_{ij}| = r_i$. Minden sajátértékre felírhatunk egy ilyen összefüggést, ez adja az állítást. ■

7.2.2 Második tétel

Ha a Gersgorin köröknek vannak diszjunkt részhalmazai, akkor minden ilyen részhalmazban annyi sajátérték található, amennyi a hozzá tartozó Gersgorin körök száma.

Bizonyítás. Fel kell használnunk azt az itt nem bizonyított eredményt, hogy a mátrix sajátértékei a mátrixelemek folytonos függvényei. Bontsuk a mátrixot két részre, és képezzük az $A(\varepsilon) = D + \varepsilon A_1$ mátrixot, ahol D a főátlót tartalmazó diagonálmátrix, A_1 pedig a nemdiagonális rész. Ha most $\varepsilon = 0$, akkor minden kör sugara zérus. Ha ε 1-hez tart, akkor minden sajátérték kifuthat a középpontból, de a folytonosság miatt nem ugorhat át egy másik diszjunkt körhalmazba. ■

7.2.3 Példa

A Gersgorin-tétel alkalmazását kombinálhatjuk diagonálmátrixszal készített hasonlósági transzformációval. Ezzel változtatni tudjuk a körök sugarát, és egyszerűen készíthetünk a célnak megfelelő becslést. Például mutassuk meg, hogy a

$$A = \begin{bmatrix} 8 & 5 & 3 \\ 1 & 4 & 1 \\ 1 & 2 & 5 \end{bmatrix}$$

mátrixnak nincs zérus sajátértéke!

Az első Gersgorin kör középpontja 8, sugara szintén 8, így ez a kör tartalmazza a zérust. A többi kör nem. Alkalmazzuk a $D^{-1}AD$ hasonlósági transzformációt, ahol $D = \mathrm{diag}(2 \quad 1 \quad 1)$:

$$D^{-1}AD = \begin{bmatrix} 8 & 5/2 & 3/2 \\ 2 & 4 & 1 \\ 2 & 2 & 5 \end{bmatrix}$$

ezzel a kívánt célt elértük, mert az első kör sugara 4-re csökkent és a másik két kör továbbra sem tartalmazza a zérust. Figyeljük meg, milyen sor és oszlopban lesz változás, ha az alkalmazott diagonálmátrixban csak egy elem különbözik 1-től!

7.2.4 Feladatok

Bizonyítsuk be:

1. A mátrix főátló-domináns, ha a Gersgorin körök nem tartalmazzák a zérust.

2. A főátló-domináns mátrixok invertálhatók, mert nincs zérus sajátértékük.

3. Az i- és j-edik sorok és oszlopok cseréje a diagonál-dominanciát nem változtatja meg.

4. A mátrix rangja legalább akkora, mint azon Gersgorin körök száma, amelyek nem tartalmazzák a zérust.

5. A baloldali sajátvektorok segítségével is készíthetünk Gersgorin-köröket a mátrix oszlopai szerint.

6. Döntsük el Gersgorin tétele és diagonálmátrix hasonlósági transzformáció segítségével, hogy A invertálható-e: $A=\begin{bmatrix}7 & 6 & -3\\ 1 & 5 & 1\\ 4 & -2 & 6\end{bmatrix}$.

7.3. A karakterisztikus polinom számítása

Tekintsük az un. *Frobenius-féle kisérő mátrixot*:

$$F=\begin{bmatrix}0 & 0 & \dots & 0 & -a_0\\ 1 & 0 & \dots & 0 & -a_1\\ & 1 & \ddots & \vdots & \vdots\\ & & \ddots & 0 & -a_{n-2}\\ & & & 1 & -a_{n-1}\end{bmatrix}. \tag{6.4}$$

Az utolsó oszlopa mentén kifejtve igazolható, hogy $\det(\lambda I-F)=\lambda^n+\sum_{i=0}^{n-1}a_i\lambda^i$. Eszerint a karakterisztikus polinom együtthatóinak előállítása könnyű, ha van valamilyen hasonlósági transzformáció, ami az A mátrixot ilyen alakra hozza. Danyiljevszkij ötelete szerint ez megvalósítható a Gauss-Jordan módszernél megismert egyszerű transzformációs mátrixszal, ha a mátrix első oszlopát nem e_1-be, hanem e_2-be visszük. Legyen tehát az első transzformációs mátrix $T_1=I+(Ae_1-e_2)e_2^T$, $A_2=T_1^{-1}AT_1$ és ekkor a hasonlósági transzformáció eredményeként az első oszlop e_2 lesz:

$$A_2e_1=T_1^{-1}AT_1e_1=\left(I-\frac{(Ae_1-e_2)e_2^T}{e_2^TAe_1}\right)A\left(I+(Ae_1-e_2)e_2^T\right)e_1=\left(I-\frac{(Ae_1-e_2)e_2^T}{e_2^TAe_1}\right)Ae_1=e_2.$$

Általában a k-adik lépésben $T_k=I+(A_ke_k-e_{k+1})e_{k+1}^T$, és a korábbi oszlopvektorok sem romlanak el, mert az előbbihez hasonlóan kapjuk:

$$\left(I-\frac{(A_ke_k-e_{k+1})e_{k+1}^T}{e_{k+1}^TA_ke_k}\right)A_k\left(I+(A_ke_k-e_{k+1})e_{k+1}^T\right)e_l=e_{l+1},\quad l\le k.$$

Egy transzformációs lépés végrehajthatóságának feltétele, hogy a diagonálelem alatti elem legyen zérustól különböző. Ha nem így volna, sor-cserével mozgassunk egy átló alatti nemzérus elemet az átlóelem alá és hajtsuk végre a hasonló oszlopok cseréjét is a hasonlósági transzformáció megőrzése végett. Az algoritmus $n-1$-edik lépésében a (6.4) alakhoz jutunk.

Ha a mátrix háromátlójú, a karakterisztikus polinom egyszerű rekurzióval számolható. Például a

$$\begin{vmatrix}\lambda-d_1 & -c_1 & & \\ -a_1 & \lambda-d_2 & \ddots & \\ & \ddots & \ddots & -c_{n-1}\\ & & -a_{n-1} & \lambda-d_n\end{vmatrix}$$

determináns rekurziója

$$p_{i+1}(\lambda)=(\lambda-d_{i+1})p_i(\lambda)-a_ic_ip_{i-1}(\lambda),\ \ p_0=1,\ \ p_1=\lambda-d_1, \tag{6.5}$$

ahol $p_i(\lambda)$ a bal felső i-edrendű blokk determinánsa. Az $i+1$-edrendű determinánst az $i+1$-edik oszlop szerinti kifejtéssel kapjuk és az eredmény a kapott rekurzió. A rekurzióval a polinom helyettesítési értékét is könnyen számolhatjuk.

A rekurziót meg lehet csinálni a felső Hessenberg-mátrix karakterisztikus polinomjára is. Legyen például $p_3 = 1$ és alulról felfelé oldjuk meg a következő egyenletrendszert:

$$\begin{bmatrix} \lambda-2 & 1 & 3 \\ 2 & \lambda+1 & 2 \\ 0 & 2 & \lambda-1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} p(\lambda) \\ 0 \\ 0 \end{bmatrix}.$$

Az utolsó sorból $p_2(\lambda) = (1-\lambda)/2$, a második sorból pedig $2p_1(\lambda) + (\lambda+1)p_2(\lambda) + 2 = 0$, ahonnan p_1 kifejezhető. Ezekkel az első sor adja $p(\lambda)$ kifejezését. A mátrix determinánsa akkor lesz zérus, ha $p(\lambda)$ zérus, tehát $p(\lambda)$ gyökei megegyeznek a mátrix sajátértékeivel. Figyeljük meg, $p(\lambda) = 0$ esetén $p_1(\lambda), p_2(\lambda), p_3(\lambda)$ a λ sajátértékhez tartozó sajátvektor elemei.

7.3.1 Feladat

Igazoljuk, hogy a 2×2-es A mátrix sajátértékei: $\lambda_{1,2} = \frac{a_{11}+a_{22}}{2} \pm \sqrt{\left(\frac{a_{11}-a_{22}}{2}\right)^2 + a_{12}a_{21}}$.

7.4. Tétel

Minden mátrix unitér hasonlósági transzformációval felső Hessenberg-alakra hozható.

Bizonyítás. Az első lépésben legyen $u_1 = (A - e_1 e_1^T A)e_1$, $\|u_1\|_2 = |\sigma_1|$, tehát A első oszlopából az átlóelemet elhagyjuk. A hasonlósági transzformációt az $R(u_1 - \sigma_1 e_2)$ tükröző mátrixszal végezzük, ahol a kivonási jegyveszteség elkerülése érdekében σ_1 előjelét aszerint választjuk, hogy u_1/σ_1 második elemének valós része legyen negatív. Ekkor $R(u_1 - \sigma_1 e_2)A$ az első oszlopot $a_{11}e_1 + \sigma_1 e_2$-be viszi, (az első elem változatlan az első oszlopvektor második elemmel kezdődő részét pedig $\sigma_1 e_2$-be tükröztük). Ugyanezzel a tükröző mátrixszal jobbról szorozva az első oszlop már nem fog változni, mert $R(u_1 - \sigma_1 e_2)$ első sora és oszlopa e_1^T és e_1. Ezzel $A_2 = R(u_1 - \sigma_1 e_2)AR(u_1 - \sigma_1 e_2)$ első oszlopa mutatja a Hessenberg-alakot. A következő lépésben az imént látottakat alkalmazzuk A_2 eggyel kisebb méretű jobb alsó blokkjára. Az eljárást folytatva végül a kívánt teljes Hessenberg-alakra jutunk. ■

7.5. Iterációs módszerek

7.5.1 A hatványiteráció

A módszer azon az észrevételen alapul, hogy k növekedésével $A^k x_0$-ben a legnagyobb sajátértékhez tartozó komponens fog felerősödni. A konvergenciára kimondhatjuk a következő tételt:

Tegyük fel A n-edrendű valós vagy komplex mátrix és a sajátértékeire teljesül

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|.$$

Továbbá a mátrix egyszerű struktúrájú, azaz annyi sajátvektora van, mint a mátrix rendje. Ekkor a spektrálfelbontás $A = \sum_{i=1}^{n} \lambda_k u_k v_k^T$, ahol v_k, u_k a bal és jobb sajátvektorok és kifejthető a sajátvektorok szerint: $x_0 = \sum_{k=1}^{n} \alpha_k u_k$. Ekkor

$$\lim_{m\to\infty}\frac{1}{\lambda_1^m}A^m x_0 = \alpha_1 u_1 \tag{6.6}$$

Bizonyítás. A módszer szerint képezzük az $x_m = Ax_{m-1} = \sum_{k=1}^{n}\alpha_k\lambda_k^m u_k$ vektorokat és innen kapjuk $\frac{A^m x_0}{\lambda_1^m} = \sum_{k=1}^{n}\alpha_k\left(\frac{\lambda_k}{\lambda_1}\right)^m u_k$. Az $m\to\infty$ határátmenettel kapjuk az állítást, mivel a többi sajátvektor szorzója zérushoz tart. ■

Látjuk, a konvergencia gyorsaságát $\alpha_1 \neq 0$ esetén lényegében a λ_2/λ_1 hányados szabja meg. Az algoritmus:

$$m = 1,2,\ldots - re:$$
$$y_{m+1} = Ax_m,$$
$$x_{m+1} = \frac{y_{m+1}}{\|y_{m+1}\|}.$$

A normának célszerű például a végtelen normát választani. A sajátérték a

$$\lambda_1^{(m)} = \frac{x_m^T y_{m+1}}{x_m^T x_m} = \frac{x_m^T A x_m}{x_m^T x_m}$$

kifejezéssel becsülhető. A hatványmódszerrel a spektrum (= a mátrix sajátértékeinek összessége) szélein lévő egyszeres sajátértékeket kereshetjük sikerrel. Az *inverz hatányiterációval* azonban kereshetjük a spektrum belsejében elhelyezkedő sajátértékeket is. Ekkor az iteráció egy lépésében az

$$x_{m+1} = (\lambda I - A)^{-1} x_m$$

vektort számítjuk. A $(\lambda I - A)^{-1}$ mátrix sajátértékei $1/(\lambda - \lambda_k)$, $k = 1,\ldots,k$, innen látható: ez is hatványiteráció, ami a λ paraméter értékéhez legközelebb eső sajátértékhez és a hozzátartozó sajátvektorhoz fog tartani.

A sajátprobléma megoldására az egyik legjobb módszer a QR-módszer. Ekkor elkészítjük az $A = Q_1 R_1$ QR-felbontást és a következő mátrix $A_2 = R_1 Q_1 = Q_1^T A Q_1$, tehát egy ortogonális hasonlósági transzformáció eredménye. A k-edik lépésben $A_k = R_{k-1}Q_{k-1}$. Megmutatható, hogy amennyiben a mátrix egyszerű struktúrájú és a sajátvektorok mátrixának van LU-felbontása, akkor a QR-módszer egy felső háromszögmátrixhoz konvergál. A konvergencia még gyorsítható, ha a felbontásokat kombináljuk egy κI mátrixszal való eltolással is, ahol κ a sajátérték egy becslése. Ekkor a QR-módszer konvergencia-sebessége másodrendű, szimmetrikus mátrixoknál harmadrendű lesz.

7.6. A sajátértékfeladattal kapcsolatos egyenlőtlenségek

A Gersgorin-körök ismertetése során már megismerkedtünk ilyen összefüggésekkel. Itt folytatjuk a vizsgálatainkat. Arra vagyunk kíváncsiak, hogyha van egy közelítő (λ, u) sajátpárunk, mit mondhatunk a jóságának jellemzésére. Egy másik feladat, hogyha a mátrixelemeket kissé megváltoztatjuk (– perturbáljuk), hogyan változik meg a sajátpár?

A következő *jelöléseket* alkalmazzuk: $M = \max_{(i)}|\lambda_i(A)|$, $m = \min_{(i)}|\lambda_i(A)|$ és feltesszük, hogy a mátrix invertálható. Mindig indukált mátrixnormát használunk.

A spektrálsugár és az indukált normák összefüggéséből már ismerjük: $M \le \|A\|$, $1/m \le \|A^{-1}\|$, a kettő összeszorzásából:

$$\frac{M}{m} \le \mathrm{cond}(A) = \|A\|\|A^{-1}\|. \tag{6.7}$$

Ez jelzi számunkra, hogyha abszolút értékben a legnagyobb és legkisebb sajátérték hányadosa nagy, akkor a mátrix kondiciószáma nagy.

7.6.1 Lemma

Legyen $D = \mathrm{diag}(d_1,\ldots,d_n)$ diagonálmátrix, ekkor $\|D\|_p = \max_{(i)} |d_i|$, $1 \le p \le \infty$.

Bizonyítás. Legyen $|d_k| \ge |d_i|$ minden i-re. Az indukált norma definiciót alkalmazva

$$\|D\|_p^p = \sup_{(x\neq 0)} \frac{\sum_{i=1}^n |d_i x_i|^p}{\sum_{i=1}^n |x_i|^p} = |d_k|^p \sup_{(x\neq 0)} \frac{\sum_{i=1}^n |x_i d_i / d_k|^p}{\sum_{i=1}^n |x_i|^p} = |d_k|^p,$$

mert a nevező nagyobb a számlálónál, ha van olyan nemzérus x_i elem, amelyre $|d_i / d_k| < 1$. ■

7.6.2 Tétel, sajátpár jósága

Legyen A egyszerű szerkezetű: $AU = U\Lambda$, ahol U a sajátvektorok mátrixa és Λ a sajátvektorokat tartalmazó diagonálmátrix, továbbá (λ, x) a sajátpár egy közelítése. Ekkor az $r = Ax - \lambda x$ jelöléssel

$$\min_{(i)} |\lambda_i - \lambda| \le \frac{\|r\|}{\|x\|} \mathrm{cond}(U), \quad \|\,.\,\| \ p\text{-norma}. \tag{6.8}$$

Bizonyítás. Ha $\lambda_i = \lambda$ valamely i-re, akkor az állítás igaz. Tegyük fel, $\lambda_i \neq \lambda$, ezzel $A - \lambda I$ invertálható: $x = (A - \lambda I)^{-1} r = U(\Lambda - \lambda I)^{-1} U^{-1} r$. A normákra áttérve és az előző lemmát alkalmazva

$$\|x\| \le \frac{\mathrm{cond}(U)}{\min_i |\lambda_i - \lambda|} \|r\|.$$

A kapott egyenlőtlenséget rendezve kapjuk a állítást. ■

7.6.3 Következmény

Hermitikus mátrixokra U unitér, emiatt $\mathrm{cond}_2(U) = 1$ és $\min_{(i)} |\lambda_i - \lambda| \le \|r\|_2 / \|x\|_2$, ami nagyon egyszerűen számolható.

7.6.4 Tétel, alsó becslés cond(U)-ra

Ha A egyszerű szerkezetű és invertálható,

$$\frac{\|A\|}{M} \le \mathrm{cond}(U), \quad \|A^{-1}\| m \le \mathrm{cond}(U), \quad \sqrt{\frac{\mathrm{cond}(A)}{\mathrm{cond}(\Lambda)}} \le \mathrm{cond}(U) \tag{6.9}$$

Bizonyítás. A harmadik összefüggés az első kettő összeszorzásával adódik. Az első egyenlőtlenség az $A = U\Lambda U^{-1}$ normáját képezve adódik, a második pedig az $A^{-1} = U\Lambda^{-1}U^{-1}$ kifejezésből. ■

7.6.5 Feladatok

Mutassuk meg, hogy

1. $\|U\| \le \|AU\| / m$.

2. $\left\|U^{-1}\right\| \le \left\|U^{-1}A^{-1}\right\| M$.

3. $\mathrm{cond}(U) \le \mathrm{cond}(AU)\,\mathrm{cond}(\Lambda)$.

7.6.6 Tétel, (Bauer, Fike)

Legyen A egyszerű szerkezetű és E egy ugyanolyan méretű mátrix. Ha μ az $A+E \in \mathbb{C}^{n\times n}$ egy sajátértéke és $AU = U\Lambda$, akkor

$$\min_{(i)} \left|\lambda_i - \mu\right| \le \left\|E\right\|_p \mathrm{cond}_p(U) . \tag{6.10}$$

Bizonyítás. Tegyük fel $\mu \notin \{\lambda_i(A)\}$, mert különben igaz az állítás. Mivel μ sajátérték, következik, hogy $U^{-1}(A+E-\mu I)U = \Lambda - \mu I + U^{-1}EU$ szinguláris, ahonnan átrendezéssel $I + \left(\Lambda - \mu I\right)^{-1} U^{-1} EU$ adódik. Ez az utóbbi mátrix pedig csak akkor lehet szinguláris, ha $\left(\Lambda - \mu I\right)^{-1} U^{-1} EU$ -nek van egy 1 abszolút értékű sajátértéke, amiből $1 \le \left\|\left(\Lambda - \mu I\right)^{-1} U^{-1} EU\right\|_p \le \left\|\left(\Lambda - \mu I\right)^{-1}\right\|_p \left\|E\right\|_p \mathrm{cond}_p(U)$. Ezt rendezve kapjuk az állítást. ■

7.6.7 Tétel, inverz perturbáció

Legyen (λ, x) egy közelítő sajátpár, $r = Ax - \lambda x$. Ekkor az $E = -\dfrac{rx^H}{\left\|x\right\|_2^2}$, $\left\|E\right\|_2 = \dfrac{\left\|r\right\|_p}{\left\|x\right\|_p}$, $p = 2, F$ (F a Frobenius-norma) mátrixszal a (λ, x) sajátpár az

$$\left(A+E\right)x = \lambda x \tag{6.11}$$

egyenlet pontos megoldása.

Bizonyítás. $\left(A - \dfrac{rx^H}{x^H x}\right)x = Ax - r = \lambda x$. ■

Például, ha $\left\|r\right\|_2 / \left\|x\right\|_2 \approx 10^{-9}$, a mátrix elemei 1 körüliek, akkor (λ, x) pontos megoldása egy mátrixnak, ami A-tól csak a kilencedik jegyében különbözik. Ha A-t csak 7 jegyre ismerjük, akkor nincs értelme tovább folytatni az iterációt.

7.6.8 Tétel, egyszeres sajátérték perturbációja

Legyen (λ, x, y) egy A-hoz tartozó sajáthármas, ahol λ egyszeres sajátérték. Az $A+E$ mátrix sajátértékének megváltozása első rendben

$$\tilde{\lambda} = \lambda + \frac{y^T Ex}{y^T x} + \mathcal{O}(\left\|E\right\|_2^2) \tag{6.12}$$

és

$$\left|\tilde{\lambda} - \lambda\right| \le \frac{\left\|y\right\|_2 \left\|x\right\|_2}{\left|y^T x\right|} \left\|E\right\|_2 + \mathcal{O}(\left\|E\right\|_2^2) . \tag{6.13}$$

Bizonyítás. A második összefüggés az első következménye, ha normákra térünk át. Az első bizonyításához legyen a sajátérték megváltozása μ, a sajátvektoré pedig h:

$$\left(A+E\right)\left(x+h\right) = \left(\lambda+\mu\right)\left(x+h\right).$$

Feltesszük, hogy $E \to 0$ esetén $\mu \to 0$ és $h \to 0$. Beszorzás után a másodrendű tagokat hagyjuk el:

$$Ah + Ex \approx \lambda h + \mu x .$$

Szorozzuk ezt a kifejezést balról az y^T sajátvektorral, ekkor mindkét oldalon az első vektorok is kiesnek és az első bizonyítandó összefüggéshez adódik

$$\mu \approx \frac{y^T E x}{y^T x} . \tag{6.14}$$

Itt a nevező nem lehet zérus a 7.1.7 Tétel miatt. (6.13)-ban $\|E\|_2$ szorzója nem más, mint az x és y vektoroknál a bezárt szög koszinuszának reciproka: $\sec \angle(x, y) = \|x\|_2 \|y\|_2 / |y^T x|$. Szokás ezt az értéket a λ *sajátérték kondiciószámának* nevezni. ■

8. A legkisebb négyzetek módszere

8.1. Egy illesztési feladat.

Gyakran találkozhatunk a következő feladattal: adottak a (t_i, y_i), $i = 0,1,\ldots,n$ pontok, ahol a t_i helyhez tartozó y_i értéket valamely merésből kapjuk. A mért függvényértékeket hiba terheli. Az előálló pontsorozatot – vagy annak egy részét - szeretnénk

egy $f(t) = \sum_{j=0}^{n} c_j \varphi_j(t)$ függvénnyel közelíteni (pl. $\varphi_j(t) = t^j$):

$$\sum_{j=0}^{n} c_j \varphi_j(t_i) \approx y_i . \tag{7.1}$$

Ésszerűnek látszik ezt a feladatot úgy megoldani, hogy az eltérések négyzetösszege minimális legyen:

$$\sum_{i=0}^{m} (y_i - \sum_{j=0}^{n} c_j t_i^j)^2 = \min \tag{7.2}$$

vagyis, (6.1)-ben a c_j lineárkombinációs együtthatókat e feltétel szerint keressük. A (6.1) feladat a c_j ismeretlenekre egy lineáris egyenletrendszer, így tekintsük általánosan az

$$Ax = b,\ A \in \mathbb{R}^{m,n},\ x \in \mathbb{R}^n, b \in \mathbb{R}^m \tag{7.3}$$

egyenletrendszert, ahol most az együtthatómátrix nem kvadratikus, hanem téglalap alakú, és az sem biztos, hogy mindig van megoldása. A legkisebb négyzetes tulajdonságnak eleget tevő megoldáshoz ($\|b - Ax\|_2^2 = \min$) vizsgáljuk meg a projektor (vetítő) mátrixokat!

8.2. Vetítő mátrixok, projektorok

8.2.1 Definició

A P mátrixot projektor vagy vetítő mátrixnak nevezzük, ha eleget tesz a $P^2 = P$ összefüggésnek.

Innen rögtön következik: $P(I - P) = (I - P)P = 0$. Ha P invertálható, akkor P^{-1}-et a definiáló egyenletre alkalmazva kapjuk: $P = I$. Ugyancsak egyszerű ellenőrízni, hogyha P projektor, akkor $I - P$ is az.

Figyeljük meg: P a transzformált vektort egyszeri alkalmazás után helybenhagyja: $P(Px) = Px$. Ezután akárhányszor alkalmazzuk a projektort, az eredmény ugyanaz marad, ugyanúgy, mint vetítéskor. Mivel P akárhányadik hatványa önmaga, emiatt használatos az *idempotens mátrix* elnevezés.

Példa projektorra: Legyen $A \in \mathbb{R}^{m,n}$, $B \in \mathbb{R}^{m,n}$, $n < m$, ^{T} jelölje a transzponáltat és legyen $B^T A$ invertálható. Ekkor

$$P = P(A, B^T) = A(B^T A)^{-1} B^T$$

olyan projektor lesz, amely A oszlopvektorainak alterébe vetít:

$$P(A, B^T): \mathbb{R}^m \to \mathrm{Im}(A).$$

$(I - P)P = 0$ -ból pedig következik: bármely z vektorra $\{I - P^{\mathrm{T}}(A, B^T)\}z \perp \mathrm{Im}(A)$.

8.2.2 Tétel.

Legyen $\mathcal{P}(\mathcal{A})$ az $\mathcal{A} \subset \mathbb{R}^m$ altérbe vetítő projektorok halmaza. Ekkor bármely $x \notin \mathcal{A}$, $Px \neq 0$ vektorra

$$\max_{P \in \mathcal{P}(\mathcal{A})} \frac{x^{\mathrm{T}} P x}{\|Px\|_2} = \|P_s x\|_2 ,$$

ahol P_s az altérbe vetítő szimmetrikus projektor.

Bizonyítás. Szemléletesen szólva: az x vektorral $P_s x$ zárja be a legkisebb szöget. Ha $P_s, P \in \mathcal{P}(\mathcal{A})$, akkor

$$P_s P = P$$

hiszen P oszlopvektorai az $\mathcal{A}$ altérbe esnek, és ezeket egy másik olyan projektor mindig helybenhagyja, amely ugyanabba az altérbe vetít. Felhasználva a Cauchy-egyenlőtlenséget, adódik

$$\frac{x^{\mathrm{T}} P x}{\|Px\|_2} = \frac{x^{\mathrm{T}} P_s P x}{\|Px\|_2} \leq \frac{\|P_s x\|_2 \|Px\|_2}{\|Px\|_2} = \|P_s x\|_2 , \tag{7.4}$$

ahol a maximum még a $\tilde{P}x = \lambda P_s x,\ \lambda > 0$ feltételt kielégítő $\tilde{P}$ projektorok mellett is előáll. ■

8.2.3 Tétel

Ugyanazon altérbe vetítő projektorok között a szimmetrikus projektor egyértelmű.

Bizonyítás. Indirekt. Tegyük fel, P_1 és P_2 két ugyanabba az altérbe vetítő különböző szimmetrikus projektor, ekkor

$$P_1 P_2 = P_2 \quad \Rightarrow \quad P_2 = P_2^{\mathrm{T}} = P_2^{\mathrm{T}} P_1^{\mathrm{T}} = P_2 P_1 = P_1 ,$$

ahonnan ellentmondásra jutottunk. ■

8.2.4 Tétel

Az x vektor $\mathcal{A}$ altértől való távolsága kettes normában:

$$\|(I - P_s)x\|_2 , \quad P_s \in \mathcal{P}(\mathcal{A}) .$$

Bizonyítás. Minthogy $P_s x$ zárja be a legkisebb szöget x -szel, így a $P_s x$ irány mentén található az a pont $\mathcal{A}$ -ban, amely legközelebb van x végpontjához. Keressük tehát azt a λ -t, amelyre $x - \lambda P_s x$ norma-négyzete

$$\|x - \lambda P_s x\|_2^2 = x^{\mathrm{T}} x - 2\lambda x^{\mathrm{T}} P_s x + \lambda^2 x^{\mathrm{T}} P_s x$$

minimális. Deriválással kapjuk, hogy a minimum helye $\lambda = 1$ -nél van. Tehát $(I - P_s)x \perp \mathcal{A}$ és a kettes normája az x vektor $\mathcal{A}$ altértől való távolsága. ■

8.3. Mátrixok általánosított inverze, a pszeudoinverz

Mátrixok általánosított inverzét akkor értelmezzük, ha az inverzük nem létezik. A pszeudoinverz a lineáris egyenletrendszernek azt a megoldását adja, amelyre az eltérés vektor, más szóval reziduum,

$r = b - Ax$ kettes normája minimális. Ha több megoldás is van, akkor a legkisebb kettes normájú megoldást szolgáltatja. Emlékeztetőül két kis lemma felidézésével kezdjük.

8.3.1 Lemma.

Legyenek az L mátrix oszlopai lineárisan függetlenek. Akkor $LB = LC$-ből $B = C$ következik.

Bizonyítás. Átrendezve $L(B-C) = 0$. L oszlopainak bármely lineáris kombinációja a lineáris függetlenség miatt csak akkor lesz zérusvektor, ha $B-C$ oszlopvektorai zérusok. ■

8.3.2 Lemma.

Legyen $A \in \mathbb{R}^{m,n}$. Ekkor $A^T A$ pozitív szemidefinit. Ha A oszlopai lineárisan függetlenek, azaz A oszloprangú, akkor $A^T A$ pozitív definit.

Bizonyítás. Legyen $y = Ax$, ekkor $x^T A^T Ax = y^T y \geq 0$. Ha A oszloprangú, az előző lemma alapján $y = 0$-ból következik $x = 0$, így ez esetben $A^T A$ pozitív definit. ■

8.3.3 A pszeudoinverz

A továbbiakban megmutatjuk, hogy minden mátrixhoz létezik pszeudoniverz, vagy Moore-Penrose féle általánosított inverz A^+, mely a közönséges inverzzel egyenlő, ha a mátrix nemszinguláris, egyéb esetben viszont minimális kettes norma tulajdonságokkal rendelkezik. Ezt a mátrix inverzet a következő tulajdoságok definiálják:

1. $AA^+A = A$, 2. $A^+AA^+ = A^+$,
3. AA^+ hermitikus, 4. A^+A hermitikus.

A definició komplex mátrixokra vonatkozik, mi most valós esetben a hermitikus tulajdonság helyett a szimmetrikusságot követeljük meg. Vegyük észre: Ha az első egyenletet jobbról, vagy balról szorozzuk A^+-tel, az adódik, hogy AA^+ és A^+A projektorok és a 3. és 4. feltétel szerint még szimmetrikusak is. Az első feltétel alapján AA^+ $\mathrm{Im}(A)$-ba vetít, az $A(I - A^+A) = 0$ alakból pedig azt látjuk, hogy $I - A^+A$ a $\ker(A)$-ba vetítő szimmetrikus projektor.

Tegyük fel: $A = LU$, ahol L és U közbülső mérete $r = \mathrm{rang}(A)$, $L \in \mathbb{R}^{m,r}$, $U \in \mathbb{R}^{r,n}$. A továbbiakban az LU rang-faktorizáció ismeretében előállítjuk a pszeudoinverzet.

8.3.4 Tétel, a pszeudoinverz előállítása

Legyen $A = LU$ egy rang faktorizáció. Akkor egyértelműen létezik $A^+ = U^+L^+$, ahol $L^+ = (L^TL)^{-1}L^T$ és $U^+ = U^T(UU^T)^{-1}$.

Bizonyítás. Az egyértelműséget indirekt úton bizonyítjuk. Tegyük fel van kettő: A_1^+ és A_2^+. Ekkor a szimmetrikus projektor egyértelműsége miatt $A_1^+A = A_2^+A$ illetve $AA_1^+ = AA_2^+$. Ezeket, és a definiáló egyenleteket felhasználva adódik

$$A_1^+ = A_1^+AA_1^+ = A_2^+AA_1^+ = A_2^+AA_2^+ = A_2^+,$$

amivel ellentmondásra jutottunk. A továbbiakban megkonstruáljuk a pszeudoinverzet.

Vegyük észre: $\mathrm{Im}(A) = \mathrm{Im}(L)$, így ebbe az altérbe vetítő egyértelmű szimmetrikus projektor $AA^+ = LL^+ = L(L^TL)^{-1}L^T$, és $L^+ = (L^TL)^{-1}L^T$ következik a 8.3.1 Lemmából. Hasonlóan az $\mathrm{Im}(A^T) = \mathrm{Im}(U^T)$ altérbe vetítő egyértelmű szimmetrikus projektor $A^+A = A^T(A^+)^T =$

$= U^T (U^+)^T = U^+ U = U^T (UU^T)^{-1} U$, ahonnan $U^+ = U^T (UU^T)^{-1}$. Ezek alapján $L^+ L = UU^+ = E_r$ r -edrendű egységmátrixok, és ezzel

$$AA^+ = LL^+ = LUU^+ L^+ \quad \text{és} \quad A^+ A = U^+ U = U^+ L^+ LU,$$

Ebből kiolvasható, hogy $A^+ = U^+ L^+$. ■

8.3.5 Megjegyzések

Ha A oszloprangú, akkor $L = A$ és $U = I_n$ megfelelő választás, és $A^+ = (A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}}$ következik. Az $A = QR$ faktorizációnál kapjuk: $A^+ = R^{-1} Q^{\mathrm{T}}$. Ha A sorrangú, akkor $L = I_m$ és $U = A$ a megfelelő választás, amivel $A^+ = A^{\mathrm{T}} (AA^{\mathrm{T}})^{-1}$. Ha most $A^{\mathrm{T}} = QR$, akkor $A^+ = Q(R^{\mathrm{T}})^{-1}$. Végül, ha A rangja kisebb, mint a legkisebb mérete, azaz, vannak lineárisan összefüggő sorok és oszlopok, akkor mindkét oldal felől végzett ortogonalizációval elérhető az $A = Q_1 B Q_2$ alak, ahol Q_1, Q_2 ortogonális mátrixok és B felső bidiagonális mátrix. Ekkor $A^+ = Q_2^{\mathrm{T}} (B)^{-1} Q_1^{\mathrm{T}}$. Mátrixoknál a rang numerikus meghatározása néha nagyon kényes feladat.

8.3.6 Tétel, lineáris egyenletrendszer megoldhatósága

Legyen P $\mathrm{Im}(A)$ -ba vetítő projektor. Ekkor az $Ax = b$ egyenletrendszer akkor és csak akkor konzisztens (megoldható), ha $Pb = b$.

Bizonyítás. Szükségesség. Ha a rendszer megoldható, akkor $b \in \mathrm{Im}(A)$ és $Pb = b$ -nek teljesülni kell. Az elégségességhez válasszuk az AA^+ szintén $\mathrm{Im}(A)$ -ba vetítő projektort, amelyre $AA^+ b = b$ teljesül. Innen kiolvasható, hogy $x = A^+ b$ egy megoldás. ■

8.3.7 Tétel, a pszeudoinverzes megoldás tulajdonságai

Az $Ax = b$ lineáris egyenletrendszer általános megoldása a pszeudoinverz segítségével a következőképp állítható elő:

$$x = x_p + x_h = A^+ b + (I - A^+ A)t, \quad t \in \mathbb{R}^n, \tag{7.5}$$

ahol x_p egy partikuláris megoldás és x_h a homogén egyenlet általános megoldása. Ha a rendszer megoldható, akkor $A^+ b$ egy partikuláris megoldás és $(I - A^+ A)t$ a homogén egyenlet általános megoldása. Ha a rendszer inkonzisztens, akkor $A^+ b$ az a legkisebb négyzetes megoldás, melyre $b - AA^+ b$ kettes normája minimális. Minden esetben $A^+ b$ a minimális kettes normájú megoldás.

Bizonyítás. A 8.1.4 Tétel alapján $\left\| b - AA^+ b \right\|_2$ a b vektor $\mathrm{Im}(A)$ -tól való távolsága. Továbbá vegyük észre, (7.5)-ben két ortogonális vektor van, mert $A^+ A$ az első vektort a pszeudoinverz tulajdonságok miatt helyben hagyja, a másodikat pedig zérusba viszi. Emiatt írhatjuk: $\|x\|_2^2 = \left\| A^+ b \right\|_2^2 + \left\| (I - A^+ A)t \right\|_2^2$, ami akkor a legkisebb, ha $t = 0$, vagy $I - A^+ A = 0$. ■

8.4. Feladatok

1. Legyen $A = LU$ egy rang-faktorizáció. Ekkor írjuk fel azt a szimmetrikus vetítőmátrixot, amely $\mathrm{Im}(A)$ -ba vetít.
2. Írjuk fel az A mátrix null-terébe vetítő szimmetrikus projektort! Adjuk meg az x vektor $\mathrm{Nul}(A)$ -tól való távolságát!

3. Egy egyenes áthalad az r_0 és r_1 ponton. Adjuk meg az x vektor és ez az egyenes távolságát!

4. Igazoljuk, hogyha a mátrix invertálható, akkor a pszeudoinverze megegyezik az inverzével.

5. $A^T = \begin{bmatrix} 2 & -4 & 6 \\ 0 & 5 & -5 \end{bmatrix}$. Írjuk fel az $\mathrm{Im}(A)$ -ba vetítő szimmetrikus projektort!

6. Két sík normálvektorát az 5. feladat sorvektorai adják. Melyik az a szimmetrikus projektor, amely a két sík közös részébe vetít?

7. $r^T = [1 \quad -1 \quad 1]$. r végpontja milyen távol van az előbbi két sík közös részétől?

8. $A = \begin{bmatrix} 2 & 0 & 3 \\ -4 & 5 & -2 \\ 6 & -5 & 5 \end{bmatrix}$, $\mathrm{rang}(A) = 2$. $A^+ = ?$

9. Az előbbi mátrixszal mi lesz $Ax = b$ pszeudoinverzes megoldása, ha $b^T = [1 \quad -1 \quad 1]$?

10. Igazoljuk, hogy $I - A^+A = 0$, ha A oszlopai lineárisan függetlenek.

11. A pszeudoinverz 4 tulajdonságából vezessük le: $(A^+)^T = (A^T)^+$.

12. Az A mátrix közelítő sajátvektora x. A hozzátartozó λ sajátértéket úgy szeretnénk közelíteni, hogy $\|Ax - \lambda x\|_2$ minimális legyen. Mi lesz ekkor λ kifejezése?

9. Ortogonális polinomok

Gyakran polinommal kell végezni a legkisebb négyzetes illesztést. Ilyenkor speciális módszert készíthetünk ortogonális polinomok segítségével. Később a numerikus integrálási módszereknél is szükségünk lesz az ortogonális polinomokra, így most röviden megismerkedünk velük.

9.1. Függvények skaláris szorzata.

Az f és g függvény skaláris szorzatát a következő utasítással definiáljuk:

$$(f,g)=\int_a^b f(x)g(x)\alpha(x)dx, \quad \alpha(x)>0 \tag{8.1}$$

ahol $\alpha(x)$-et súlyfüggvénynek nevezzük és feltesszük, hogy a kijelölt integrál létezik. Mi most a polinomok használatával összefüggésben egyszerűbb skalárszorzatot fogunk használni, nevezetesen:

$$(f,g)=\sum_{i=1}^{m} f(x_i)g(x_i)w_i \tag{8.2}$$

ahol x_i, $i=0,1,\ldots,m$ az illesztés alappontjait, w_i pedig a hozzátartozó súlyokat jelenti. Gyakran $w_i=1$ minden i-re.

Ellenőrízzük, hogy a fenti definiciók rendelkeznek a skaláris szorzat tulajdonságaival!

Természetesen most is igaz, hogy a skaláris szorzat normát definiál:

$$\|f\|^2=(f,f). \tag{8.3}$$

Ezek után semmi akadálya sincs annak, hogy az x^i, $i=0,1,\ldots$ lineárisan független rendszerből Gram-Schmidt ortogonalizációval ortogonális rendszert készítsünk: így kapjuk az *ortogonális polinomokat*.

9.1.1 Definició.

1-*főegyüttható*s az a polinom, amelynél 1 a legmagasabb fokú tag együtthatója.

9.1.2 Tétel

Legyenek $p_i(x)$, $i=0,1,\ldots$ 1-főegyütthatós i-edrendű polinomok. Ekkor bármely $q(x)$ polinom egyértelműen előállítható a p_i polinomok lineáris kombinációjaként:

$$q(x)=\sum_{j=0}^{n} b_j p_j(x). \tag{8.4}$$

Bizonyítás. Legyen $p_i(x)=x^i+p_{i,i-1}x^{i-1}+\ldots+p_{i,0}$, ekkor a b_j együtthatókat meghatározó lineáris egyenletrendszer:

$$\begin{bmatrix} 1 & p_{10} & p_{20} & \cdots & p_{n0} \\ & 1 & p_{21} & \cdots & p_{n1} \\ & & 1 & \cdots & \vdots \\ & \bigcirc & & \ddots & \vdots \\ & & & & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ \vdots \\ a_n \end{bmatrix} \tag{8.5}$$

.

Ez alulról felfelé haladva egyértelműen megoldható.

9.1.3 Következmény

Legyenek p_i-k ortogonális polinomok. Akkor p_{n+1} ortogonális minden legfeljebb n-edfokú polinomra.

9.2. Az ortogonális polinomok rekurziója

Az 1-főegyütthatós ortogonális polinomok a $p_0(x)$ és $p_1(x)$ polinomok ismeretében rekurzíve felépíthetők:

$$p_{n+1} = (x - \alpha_{n+1}) p_n - \beta_n p_{n-1}. \tag{8.6}$$

Bizonyítás. Vizsgáljuk az

$$(xp_k, p_n) = (p_k, xp_n)$$

skaláris szorzatot! Az eredmény zérus, ha

$$k+1 < n \quad \text{és} \quad n+1 < k$$

a *9.1.3 Következmény* miatt. Nemzérus az eredmény, ha

$$n-1 \le k \le n+1,$$

így xp_n kifejthető a p_{n-1}, p_n, p_{n+1} polinomokkal:

$$xp_n = p_{n+1} + \alpha_{n+1} p_n + \beta_n p_{n-1},$$

ahonnan p_{n+1}-re rendezve kapjuk (8.6)-öt. ■

9.2.1 Tétel

Az α_{n+1} és β_n kifejtési együtthatókra érvényes

$$\alpha_{n+1} = \frac{(xp_n, p_n)}{(p_n, p_n)}, \tag{8.7}$$

$$\beta_n = \frac{(xp_n, p_{n-1})}{(p_{n-1}, p_{n-1})} = \frac{(p_n, p_n)}{(p_{n-1}, p_{n-1})}. \tag{8.8}$$

Bizonyítás. Helyettesítsük xp_n értékét a rekurzióból. Az ortogonalitás miatt α_{n+1} értéke rögtön adódik, ha a (8.6) rekurzió mindkét oldalán skaláris szorzatot képezünk p_n-nel. β_n értéke hasonlóan készül, csak most a skaláris szorzatot p_{n-1}-gyel vesszük. Az átalakításban x-et átvisszük p_{n-1}-hez, és az $xp_{n-1} = p_n + \alpha_n p_{n-1} + \beta_{n-1} p_{n-2}$ 1-gyel kisebb indexű rekurziós összefüggést helyettesítjük:

$$\beta_n = \frac{(xp_n, p_{n-1})}{(p_{n-1}, p_{n-1})} = \frac{(p_n, xp_{n-1})}{(p_{n-1}, p_{n-1})} = \frac{(p_n, p_n)}{(p_{n-1}, p_{n-1})}.$$ ■

9.3. Legkisebb négyzetes közelítés ortogonális polinomokkal

A (8.2) skaláris szorzat mellett az induló polinomok, ha $w_i = 1$ minden i-re

$$p_0 \equiv 1, \quad \|p_0\|^2 = \sum_{j=0}^{m} 1 = m+1. \tag{8.9}$$

A következő polinomot p_1-et keressük $x-\alpha_1$ alakban! Ekkor az ortogonalitás miatt

$$(p_0, p_1) = 0 = (p_0, x-\alpha_1) \;\rightarrow\; (p_0, x) = \alpha_1 (p_0, p_0,)$$

ahonnan

$$\alpha_1 = \frac{1}{m+1} \sum_{j=0}^{m} x_j \,, \tag{8.10}$$

így β_0-t zérusnak vehetjük.

A rekurzióból felépített ortogonális polinomokkal a mért y_i, $i = 0,1,\ldots,m$ függvényértékek a következőképp közelíthetők:

$$y \approx \sum_{j=0}^{k} p_j(x) \frac{(p_j, y)}{(p_j, p_j)}, \quad (p_j, y) = \sum_{i=0}^{m} p_j(x_i) y_i \,. \tag{8.11}$$

Ez az előállítás formálisan ugyanúgy néz ki, mint a lineáris algebrában egy $\{q_j\}$ ortogonális vektorrendszer szerinti kifejtés: legyen y $m+1$-dimenziós vektor, ekkor

$$y = \sum_{j=1}^{k} \frac{q_j q_j^T y}{q_j^T q_j} \,. \tag{8.12}$$

A (8.11)-ben látható $P_k = \sum_{j=0}^{k} p_j(x) p_j(t) / (p_j, p_j)$ kifejezés is szimmetrikus projektor, így a (8.11) közelítés rendelkezik a legkisebb négyzetes tulajdonsággal: $\|(I - P_k) y\|$ a p_j, $j = 0,\ldots,k$ polinomok által kifeszített altértől való távolságot jelenti.

9.3.1 Példa

Ortogonális polinomokkal állítsuk elő azt az elsőfokú polinomot, amely az alábbi pontsort legkisebb négyzetesen közelíti:

x_i	-1	0	1	2
y_i	1	2	2	4

Megoldás. Először előállítjuk az ortogonális polinomokat. $p_0(x) = 1$, innen $(p_0, p_0) = 4$. Következik $p_1(x) = x - \alpha_1$, ahol $\alpha_1 = (xp_0, p_0)/(p_0, p_0) = \sum_{j=0}^{3} x_j / (p_0, p_0) = 1/2$. Még meg kell állapítanunk $p_1(x)$ önmagával vett skaláris szorzatát: $(p_1, p_1) = \sum_{j=0}^{3} (x_j - \alpha_1)^2 = \; = \frac{1}{4}(9+1+1+9) = 5$. Ezzel a legkisebb négyzetesen közelítő elsőfokú polinom:

$$P_1(x) = \frac{(p_0, y)}{(p_0, p_0)} p_0 + \frac{(p_1, y)}{(p_1, p_1)} p_1 = \frac{9}{4} + \frac{1}{5} \cdot \frac{1}{2}(-3-2+2+12)(x - \frac{1}{2}) = \frac{9}{4} + \frac{9}{10}(x - \frac{1}{2}).$$

9.4. Feladatok

1. Bizonyítsuk be, hogyha az alappontok $x=0$ -ra szimmetrikusan helyezkednek el, akkor $\alpha_j = 0,\ i = 1,2,\ldots$, és a polinomok váltakozva páros és páratlan függvények.

2. Készítsük el a p_0, p_1, p_2 ortogonális polinomokat a $\{-2,-1,0,1,2\}$ alappontokra!

3. A Csebisev polinomok is ortogonális polinomok, amelyek a következő rekurzióval állíthatók elő: $T_0 = 1,\ T_1 = x,\ T_{n+1} = 2xT_n - T_{n-1}$. Ez a szokásos alak, bár így nem 1-főegyütthatósak. Állítsuk elő a $4x^2 - 3x + 2$ polinomot Csebisev-polinomok szerint!

4. $P(x) = \sum_{j=0}^{k} (2j+1)T_j(x)$. Az x_0 helyen ekkor mi a célszerű kiszámítási módja $P(x_0)$ -nak?

5. Mutassuk meg, hogy $(p_i, p_i) = \mu_0 \beta_1 \beta_2 \ldots \beta_i$, ahol $\mu_0 = (p_0, p_0)\left[= \int_a^b \alpha(x)dx \right]$ a 0-adik momentum.

6. Mutassuk meg, hogy a

$$\begin{pmatrix} x-\alpha_1 & -\beta_1 & & \\ -\beta_1 & x-\alpha_2 & \ddots & \\ & \ddots & \ddots & -\beta_{n-1} \\ & & -\beta_{n-1} & x-\alpha_n \end{pmatrix}$$

háromátlójú mátrix bal felső sarok-aldeterminánsai α_j és β_j^2 paraméterekkel bíró ortogonális polinomokat adnak.

10. Lineáris egyenletrendszerek megoldása iterációval

A lineáris egyenletrendszerek megoldását nem mindig célszerű véges módszerrel készíteni. Ha a mátrix nagyméretű és ritka – azaz soronként csak kevés nemzérus elem található – akkor az *LU*-felbontás hátránya, hogy felbontáskor a mátrix nemzérus elemeinek száma megnő – besűrűsödik – ez egyrészt tárolási kérdéseket vet fel, másrészt a sok nemzérus elem miatt megnő a munkaigény. Az iterációs módszereknél ilyen nehézségek nem lépnek fel, de probléma, ha a konvergencia lassú.

10.1. Egyszerű iteráció

Legyen $A \in \mathbb{R}^{n\times n}$ egy felbontása

$$A = M - N \tag{9.1}$$

Ha M invertálható, akkor a következő iterációs módszert készíthetjük: $Ax = (M-N)x = b \rightarrow x = M^{-1}(b + Nx)$, azaz

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b = Bx^{(k)} + c, \tag{9.2}$$

ahol k a vektor felső indexében az iterációszámot jelöli. A B mátrixot *iterációs mátrixnak* nevezzük. Kérdés, mikor remélhető, hogy a fenti iteráció konvergens és az milyen sebességű?

10.1.1 A konvergencia vizsgálata

Az $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ függvény kontrakció, ha van olyan $0 \le q < 1$ szám, amelyre $\forall x, y \in \mathbb{R}^n$ esetén

$$\|F(x) - F(y)\| \le q\|x - y\| \tag{9.3}$$

teljesül. Itt q a *kontrakciós állandó*, vagy *kontrakciószám.* Figyeljük meg, $q < 1$ azt jelenti, hogy a leképezett vektorok közelebb kerülnek egymáshoz.

10.1.2 Tétel. (Banach fixponttétel)

Legyen $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ egy leképezés $q < 1$ kontrakciós állandóval. Akkor

1) $\exists x^* \in \mathbb{R}^n: \; x^* = F(x^*)$, azaz van az iterációnak fixpontja és ez egyértelmű.

2) $\forall x^{(0)} \in \mathbb{R}^n$ kezdőértékre $x^{(k+1)} = F(x^{(k)})$ konvergens sorozat és $\lim_{k\to\infty} x^{(k)} \rightarrow x^*$.

3) Fennáll a hibabecslés: $\|x^{(k)} - x^*\| \le \frac{q^k}{1-q}\|x^{(1)} - x^{(0)}\|$.

Bizonyítás. Az $x^{(k+1)} = F(x^{(k)})$ sorozat Cauchy-sorozat: $\|x^{(k+1)} - x^{(k)}\| = \|F(x^{(k)}) - F(x^{(k-1)})\| \le$ $\le q\|x^{(k)} - x^{(k-1)}\| \le q^2\|x^{(k-1)} - x^{(k-2)}\| \le \ldots \le q^k\|x^{(1)} - x^{(0)}\|$. Így a sorozat egymás utáni tagjai egyre közelebb vannak egymáshoz: van határérték. Legyen most $m \ge k \ge 1$. Ekkor a teleszkópikus összeg képzésével

$$\|x^{(m)} - x^{(k)}\| = \|x^{(m)} - x^{(m-1)} + x^{(m-1)} - x^{(m-2)} + \ldots + x^{(k+1)} - x^{(k)}\| \le (q^{m-1} + q^{m-2} + \ldots + q^k)\|x^{(1)} - x^{(0)}\| =$$

$$= \frac{q^k - q^m}{1-q}\|x^{(1)} - x^{(0)}\| < \frac{q^k}{1-q}\|x^{(1)} - x^{(0)}\|.$$

$m \rightarrow \infty$ mellett kapjuk a 3) állítást.

Az egyértelműség igazolásához indirekt módon tegyük fel: van két fixpont, x_1^* és x_2^*. De ekkor a kontrakció felhasználásával $\|x_1^* - x_2^*\| = \|F(x_1^*) - F(x_2^*)\| < q\|x_1^* - x_2^*\|$, ami ellentmondás, hiszen a különbség normája nem lehet önmagánál kisebb. ■

A tétel szerint az $x^{(k+1)} = Bx^{(k)} + c$ iteráció konvergens, ha az $F(x) = Bx + c$ leképezés kontrakció:

$$\|F(x) - F(y)\| = \|Bx + c - By - c\| = \|B(x - y)\| \le \|B\|\|x - y\|$$

Innen látható, kontrakció van, ha $\|B\| < 1$. Bizonyítás nélkül megjegyezzük: a spektrál sugár az indukált normák infimuma. Emiatt mondhatjuk: $Bx + c$ konvergens, ha B spektrál sugara $\rho(B) < 1$. A (9.1) felbontást *regulárisnak* nevezzük, ha M invertálható és $\rho(M^{-1}N) < 1$.

10.2. Jacobi-iteráció

Legyen A felbontása $A = L + D + U$, ahol $D = \mathrm{diag}(A)$, L a mátrix főátló alatti, U a főátló feletti része (szigorúan *alsó* ill. *felső* Δ-mátrixok).

A Jacobi-iterációnál $M = D,\ N = -L - U$ a választás, ezzel

$$B_J = -D^{-1}(L + U),\ \ c_J = D^{-1}b. \tag{9.4}$$

Komponensenkénti alak: $x_i^{(k+1)} = -\frac{1}{a_{ii}}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^{(k)} - b_i\right)$.

Tárigény: $A, b, x^{(k)}, x^{(k+1)}$. Célszerű kezdővektor, ha nincs jobb: $x^{(0)} = c_J$.

10.2.1 Tétel

Ha A sor szerint szigorúan főátló-domináns, akkor a Jacobi-iteráció konvergens.

Bizonyítás. $\|B_J\|_\infty = \max_{(k)} \|e_k^T D^{-1}(L + U)\|_\infty = \max_{(k)} \sum_{j\neq k} \left|\frac{a_{kj}}{a_{kk}}\right| < 1$, tehát van kontrakció.

10.3. Gauss-Seidel iteráció

A Gauss-Seidel iterációnál $M = L + D,\ N = -U$ a választás, ezzel

$$B_{GS} = -(L + D)^{-1}U,\ \ c_{GS} = (L + D)^{-1}b. \tag{9.5}$$

A Gauss-Seidel iteráció komponensenkénti alakját $(L + D)x^{(k+1)} = -Ux^{(k)} + b$ i-edik sorának kiírásából kapjuk:

$$x_i^{(k+1)} = -\frac{1}{a_{ii}}\left(\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} - b_i\right),\ \ i = 1, 2, \ldots, n. \tag{9.6}$$

Olyan a műveletek sorrendje, hogy $x_i^{(k)}$ értéke $x_i^{(k+1)}$ értékével felülírható. Így a tárigény: A, b, x előnyösebb, mint a Jacobi-iterációnál. Célszerű kezdővektor: $x^{(0)} = c_{GS}$.

10.3.1 Tétel. Felhasításból származó iterációs mátrix normájának becslése

Legyenek A_1, A_2, D $n \times n$-es valós mátrixok, D diagonálmátrix: $e_i^T D e_i = d_i$ és $\|e_i^T A_1\|_\infty < |d_i|$ $\forall$ i-re. Ekkor fennáll:

$$\left\|(A_1+D)^{-1}A_2\right\|_\infty \le \max_{(i)} \frac{\left\|e_i^T A_2\right\|_\infty}{|d_i| - \left\|e_i^T A_1\right\|_\infty},$$

ahol a maximum-keresésnél elegendő a nemzérus számlálókat tekinteni.

7. *Bizonyítás.* Mivel A_1+D főátlódomináns, a kijelölt inverz létezik. A norma definiciója szerint $\left\|(A_1+D)^{-1}A_2\right\|_\infty = \max_{\|x\|_\infty=1}\left\|(A_1+D)^{-1}A_2x\right\|_\infty$. Legyen $y=(A_1+D)^{-1}A_2x$ és tegyük fel, a maximum y i-edik indexénél valósul meg: $\|y\|_\infty = |y_i|$. Átrendezéssel $A_2x=(A_1+D)y \;\rightarrow\; Dy = A_2x - A_1y$, ahonnan az i-edik sorra $\left\|e_i^T Dy\right\|_\infty = |d_i y_i| \le \left\|e_i^T A_2\right\|_\infty \|x\|_\infty + \left\|e_i^T A_1\right\|_\infty |y_i|$. Figyelembe véve, hogy $\|x\|_\infty = 1$, innen átrendezéssel kapjuk az egyenlőtlenséget. Mivel nem tudjuk, melyik i-re valósul meg $\|y\|_\infty$, ezért a törtet a sorok szerint maximalizáljuk. Ha A_2 i-edik sora zérus, akkor $|d_i||y_i| \le \left\|e_i^T A_1\right\||y_i|$ adódik, ami feltevésünkkel ellentmondó nemzérus $|y_i|$ mellett, így ezeket a sorokat elhagyhatjuk. ■

10.3.2 Tétel

Ha A sor szerint szigorúan domináns átlójú, akkor

$$\|B_{GS}\|_\infty \le \max_{(i)} \frac{\sum_{j=i+1}^{n} |a_{ij}|}{|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}|} \le \|B_J\|_\infty < 1. \tag{9.7}$$

Bizonyítás. Az előző tételben legyen $A_1 = L$, $U = A_2$ és D a mátrix főátlójából alkotott diagonálmátrix. Ezzel a választással közvetlenül adódik az első egyenlőtlenség.

A második egyenlőtlenség igazolásához vezessük be az

$$\alpha_i = \frac{1}{|a_{ii}|}\sum_{j=1}^{i-1}|a_{ij}|, \quad \text{és} \quad \beta_i = \frac{1}{|a_{ii}|}\sum_{j=i+1}^{n}|a_{ij}| \tag{9.8}$$

jelöléseket. Eszerint $\|B_{GS}\|_\infty \le \max_{(i)} \frac{\beta_i}{1-\alpha_i}$ és igazolandó, hogy ez nem nagyobb, mint $\|B_J\|_\infty = \max_{(j)}(\alpha_j+\beta_j)$. Tegyük fel, $\alpha_j > 0$. Ekkor $(\alpha_j+\beta_j) > \beta_j/(1-\alpha_j)$-ből átrendezéssel $\alpha_j + \beta_j - \alpha_j^2 - \beta_j\alpha_j > \beta_j$, ahonnan $1-\alpha_j-\beta_j > 0$ következik. Ez éppen a szigorú főátló-dominancia feltétele, amit feltettünk. Egyenlőség csak akkor lehetséges, ha $\alpha_j = 0$. Így

$$\|B_{GS}\|_\infty \le \max_{(j)} \frac{\beta_j}{1-\alpha_j} = \frac{\beta_k}{1-\alpha_k} \le \alpha_k + \beta_k \le \max_{(j)}(\alpha_j+\beta_j) = \|B_J\|_\infty.$$

Ha balról az első maximumnál $\alpha_k > 0$, akkor biztosan $\|B_{GS}\|_\infty < \|B_J\|_\infty$. ■

10.4. Gauss-Seidel (GS-) relaxáció

Ekkor a gyorsabb konvergencia reményében D szerepét megosztjuk L és U között:

$$(L+D)x = -Ux + b \qquad / \text{ szorozzuk } \omega\text{-val}$$
$$Dx = Dx \qquad / \text{ szorozzuk } (1-\omega)\text{-val}$$

Összeadva, majd x-re rendezve:

$$(D+\omega L)x^{(k+1)} = (1-\omega)Dx^{(k)} - \omega U x^{(k)} + \omega b \tag{9.9}$$

$$x^{(k+1)} = (D+\omega L)^{-1}\left[(1-\omega)D - \omega U\right]x^{(k)} + (D+\omega L)^{-1}\omega b.$$

Innen az iterációs mátrix

$$B_{GS}(\omega) = (D+\omega L)^{-1}\left[(1-\omega)D - \omega U\right]. \tag{9.10}$$

Ha $\omega = 1$, akkor visszakapjuk a Gauss-Seidel iterációt. A komponensenkénti alakot (9.9) i-edik sorából kapjuk:

$$x_i^{(k+1)} = -\frac{\omega}{a_{ii}}\left(\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} - b_i\right) + (1-\omega)x_i^{(k)}, \quad i = 1,2,\ldots,n. \tag{9.11}$$

Eszerint a Gauss-Seidel relaxáció következő lépésének eredményét megszorozzuk ω-val és ehhez hozzáadjuk a k-adik vektor $(1-\omega)$-szorosát.

10.5. A relaxációs módszerekre vonatkozó néhány tétel

1. Ha A egyik diagonáleleme sem 0, egyébként tetszőleges, akkor $\rho(B_{GS}(\omega)) \geq |\omega - 1|$, azaz csak akkor remélhető konvergencia, ha ω 0 és 2 közé esik.
2. Legyen $A \in \mathbb{R}^{n\times n}$ szimmetrikus, pozitív definit mátrix és $0 < \omega < 2$. Ekkor $\rho(B_{GS}(\omega)) < 1$, vagyis minden ilyen ω-ra konvergens a GS-relaxáció.

A következő két tétel blokk-háromátlós mátrixokra vonatkozik. Természetesen 1×1-es blokkok esetén a megszokott mátrixot kapjuk vissza.

3. Legyen $A \in \mathbb{R}^{n\times n}$ blokk-háromátlós mátrix. Akkor a megfelelő blokk Jacobi (J) és GS-iteráció mátrixaira

$$\rho(B_{GS}^b) = \left[\rho(B_J^b)\right]^2.$$

   Ez azt jelenti, a kettő egyszerre konvergens, vagy divergens és konvergencia esetén a GS-iteráció kétszer gyorsabb.

4. Legyen A blokk-háromátlós, szimmetrikus és pozitív definit. Ekkor a blokk-Jacobi iteráció, valamint a blokk-GS relaxáció $0 < \omega < 2$ mellett konvergens. Utóbbinál az optimális relaxációs paraméter

$$\omega_0 = 2/\left(1 + \sqrt{1 - \left(\rho(B_J^b)\right)^2}\right) \in (0,2)$$

és erre az optimális paraméterre a spektrál sugár

$$\rho(B_{GS}^b(\omega_0)) = |\omega_0 - 1| < \rho(B_{GS}^b) = \left(\rho(B_J^b)\right)^2.$$

10.6. Egy lépésben optimális ω paraméter meghatározása

Láttuk, (9.11)-ben az x_k vektorból kiindulva az

$$x_{k+1}^{\omega} = \omega x_{k+1} + (1-\omega)x_k = x_k + \omega(x_{k+1} - x_k) \tag{9.12}$$

vektort készítjük, a Gauss-Seidel módszer x_{k+1} vektora helyett. Vezessük be az $y_k = x_{k+1} - x_k$, $r_k = b - Ax_k$ jelöléseket és határozzuk meg az ω paramétert általánosan az $A = M - N$ felbontás mellett! (9.12)-ból kapjuk:

$$r_{k+1} = b - Ax_{k+1}^{\omega} = r_k - \omega A y_k \,. \tag{9.13}$$

Határozzuk meg a k -adik lépésben ω_k -t abból a feltételből, hogy $\|r_{k+1}\|_2$ minimális! Ehhez nem kell mást tenni, mint az $Ay_k\omega = r_k$ „egyenletet" a pszeudoinverzzel ω -ra megoldani:

$$\omega_k = \left(Ay_k\right)^+ r_k = \frac{y_k^T A^T r_k}{\|Ay_k\|_2^2} = \frac{r_k^T A y_k}{\|Ay_k\|_2^2} \tag{9.14}$$

Hogy ne kelljen x_{k+1} -et explict módon előállítani, a relaxáció nélküli alakból kifejezzük y_k -t:

$$x_{k+1} = M^{-1}(Nx_k + b) = x_k + M^{-1}\left(b - (M-N)x_k\right) = x_k + M^{-1}r_k \,, \tag{9.15}$$

innen

$$y_k = M^{-1}r_k \,. \tag{9.16}$$

Az ω_k meghatározásához vezessünk be egy újabb vektort:

$$c_k = Ay_k = (M-N)M^{-1}r_k = r_k - Ny_k \,, \tag{9.17}$$

és ekkor a következő iterációs algoritmust készíthetjük:

Kezdés: $r_0 = b - Ax_0$;

$k = 1,2,3,\ldots,$-ra:

$y_k = M^{-1}r_k$;

$c_k = r_k - Ny_k$;

$\omega_k = \dfrac{r_k^T c_k}{c_k^T c_k}$;

$x_{k+1} = x_k + \omega_k * y_k$;

$r_{k+1} = r_k - \omega_k * c_k \quad (= b - Ax_{k+1})$;

Két lehetőség is van r_{k+1} számítására. Természetesen az első az olcsóbb. Az iteráció előrehaladtával lehet, hogy a második módszer eredménye jelentősen eltér az elsőétől. Célszerű ilyenkor r_{k+1} értékét a második, pontosnak tekinthető módszerrel feljavítani. Az algoritmusban a vektorokat indexeltük, bár nem szükséges, mivel minden lépésben az előző vektor az újjal felülírható.

10.7. A Richardson iteráció

Ha a mátrixunk sajátértékei valós, pozitív számok, akkor egy iterációt készíthetünk a következő észrevétel alapján:

$$(I - pA)r_i = r_{i+1} = b - A(x_i + pr_i) = b - Ax_{i+1}, \quad x_{i+1} = x_i + pr_i \,, \tag{9.18}$$

ahol a p számot úgy választjuk, hogy $I - pA$ spektrál sugara minél kisebb legyen. Az $I - pA$ mátrix sajátértékei most $1 - p\lambda_i$-k. Látjuk, az $1 - px$ leképező függvény a (0,l) ponton áthaladó, pozitív p mellett negatív meredekségű egyenes. Legyen a legkisebb sajátérték m, a legnagyobb M. A Richarson iterációnál az optimális p értéket abból a feltételből határozzuk meg, hogy a legkisebb és a legnagyobb sajátérték ugyanakkora abszolút értékű számokba képződjön le:

$$1 - pm = -(1 - pM) \quad \rightarrow \quad p = 2/(m + M). \tag{9.19}$$

Ezzel a választással $I - pA$ spektrál sugara $(M - m)/(M + m)$ lesz.

H nem ismerjük a mátrix sajátértékeit, de tudjuk, hogy a sajátértékek pozitívak, pl. mert A szimmetrikus, pozitív definit, a p számot abból a feltételből is kereshetjük, hogy $\|r_{i+1}\|_2$ legyen minimális. Ekkor az $r_i = pAr_i$ egyenlet pszeudo-megoldása

$$p = \frac{r_i^T A^T r_i}{\|Ar_i\|_2^2} = \frac{r_i^T A r_i}{\|Ar_i\|_2^2}. \tag{9.20}$$

Az iteráció során elég néhányszor kiszámolni p értékét, hiszen az az előbb megállapított optimális érték körül fog ingadozni.

10.8. Feladatok

1. Bizonyítsuk be, szimmetrikus mátrixokra a Rayleigh-hányados legkisebb értéke a legkisebb sajátérték.
2. Hogy hajtsuk végre a Jacobi-iterációt, ha a mátrix az oszlopai szerint szigorúan főátló-domináns?
3. Mutassuk meg, a 10.3.1 tétel átfogalmazható arra az esetre, amikor a mátrix oszlopai szerint szigorúan főátló-domináns.
4. Mit ad a (9.7) becslés arra az esetre, ha a sor szerinti főátló-dominancia csak úgy teljesül, hogy néhány sorban van egyenlőség? És ha csak az utolsó sorban van egyenlőség?
5. $A = \begin{pmatrix} 5 & -1 & 2 & 1 \\ -3 & 7 & -2 & 0 \\ 3 & 0 & 5 & -1 \\ 0 & 2 & -4 & 6 \end{pmatrix}$. $\|B_J\|_\infty = ?$ $\|B_{GS}\|_\infty \le ?$
6. A 10.3.1 Tétel segítségével bizonyítsuk be: $\|A^{-1}\|_\infty \le \max_i \frac{1}{|a_{ii}|(1-\alpha_i-\beta_i)}$, ld. (9.8)-at is, ha A sorai szerint szigorúan főátló-domináns. Oszlopok szerinti főátló-dominancia esetén hogyan módosítsuk az állítást?
7. Feltéve, hogy $D+\omega L$ főátlódomináns a sorai szerint, a 10.3.1 Tétel segítségével mutassuk meg: $\|B_{GS}(\omega)\|_\infty \le \max_{(j)} \frac{|1-\omega|+\omega\beta_j}{1-\omega\alpha_j}$.
8. Ha $\|D^{-1}A_1\| < 1$, a 10.3.1 Tételhez hasonló egyenlőtlenséget származtathatunk (2.15) fellhasználásával, mivel $(A_1+D)^{-1}A_2 = (I+D^{-1}A_1)^{-1}D^{-1}A_2$. Mutassuk meg, hogy ekkor indukált normával érvényes: $\|(A_1+D)^{-1}A_2\| \le \frac{\|D^{-1}A_2\|}{1-\|D^{-1}A_1\|}$. Szükséges, hogy most D diagonálmátrix legyen? Az 5. Példa mátrixára melyik módszer ad jobb becslést?

11. A Lagrange interpoláció és hibája

Az interpoláció a függvény közelítések olyan módja, ahol azt írjuk elő, hogy az interpoláló függvény a közelíteni kívánt függvény értékét vegye fel a megadott helyeken. Az interpoláció alappontjait gyűjtsük az $\Omega_n = \{x_0, x_1, \ldots, x_n\}$ halmazba, ahol az x_i-k nem szükségképpen rendezettek. A tulajdonságokat az $[a,b]$ intervallumban fogjuk vizsgálni. Sokszor $[a,b] = [\min_i x_i, \max_i x_i]$, de az is lehetséges, hogy minden alappont $[a,b]$ belső pontja.

11.1. Interpoláló függvény lineáris paraméterekkel

Legyen $n \in \mathbb{N}$, és tegyük fel, az $x_k \in \mathbb{R}, \quad k = 0,1,...,n$ pontokban ismerjük az $f(x)$ függvény értékeit. Az interpoláció alkalmával eljárhatunk úgy, hogy felveszünk egy

$$\Phi(x) = \sum_{i=0}^{n} a_i \varphi_i(x) \tag{10.1}$$

alakú próbafüggvényt, ahol az a_i paraméterek meghatározandók az

$$f(x_i) = \Phi(x_i), \quad i = 0,...,n \tag{10.2}$$

feltételekből. Az (1.1)-ben szereplő $\varphi_i(x)$ függvények lehetnek például hatványfüggvények, $\varphi_i(x) = x^i$, amivel interpolációs polinomhoz jutunk, de választhatunk más függvényeket: $\varphi_i(x) = \sin(i\omega x), \quad \varphi_i(x) = \cos(i\omega x), \quad \varphi_i(x) = \exp(i\omega x)$. Ha $n = 2$, az (1.2) interpolációs feladat a következő lineáris egyenletrendszerre vezet:

$$\begin{pmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \varphi_2(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \varphi_2(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \varphi_2(x_2) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \end{pmatrix}. \tag{10.3}$$

Látjuk tehát, hogyha a függvények lineáris kombinációját vesszük, akkor az interpolációs feladat lineáris egyenletrendszerre vezet. A feladat egyértelműen megoldható, ha a kapott rendszer együtthatómátrixának van inverze.

11.2. Polinom-interpoláció

Ekkor a $\varphi_i(x) = x^i$ választással az (1.2) rendszerben az együtthatómátrix Vandermonde mátrix,

$$\begin{pmatrix} 1 & x_0 & \ldots & x_0^n \\ 1 & x_1 & \ldots & x_1^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^n \end{pmatrix} \tag{10.4}$$

amiről tudjuk, hogy determinánsa nemzérus, ha az x_i alappontok különbözők. Következik, hogy az egyváltozós polinom-interpoláció feladata különböző alappontokra egyértelműen megoldható.

11.3. Interpoláció Lagrange-alappolinomokkal

Az Ω_n -ben szereplő alappontokhoz rendeljük a kövekező polinomot:

$$\omega_n(x) = \prod_{j=0}^{n} (x - x_j). \tag{10.5}$$

Ez $n+1$-edfokú és az alappontokon eltűnik. Segítségével könnyen bevezethetünk egy olyan n-edfokú Lagrange-alappolinomot, amely minden alappontban zérust ad, egyet kivéve, - legyen ez az i-edik, és e helyen az értéke legyen 1:

$$l_i(x) = \frac{\omega_n(x)}{(x - x_i) \prod_{j=0, j\neq i}^{n} (x_i - x_j)} = \frac{\omega_n(x)}{(x - x_i)\omega_n'(x_i)} = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \tag{10.6}$$

Itt az $(x - x_i)$ tényezővel azért osztottunk, hogy az (1.5) szorzatból kihagyjuk, és a produktum a nevezőben azért szerepel, hogy $l_i(x_i) = 1$ legyen. Ha most (1.1)-ben $\varphi_i(x) = l_i(x)$, akkor az interpolációs feladat együtthatómátrixa az E egységmátrix, mert $l_i(x_j) = \delta_{ij}$ - ahol δ_{ij} a Kronecker delta. Innen adódik:

$$a_i = f(x_i), \tag{10.7}$$

és a Lagrange-interpoláció polinomja:

$$L_n(x) = \sum_{i=0}^{n} f(x_i) l_i(x). \tag{10.8}$$

Ekkor a Lagrange-alappolinomok tulajdonsága alapján $L_n(x_i) = f(x_i)$.

11.3.1 Tétel. Az interpoláció hibája.

Legyen az interpolált függvény legalább $n+1$-szer differenciálható az $[a,b]$ intervallumon: $f(x) \in C^{n+1}[a,b]$, ahol az alappontok az $[a,b]$-ben vannak . Akkor $\forall x \in [a,b]$ esetén $\exists \xi_x \in [a,b]$, melyre

$$f(x) - L_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \omega_n(x), \tag{10.9}$$

továbbá

$$\left| f(x) - L_n(x) \right| \leq \frac{M_{n+1}}{(n+1)!} \left| \omega_n(x) \right|, \tag{10.10}$$

ahol $M_k \leq \left\| f^{(k)} \right\|_\infty = \max_{x\in[a,b]} \left| f^{(k)}(x) \right|$.

Bizonyítás. Ha $x \in \Omega_n$, akkor (1.9) mindkét oldala zérus és az egyenlőség fennáll. A továbbiakban tegyük fel, $x \notin \Omega_n$ és vezessük be a

$$g_x(z) = f(z) - L_n(z) - \frac{\omega_n(z)}{\omega_n(x)} (f(x) - L_n(x)), \quad z \in [a,b] \tag{10.11}$$

függvényt. Szintén teljesül $g_x(z) \in C^{n+1}[a,b]$ és $g_x(z) = 0$, $z \in \Omega_n$, de ezen felül $z = x$ is zérushely, így összesen $n+2$ db zérus van. A zérushelyek között többszörösen alkalmazva a Rolle-tételt, az $(n+1)$-edik deriválás után kapjuk: $\exists \xi_x \in [a,b]$, amelyre

$$g_x^{(n+1)}(\xi_x) = 0 = f^{(n+1)}(\xi_x) - 0 - \frac{(n+1)!}{\omega_n(x)} (f(x) - L_n(x))$$

és innen rendezéssel kapjuk (1.9)-et. A második állítás úgy adódik, hogy mindkét oladolon vesszük az abszolút értéket és az $(n+1)$-edik deriváltat felülről becsüljük az $[a,b]$ intervallumban. ■

Megjegyzés. Rolle tétele szerint, ha $f(a)=f(b)=0$ és $[a,b]$-ben f deriválható, akkor $[a,b]$-ben van egy olyan pont, ahol a függvény deriváltja zérus. E tétel egyszerű következménye a Lagrange középérték-tételnek és akkor is igaz, ha $f(a)=f(b)$. Hasonlóan (1.10)-hez, az $(n+1)$-edik derivált abszolút érték minimumával az alsó becslés is elkészíthető.

11.4. Példa

Az $\Omega_n=\{x_0,x_1,\ldots,x_n\}$ alappontokon adott y_i értékek egy n-edfokú $p_n(x)$ polinom értékei. Mutassuk meg, hogy ezen pontokra készített $L_n(x)$ interpolációs polinomra $L_n(x)=p_n(x)$.

Megoldás. Vizsgáljuk meg az interpoláció hibáját:

$$p_n(x)-L_n(x)=\frac{p_n^{(n+1)}(\xi_x)}{(n+1)!}\omega_n(x)=0,$$

mert az n-edfokú polinom $(n+1)$-edik deriváltja mindenütt zérus.

11.5. Feladatok

1. Egy függvény 3 pontban adott: (-1,-1), (1,1), (2,3). Készítsük el a Lagrange-alappolinomokat és azt az $L_2(x)$ polinomot, mely áthalad e pontokon.

2. Az $f(x)=(x+1)^{-2}$ függvényt a [0,1] intervallumban interpoláljuk az $\Omega=\{0,\ 0.2,\ \ 0.5,\ 0.8,\ 1\}$ alappontokon. Becsüljük meg az $\left|f(x)-L_4(x)\right|$ hibát az $x=0.4$ helyen!

3. Lássuk be: $\sum_{j=0}^{n} x_j^k l_j(x)=x^k$, ha $k\le n$.

4. Igazoljuk: $x^{n+1}-\sum_{j=0}^{n} x_j^{n+1} l_j(x)=\omega_n(x)$.

12. A polinom-interpoláció tulajdonságai

Természetesen adódik a kérdés: Pontosabb az interpoláció közelítése, ha növeljük a polinom fokszámát? Ekkor konvergál-e a polinom a függvényhez? A válasz nem mindig igenlő, de van eset, amikor az.

12.1. Tétel, egyenletes konvergencia

Legyen $f \in C^{\infty}[a,b]$ és legyen $x_k^{(n)}$, $k = 0,1,\ldots,n$; $n = 0,1,2,\ldots$ az $[a,b]$ intervallumot kifeszítő alappontrendszerek egy sorozata. Jelölje $L_n(x)$ az $x_0^{(n)}, x_1^{(n)}, \ldots, x_n^{(n)}$ alappontrendszerre illesztett Lagrange interpolációs polinomot, $n = 0,1,2,\ldots$ Ha $\exists M > 0$ úgy, hogy $M_n \le M^n$ $\forall n$-re, akkor az L_n interpolációs polinomok sorozata egyenletesen konvergál az $f(x)$ függvényhez.

Bizonyítás. Alkalmazzuk az egész intervallumra érvényes hibakorlátot, majd becsüljük felülről az $\|\omega_n\|_{\infty}$ normát:

$$|f(x) - L_n(x)| \le \frac{M_{n+1}}{(n+1)!}\|\omega_n\|_{\infty} \le \frac{M^{n+1}(b-a)^{n+1}}{(n+1)!} = \frac{[M(b-a)]^{n+1}}{(n+1)!}.$$

A nevezőben lévő faktoriális függvény a hatványfüggvénynél gyorsabban tart ∞-hez, emiatt a jobb oldal zérushoz tart, így igaz az egyenletes konvergencia:

$$\|f - L_n\|_{\infty} \to 0,$$

ami azt jelenti, hogy a két függvény maximális abszolút eltérése zérushoz tart. ■

12.2. Lemma

Rendezzük nagyság szerint az alappontokat: $x_{k-1} < x_k$ és legyen $h = \max\limits_{k=1,2,\ldots,n} |x_k - x_{k-1}|$. Ekkor $|\omega_n(x)|$-re a következő becslés adható:

$$|\omega_n(x)| \le \frac{n!}{4}h^{n+1}, \quad x \in [a,b]. \tag{11.1}$$

Bizonyítás. Átvizsgáljuk az egyes intervallumokat. Először legyen $x \in [x_0, x_1]$. Ekkor deriválva és x értékét a maximum helyen véve kapjuk: $|(x-x_0)(x-x_1)| \le h^2/4$. Tovább felhasználva adódik:

$$|\omega_n(x)| \le (h^2/4)(2h)(3h)\ldots(nh) = \frac{h^{n+1}n!}{4}.$$

Másodszor legyen $x \in [x_1, x_2]$. Hasonlóan kapjuk:

$$|\omega_n(x)| \le (2h)(h^2/4)(2h)\ldots((n-1)h) < \frac{h^{n+1}n!}{4}.$$

A többi belső intervallumra is azt kapjuk, hogy kisebb a becslés eredménye, mint az első intervallumra. Végül az utolsó intervallumra az elsőével azonos becslésre jutunk, így (11.1) a végső eredmény. ■

Megjegyzés. Az itt látott becslés alapján kisebb hibára számítunk, ha x az $[a,b]$ intervallum közepén van, mint amikor az $[a,b]$ intervallum széleinél volna. Ez akkor igaz, ha az alappontok közel

egyenletesen helyezkednek el. Sejthető, jobb lesz a közelítés, ha az alappontok az intervallum széleinél sűrűbben helyezkednek el.

A megismert lemma segítségével az egész intervallumra érvényes hibakorláthoz jutunk:

12.3. Tétel

Az alappontok legyenek nagyság szerint rendezettek: $x_{k-1} < x_k$, ahol $h = \max\limits_{k=1,2,\ldots,n} |x_k - x_{k-1}|$. Ekkor fennáll

$$|f(x) - L_n(x)| \le \frac{M_{n+1}}{4(n+1)} h^{n+1}, \quad x \in [a,b]. \tag{11.2}$$

Bizonyítás. (11.1)-et beírva a hibatételbe kapjuk az állítást. ■

12.4. Az alappontok ügyes megválasztása, Csebisev polinomok

Az n-edfokú Csebisev polinom a következő összefüggéssel adható meg:

$$T_n(x) = \cos(n \arccos x), \quad x \in [-1,1], \quad n = 0,1,\ldots \tag{11.3}$$

Belátjuk, hogy ez polinom. Legyen $\vartheta = \arccos x$, ekkor

$$\begin{aligned} T_{n\pm1}(x) &= \cos((n\pm1)\vartheta) = \cos(n\vartheta \pm \vartheta) = \\ &= \cos(n\vartheta)\cos\vartheta \mp \sin(n\vartheta)\sin(\vartheta) = xT_n(x) \mp \sin(n\vartheta)\sin\vartheta. \end{aligned}$$

A + és − előjelekhez tartozó kifejezéseket összeadva:

$$T_{n+1}(x) + T_{n-1}(x) = 2xT_n(x), \tag{11.4}$$

azaz a Csebisev polinomok előállítására a következő rekurziót kapjuk:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad x \in [-1,1]. \tag{11.5}$$

Ha az első néhány polinomot kiírjuk, azt találjuk, hogy $T_n(x) = 2^{n-1}x^n + \ldots$, $n > 0$. Így vezessük be az 1-főegyütthatós Csebisev polinomokat a

$$\tilde{T}_n(x) = 2^{1-n}T_n(x), \quad 0 < n$$

utasítással. Ekkor igaz a következő tétel:

12.5. Tétel

$\|\tilde{T}_n\|_\infty \le \|p\|_\infty$, $p \in \mathcal{P}_n^1[-1,1]$, azaz szavakban: Jelölje $\mathcal{P}_n^1[-1,1]$ az 1-főegyütthatós n-edfokú polinomokat $[-1,1]$-ben, akkor e polinomok között az 1-re normált Csebisev polinom lesz az, amelyik a $[-1,1]$ intervallumban a legkisebb maximális értéket veszi fel, azaz ott legjobban közelíti a 0 függvényt.

Bizonyítás. A $T_n(x)$ polinomok a cos függvénynek megfelelően -1 és 1 között oszcillálnak. A szélsőérték helyek:

$$\cos(n \arccos z_k) = (-1)^k, \quad \text{ahonnan} \quad z_k = \cos\frac{k\pi}{n}, \quad k = 0,1,\ldots n, \tag{11.6}$$

$n+1$ különböző pont. Az 1-re normált polinomok szélsőérték helyei ugyanitt vannak. Most indirekt módon tegyük fel, hogy $\exists p \in \mathcal{P}_n^1[-1,1]$, amelyre $\|p\|_\infty < \|\tilde{T}_n\|_\infty$. De ekkor az $r = \tilde{T}_n - p$ különbség

polinom $n-1$-edfokú és a szélsőérték helyek közt előjelet kéne váltani, $n+1$ hely között összesen n-szer. De ez ellentmondás, mert $r(x)$-nek legalább n-ed-fokúnak kéne lennie. ■

A Csebisev polinomok gyökei. $\cos(n \arccos x_k) = 0 \to n \arccos x_k = \frac{\pi}{2} + k\pi \to$

$$x_k = \cos\frac{(2k+1)\pi}{2n}, \quad k = 0,1,\ldots,n-1 \text{ különböző hely.} \tag{11.7}$$

Következmény. $[-1,1]$-ben az alappontokat válasszuk úgy, hogy egybeessenek a Csebisev polinomok gyökeivel. Így érjük el a legkisebb hibakorlátot az $\omega_n(x)$ polinomnál:

$$|f(x) - L_n(x)| \le \frac{M_{n+1}}{(n+1)!}\|\omega_n\|_\infty = \frac{M_{n+1}}{(n+1)!}\|\tilde{T}_{n+1}\|_\infty = \frac{M_{n+1}}{(n+1)!}\frac{1}{2^n}. \tag{11.8}$$

Ha $x \in [a,b]$ akkor a gyökök egyszerű lineáris transzformációval $[-1,1]$-ből oda átvihetők.

12.6. Feladatok

1. Döntsük el, hogy az interpoláló polinom egyenletesen tart-e az alábbi függvényekhez, ha az $[a,b]$ intervallumot kifeszítő alappontok száma $n \to \infty$:

   a) $f(x) = \sin x, \quad x \in [0,\pi]$ *b)* $f(x) = \cos x, \quad x \in [0,\pi]$

2. *c)* $f(x) = e^x, \quad x \in [0,1]$ *d)* $f(x) = (x+2)^{-1}, \quad x \in [0,1]$

   e) $f(x) = (x+2)^{-1}, \quad x \in [-1,1]$

3. Az előző feladat *e)* példájánál mi a helyzet, ha az alappontoknak mindig a Csebisev polinomok gyökeit választjuk?
4. Állapítsuk meg azt az egyszerű lineáris transzformációt, amely a $t \in [-1,1]$ változót átviszi az $x \in [a,b]$ változóba! Mi lesz az inverz transzformáció?
5. Írjuk fel, $[a,b]$-ben mik legyenek az alappontok, hogy $\|\omega_n(x)\|_\infty$ minimális legyen!
6. (11.8) abban az esetben szolgáltatja a hiba becslését, amikor $x \in [-1,1]$. Vezessük le a hibabecslést arra az esetre, ha $x \in [a,b]$!
7. A $\sin x$ függvényt az $[0,\pi/2]$ intervallumban milyen sűrűn kell egyenletesen tabellázni, hogy lineáris interpolációt használva 10^{-4} hibával tudjuk mindenütt a függvény értékét számítani?
8. Az előző feladatnál hogy módosul az eredmény, ha másodfokú polinommal interpolálunk? Ekkor $\omega_2(x)$ maximális abszolút értékű helyét pontosan meg tudjuk határozni?
9. Mutassuk meg, hogy a (11.1) becslés továbbírható: $|\omega_n(x)| \le n!(K_n(b-a))^{n+1}/(4n^{n+1})$, ahol $1 \le K_n = hn/(b-a)$ állandó. Mikor lesz $K_n = 1$?
10. A Stirling-formula szerint $n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\left(1 + \frac{1}{12n} + \frac{1}{288n^2} - \ldots\right)$. Az előző feladat segítségével mutassuk meg, hogy egyenletes felosztás mellett $|b-a| \le e$ esetén $\lim_{n\to\infty} \omega_n(x) \to 0$.
11. Igazoljuk, hogy a Csebisev polinomok ortogonálisak a $(T_i, T_j) = \int_{-1}^{1} \alpha(x) T_i(x) T_j(x) dx$ skaláris szorzat szerint, ahol a súlyfüggvény $\alpha(x) = (1-x^2)^{-1/2}$.
12. Adjuk meg (T_n, T_n) értékét!

13. Iterált interpoláció (Neville, Aitken, Newton)

13.1. A Neville- és Aitken-interpoláció.

A Lagrange-interpoláció hátránya, hogy újabb osztópontok felvételekor az alappolinomokat újra kell számolni. És van, amikor nem is az interpolációs polinom, hanem közvetlenül annak helyettesítési értéke kéne. Ilyenkor előnyös az iterált interpoláció.

Legyenek az interpoláció tartópontjai $\{(x_i, f_i = f(x_i)\}_{i=0}^{n}$, és jelöljük $p_{0,1,\ldots,k}(x)$-szel azt a k-adfokú polinomot, amelyre

$$p_{0,1,\ldots,k}(x_j) = f(x_j),\ \ j = 0,1,\ldots,k, \tag{12.1}$$

azaz interpolációs polinom a megjelölt pontokra. Megmutatjuk, hogy e polinomok rekurzióval is felépíthetők. Tekintsük a következő determinánst:

$$p_{0,1,\ldots,k,k+1}(x) = \frac{1}{x_{k+1} - x_0} \begin{vmatrix} x - x_0 & p_{0,1,\ldots,k}(x) \\ x - x_{k+1} & p_{1,\ldots,k+1}(x) \end{vmatrix} \tag{12.2}$$

Közvetlen ellenőrzéssel kapjuk, hogy az új polinom jól interpolál az $x = x_0$ és $x = x_{k+1}$ pontokban. A közbülső pontokban pedig, ahol $0 < j < k+1$,

$$p_{0,1,\ldots,k,k+1}(x_j) = \frac{1}{x_{k+1} - x_0} \begin{vmatrix} x_j - x_0 & p_{0,1,\ldots,k}(x_j) \\ x_j - x_{k+1} & p_{1,\ldots,k+1}(x_j) \end{vmatrix} = f(x_j) \frac{x_j - x_0 - (x_j - x_{k+1})}{x_{k+1} - x_0} = f(x_j).$$

E rekurzió alapján a Neville-interpolációhoz a következő számtáblázatot készítjük:

	$k=0$	1	2	3
$x - x_0$	$f_0 = p_0(x)$			
$x - x_1$	$f_1 = p_1(x)$	$p_{01}(x)$		
$x - x_2$	$f_2 = p_2(x)$	$p_{12}(x)$	$p_{012}(x)$	
$x - x_3$	$f_3 = p_3(x)$	$p_{23}(x)$	$p_{123}(x)$	$p_{0123}(x)$

Vegyük észre, hogy most egy újabb pont (x_4, f_4) hozzávételével elegendő az x_3-ig kész táblázathoz egy újabb sort kiszámolni. Ha az $x - x_j$-k számok, akkor a bal oszlop számainál a felsőből vonjuk ki az alsót a rekurziós formula nevezőjének előállításához, pl. $x - x_0 - (x - x_j)$. Például határozzuk meg a táblázat értékeit $x = 2$-re, ha a tartópontok:

x_i	0	1	3
f_i	1	3	2

A számolás menete:

	$k=0$	1	2
$2-0=2$	1		
$2-1=1$	3	$\frac{1}{2-1}\begin{vmatrix}2 & 1\\ 1 & 3\end{vmatrix}=5$	
$2-3=-1$	2	$\frac{1}{1-(-1)}\begin{vmatrix}1 & 3\\ -1 & 2\end{vmatrix}=5/2$	$\frac{1}{2-(-1)}\begin{vmatrix}2 & 5\\ -1 & 5/2\end{vmatrix}=10/3$

Az Aitken-interpoláció filozófiája hasonló, csak más köztes polinomokat állít elő. A sorrendet az Aitken-interpoláció táblázatával szemléltetjük:

	$k=0$	1	2	3
$x-x_0$	$f_0=p_0(x)$			
$x-x_1$	$f_1=p_1(x)$	$p_{01}(x)$		
$x-x_2$	$f_2=p_2(x)$	$p_{02}(x)$	$p_{012}(x)$	
$x-x_3$	$f_3=p_3(x)$	$p_{03}(x)$	$p_{013}(x)$	$p_{0123}(x)$

13.2. Osztott differenciák

Mielőtt rátérnénk a Newton-interpolációra, bevezetjük az osztott differenciákat. Legyenek most is a tartópontok $\{(x_i, f_i = f(x_i)\}_{i=0}^n$, ekkor *elsőrendű osztott differenciák:*

$$f[x_0,x_1]=\frac{f(x_1)-f(x_0)}{x_1-x_0},\quad f[x_i,x_{i+1}]=\frac{f(x_{i+1})-f(x_i)}{x_{i+1}-x_i}, \tag{12.3}$$

másodrendű oszott differenciák:

$$f[x_i,x_{i+1},x_{i+2}]=\frac{f[x_{i+1},x_{i+2}]-f[x_i,x_{i+1}]}{x_{i+2}-x_i}, \tag{12.4}$$

és általában a k *-adrendű osztott differenciák:*

$$f[x_i,x_{i+1},\ldots,x_{i+k}]=\frac{f[x_{i+1},x_{i+2}\ldots,x_{i+k}]-f[x_i,x_{i+1},\ldots,x_{i+k-1}]}{x_{i+k}-x_i}, \tag{12.5}$$

ahol a k -adrendű osztott differencia $k+1$ pontra támaszkodik. Az osztott differenciáknak a következő táblázatát készíthetjük:

	$k=0$	1	2	3
x_0	$f(x_0)$			
x_1	$f(x_1)$	$f[x_0,x_1]$		
x_2	$f(x_2)$	$f[x_1,x_2]$	$f[x_0,x_1,x_2]$	
x_3	$f(x_3)$	$f[x_2,x_3]$	$f[x_1,x_2,x_3]$	$f[x_0,x_1,x_2,x_3]$

Példa. Készítsük el az osztott differenciák táblázatát, ha a tartópontok:

x_i	1/2	1	2	3
f_i	2	1	1/2	1/3

Az alábbi táblázatban az oszlopok feletti szám az osztott differencia rendjét mutatja.

		1	2	3
1/2	2			
1	1	$\frac{1-2}{1-1/2}=-2$		
2	1/2	$\frac{1/2-1}{2-1}=-\frac{1}{2}$	$\frac{-1/2-(-2)}{2-1/2}=1$	
3	1/3	$\frac{1/3-1/2}{3-2}=-\frac{1}{6}$	$\frac{-1/6-(-1/2)}{3-1}=\frac{1}{6}$	$\frac{1/6-1}{3-1/2}=\frac{-1}{3}$

13.2.1 Lemma

Fennáll az összefüggés:

$$f[x_0,x_1,\ldots,x_k]=\sum_{j=0}^{k}\frac{f(x_j)}{\omega_k'(x_j)}, \tag{12.6}$$

itt $\omega_k(x)$ az (10.5)-ben megismert szorzatfüggvény.

Bizonyítás. Teljes indukcióval végezhető. $k=1$-re az állítás igaz. A k-ról $k+1$-re való áttérésnél az

$$f[x_0,x_1,\ldots,x_{k+1}]=\frac{f[x_1,\ldots,x_{k+1}]-f[x_0,\ldots,x_k]}{x_{k+1}-x_0}$$

összefüggés alapján az 1-gyel kisebb rendű osztott differenciákba írjuk be a tétel állítását:

$$f[x_1,\ldots,x_{k+1}]=\sum_{j=1}^{k+1}\frac{f(x_j)(x_j-x_0)}{\omega_{k+1}'(x_j)},\quad f[x_0,\ldots,x_k]=\sum_{j=0}^{k}\frac{f(x_j)(x_j-x_{k+1})}{\omega_{k+1}'(x_j)},$$

majd rendezéssel nyerjük az eredményt. ■

Látható, az osztott differencia az alappontok szimmetrikus függvénye. Értéke független attól, hogy az alappontok milyen sorrendben vannak megadva.

13.3. A rekurzív Newton-interpoláció

Jelölje $N_n(x)$ az $x_0,x_1,\ldots,x_n$ alappontokra épített interpolációs polinomot (Newton- polinom). Ekkor ezen polinomok a következő rekurzióval számíthatók:

$$N_n(x)=N_{n-1}(x)+b_n\omega_{n-1}(x), \tag{12.7}$$

ahol a b_n együttható abból a feltételből határozható meg, hogy $N_n(x)$ interpolál az x_n pontban:

$$b_n=\frac{f_n-N_{n-1}(x_n)}{\omega_{n-1}(x_n)}.$$

Azonban a b_n-ek számítására van egy ennél sokkal egyszerűbb módszer.

13.3.1 Tétel

$$b_n=f[x_0,x_1,\ldots,x_n]. \tag{12.8}$$

Bizonyítás. Az $N_n(x)-N_{n-1}(x)$ kifejezés eltűnik az $x_0,\dots,x_{n-1}$ pontokban a definició szerint, így $\omega_{n-1}(x)$ szerepeltetése jogos, aminek az együtthatóját abból a feltételből határozzuk meg, hogy $N_n(x_n)=f(x_n)$. A levezetésben $N_{n-1}(x)$ helyére a megfelelő Lagrange-polinomot írjuk:

$$b_n=\frac{N_n(x_n)}{\omega_{n-1}(x_n)}-\frac{N_{n-1}(x_n)}{\omega_{n-1}(x_n)}=\frac{f(x_n)}{\omega_{n-1}(x_n)}-\sum_{j=0}^{n-1}\frac{f(x_j)\omega_{n-1}(x_n)}{\omega_{n-1}(x_n)(x_n-x_j)\omega'_{n-1}(x_j)}=$$

$$=\frac{f(x_n)}{\omega_{n-1}(x_n)}+\sum_{j=0}^{n-1}\frac{f(x_j)}{(x_j-x_n)\omega'_{n-1}(x_j)}=f[x_0,x_1,\dots,x_n],$$

ahol az utolsó sorban figyelembe vettük (12.6)-ot. Eszerint az osztott differenciák táblázatában a jobb szélső elemek adják a kifejtéshez szükséges b_n együtthatókat. ■

A *rekurzív* jelző nélkül egyszerűen Newton-interpolációról beszélünk akkor, ha a felépítésben a b_n együtthatók helyén az osztott differencákat használjuk. A rekurzív Newton interpoláció használata szokatlan helyzetekben előnyös, például, ha többváltozós interpolációs formulát akarunk készíteni, vagy magasabb deriváltakat is interpolálunk úgy, hogy egyes alacsonyabb rendűek hiányoznak.

13.4. Feladatok

1. Részletesen ellenőrízzük a 3.3 Lemma bizonyításának lépéseit!
2. Neville-interpolációval a függvényt az x helyen kívánjuk közelíteni. A táblázat minden sorában az utolsó szám egy interpoláló polinom helyettesítési értékét adja. Milyen sorrendben írjuk fel az interpoláció alappontjait, hogy a táblázatban az utolsó elemek egyre jobb, növekvő pontosságú sorrendet adjanak?
3. Neville-interpolációval határozzuk meg azt a másodfokú polinomot, amely átmegy a $(-1,0),(1,1),(2,6)$ pontokon! (Most x paraméterként bennmarad a formulákban.)
4. Mutassuk meg, a Neville-interpoláció akkor is ugyanazt az eredményt adja, ha az első oszlopba $x-x_i$ helyett az x_i-x értékeket írjuk!
5. Ugyanezt a polinomot állítsuk elő Newton interpolációval!
6. Milyen algoritmust javasoljunk a Newton-polinom helyettesítési értékeinek számítására, ha az osztott differenciák adottak?
7. Készítsünk Matlab programot, amely a Neville-interpoláció függvényérték közelítéseit adja egy vektorban!
8. Készítsünk Matlab programot, amely az osztott differenciák táblázatát készíti el!
9. Készítsünk Matlab programot, amely az osztott differenciák értékeit felhasználva a Newton-polinom helyettesítési értékét adja!
10. Állapítsuk meg a Newton-interpoláció bázisfüggvényeit és ezek segítségével írjuk fel az interpolációs feltétel lineáris egyenletrendszerét!
11. Oldjuk meg az előző feladatban kapott egyenletrendszert úgy, hogy az osztott differencia-séma lépéseit követjük! Mit tapasztalunk?
12. Írjuk fel azt a mátrixot, ami az $[f_0,f_1,\dots,f_n]^T$ függvényértékek vektorát az elsőrendű osztott differenciák $[f[x_0,x_1],\dots,f[x_{n-1},x_n]]^T$ vektorába viszi!

14. Newton- és Hermite-interpoláció

14.1. Tétel, osztott differenciával az interpoláció hibája

Legyen $x \in [a,b]$, $x \neq x_i$, $i = 0,1,\ldots,n$, ekkor

$$f(x) - L_n(x) = f[x, x_0, x_1, \ldots, x_n]\omega_n(x), \tag{13.1}$$

ahol $[a,b]$ az alappontok által kifeszített intervallum.

Bizonyítás. Legyen N_{n+1} olyan, hogy az x helyen felveszi $f(x)$ értékét. Felhasználva, hogy $N_n(x) = L_n(x)$, a Newton interpoláció szerint írhatjuk:

$$f(x) - L_n(x) = N_{n+1}(x) - N_n(x) = f[x, x_0, x_1, \ldots, x_n]\omega_n(x),$$

és ezzel kész is vagyunk, mert minden x-re N_{n+1}-et újraválaszthatjuk úgy, hogy $N_{n+1}(x) = f(x)$ teljesüljön. ■

14.1.1 Következmény

Legyen $f(x) \in C^{n+1}[a,b]$, $x \in [a,b]$, $x \neq x_i$, ekkor létezik $\xi_x \in [a,b]$, melyre

$$f[x, x_0, x_1, \ldots, x_n] = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \tag{13.2}$$

fennáll.

Ennek belátásához elegendő, ha összehasonlítjuk az 11.3.1 Tétel (10.9) formuláját (13.1)-gyel. Speciálisan $n = 0$-ra $f[x, x_0] = f'(\xi_x)$, és ez kiírva a Lagrange középérték-tétel. Így (13.2) nem más, mint a középérték-tétel általánosítása magasabbrendű osztott differenciákra. Figyeljük meg: (13.2) – ben x is formális változóként szerepel, $n+2$ alapponthoz tartozik $n+1$-edrendű osztott differencia, és az ehhez tartozó derivált $n+1$-edrendű.

14.1.2 További következmény

A (13.2) összefüggés alkalmas arra, hogy az osztott differenciák táblázatát arra az esetre is értelmezzük, amikor egy alappont többször szerepel. Az alappontok ξ_x-szel együtt az $[a,b]$ intervallumban helyezkednek el. Ha most $[a,b]$ az x_0 pontra zsugorodik össze, akkor határátmenettel kapjuk, hogy

$$f[\underbrace{x_0, x_0, \ldots, x_0}_{n+1 \text{ db. alappont}}] = \frac{f^{(n)}(x_0)}{n!}. \tag{13.3}$$

14.2. Hermite-interpoláció

Ha előírjuk, hogy az interpoláló polinom a függvény deriváltjaira is illeszkedjen a megadott pontokban, akkor Hermite-interpolációról beszélünk. Ilyenkor a tartópontok közé a deriváltakat is felvesszük:

$$(x_k, f^{(i)}(x_k)),\ i = 0,1,\ldots,m_k - 1),\ m_k \in \mathbb{N}_+ .$$

Például, ha $m_k = 2$, akkor a nulladik és az első derivált illeszkedik x_k-ban. Általában a feltételek száma:

$$\sum_{k=0}^{n} m_k = m+1, \tag{13.4}$$

tehát m-edfokú polinom lehetséges: $H_m(x) = \mathcal{P}_m$, és az illeszkedési feltételek:

$$H_m^{(i)}(x_k) = f^{(i)}(x_k), \quad i = 0,1,\ldots,m_k - 1; \quad k = 0,1,\ldots,n. \tag{13.5}$$

14.2.1 Tétel

Ha az alappontok különbözőek, a (13.5) illeszkedési feltételeknek eleget tevő $H_m(x)$ polinom létezik és egyértelmű.

Bizonyítás. Legyen a polinom $H_m(x) = \sum_{j=0}^{m} a_j x^j$ alakú és írjuk fel az együtthatókat meghatározó lineáris egyenletrendszert:

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^m \\ 0 & 1 & 2x_0 & \ldots & mx_0^{m-1} \\ \vdots & \ldots & \ldots & \ldots & \vdots \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f'(x_0) \\ \vdots \end{bmatrix},$$

ami most egy $(m+1)\times(m+1)$-es rendszer. Ennek van megoldása, ha a determinánsa $\det(A) \neq 0$. Indirekt módon tegyük fel, hogy mégis $\det(A) = 0$. Következik, hogy a homogén egyenletnek $(b = 0)$ van nemzérus megoldása, ami ekkor m-edfokú polinom. Vegyük észre, a zérus jobb oldal most azt jelenti, hogy x_k m_k-szoros gyöke a polinomnak. De akkor ennek a polinomnak $m+1$ gyöke kéne, hogy legyen, ami ellentmondás. Ezért az egyenletrendszer mátrixa invertálható, és a megoldás egyértelmű. ■

Megjegyzés. Ha a deriváltak hiányosan vannak megadva, akkor a hiányos (idegen szóval: *lakunáris*) Hermite-interpoláció feladata nem mindig oldható meg.

14.2.2 Hibatétel a nem-hiányos Hermite-interpolációra

Legyen $f(x) \in C^{m+1}[a,b]$, $x \in [a,b]$, ekkor létezik $\xi_x \in [a,b]$, amelyre

$$f(x) - H_m(x) = \frac{f^{(m+1)}(\xi_x)}{(m+1)!}\omega_m(x), \tag{13.6}$$

ahol most $\omega_m(x) = (x-x_0)^{m_0}(x-x_1)^{m_1}\ldots(x-x_n)^{m_n}$.

Bizonyítás. Az 11.3.1 Tételhez hasonlóan tesszük. Ha $x = x_k$, $k = 0,1,\ldots,n$, akkor az állítás igaz, így a továbbiakban legyen $x \neq x_k$ minden k-ra. Vezessük be most is a

$$g_x(z) = f(z) - H_m(z) - \frac{\omega_m(z)}{\omega_m(x)}(f(x) - H_m(x)), \quad z \in [a,b] \tag{13.7}$$

függvényt, amelynek $z = x$-szel együtt $m+2$ gyöke van. A tétel állításához a Rolle-tétel $(m+1)$-szeri alkalmazása után jutunk. ■

Abban a speciális esetben, ha minden alappontban a függvényérték és az első derivált adott, *Hermite-Fejér interpolációról* beszélünk.

Érdemes még szót ejteni arról, hogy a Newton-interpolációt hogyan alkalmazhatjuk az Hermite-interpolációnál. Az osztott differenciák értelmezését többszörös ugyanazon alappont esetére már (13.3)-ban megadtuk. Eszerint például kétszer adjuk meg az x_0 pontot, ha $f(x_0)$ és $f'(x_0)$ adottak. Az $\omega(x)$ szorzatfüggvénybe minden korábban alappontként figyelembe vett x_j pont $x-x_j$ tényezőt ad, az éppen interpolált pont csak a következő lépésben ad tényezőt. A pontok sorrendje tetszőleges, de a többször megadott pontok legyenek egymás mellett a deriváltak miatt. Ne feledjük, a deriváltak megadása nem lehet hiányos, például nem hiányozhat az első, ha adott a második derivált.

14.2.3 Példa

Newton interpolációval készítsük el azt a polinomot, amely az x_0, x_1 pontokra az Hermite-Fejér interpolációt valósítja meg!

Megoldás. A osztott differenciák táblázata:

	$k=0$	1	2	3
x_0	f_0			
x_0	f_0	f_0'		
x_1	f_1	$f[x_0,x_1]$	$(f[x_0,x_1]-f_0')/(x_1-x_0)$	
x_1	f_1	f_1'	$(f_1'-f[x_0,x_1])/(x_1-x_0)$	$(f_1'-2f[x_0,x_1]+f_0')/(x_1-x_0)^2$

A keresett polinom:

$$N_3(x)=f_0+f_0'(x-x_0)+\frac{f[x_0,x_1]-f_0'}{x_1-x_0}(x-x_0)^2+\frac{f_1'-2f[x_0,x_1]+f_0'}{(x_1-x_0)^2}(x-x_0)^2(x-x_1).$$

14.3. Hermite-interpolációs alappolinomok

A Lagrange-alappolinomokhoz hasonló tulajdonságú polinomok mindig megszerkeszthetők, ha nem hiányos a deriváltak megadása. Az x_k pontban az $f_k^{(i)}$, $i=0,1,\dots,m_k-1$-edik deriváltak adottak. Vezessük be az x_k helyhez tartozó

$$h_k(x)=\prod_{j=0,\,j\neq k}^{n}\left(\frac{x-x_j}{x_k-x_j}\right)^{m_j},\quad h_k(x_k)=1$$

függvényt. Ennek a $0,1,\dots,m_j-1$-edik deriváltja eltűnik az $x_j(\neq x_k)$ pontokban, még akkor is, ha meg volna szorozva egy másik polinommal. Így $h_k(x)$ az $x_j(\neq x_k)$ pontokban már teljesíti az alappolinomtól elvárt tulajdonságokat. Az x_k ponthoz tartozó alappolinomokat keressük a következő alakban:

$$l_{k,i}(x)=p_{k,i}(x)h_k(x),\quad p_{k,i}(x)\in\mathcal{P}_{m_k-1},$$

ahol $p_{k,i}(x)$ m_k-1-edfokú polinom, $i=0,1,\dots,m_k-1$, és az i-edik polinom együtthatóit abból a feltételből kapjuk, hogy $(d/dx)^j l_{k,i}(x_k)=\delta_{ij}$, $j=0,1,\dots,m_k-1$.

A könnyebb érthetőség kedvéért tekintsük azt a példát, amikor x_k-ban a második deriváltig adottak az értékek, azaz $i=0,1,2$ lehet. A polinomokat $x-x_k$ hatványai szerint célszerű felírni. Ha $i=0$, akkor

$p_{k,0}(x)=1+\alpha_1(x-x_k)+\alpha_2(x-x_k)^2$ alakú, mert $h_k(x_k)=1$ és $l_{k,0}(x_k)=1$ miatt $p_{k,0}(x_k)=1$. α_1-et abból a feltételből határozzuk meg, hogy az első derivált zérus az x_k helyen:

$$[\alpha_1+2\alpha_2(x-x_k)]h_k(x)+p_{k,0}(x)h_k'(x)\Big|_{x=x_k}=0,$$

innen $\alpha_1=-h_k'(x_k)$. Ha a második deriváltat is zérussá tesszük:

$$2\alpha_2 h_k(x_k)+2\alpha_1 h_k'(x_k)+h_k''(x_k)=0,$$

α_1 értékét beírva $\alpha_2=\left(h_k'(x_k)\right)^2-h_k''(x_k)/2$ az eredmény.

A $p_{k,2}(x)$ polinom nulladfokú tagja zérus, mert $p_{k,2}(x_k)h_k(x_k)=0$. Hasonló ok miatt az elsőfokú tag együtthatója is zérus lesz, mert $p_{k,2}'(x_k)h_k(x_k)+p_{k,2}(x_k)h_k'(x_k)=0$. Végül $l_{k,2}''(x_k)=1$-ből $p_{k,2}(x)=(x-x_k)^2/2$ adódik. Gyakorlásképp igazoljuk, hogy a fenti példában $p_{k,1}(x)=(x-x_k)+\alpha_1(x-x_k)^2$!

Így a nem-hiányos Hermite-interpolációt a következő formula állítja elő:

$$H_m(x)=\sum_{k=0}^{n}\sum_{i=0}^{m_k-1}f_k^{(i)}l_{k,i}(x).$$

Vegyük észre, a kapott előállítás egy újabb igazolását adja annak, hogy a nem-hiányos Hermite-interpoláció feladata egyértelműen megoldható.

14.4. Inverz interpoláció

Errről beszélünk, ha a függő és független változókat felcseréljük: Így $x=x(y)$ típusú polinomot kapunk. A technikát akkor alkalmazzuk, ha arra vagyunk kíváncsiak, hogy a függvény egy adott értéket mely helyen vesz fel, például, ha a függvény gyökét keressük. Ekkor a közelítő polinomba $y=0$-t helyettesítve a gyök helyére kapunk egy közelítést.

14.5. Feladatok

1. Származtasssunk interpolációs formulát, amikor a tartópontok: (x_0,f_0), (x_1,f_1,f_1',f_1'').
2. A szinusz függvény egyenletes tabellázásakor a deriváltja is ismert a koszinusz függvénnyel való ismert összefüggés miatt, így Hermite-Fejér interpolációt alkalmazhatunk két alappont között. Milyen sűrűn kell egyenletesen tabellázni a függvényt $[0,\pi/2]$-ben, ha mindenütt 10^{-4} hibával szeretnénk a függvény értékeit megkapni?
3. Írjuk fel az Hermite-Fejér interpolációhoz tartozó hibaformulát, ha az interpoláció a Csebisev alappontokon történik.
4. Mutassuk meg, hogy Hermite-Fejér interpolációnál $h_k(x)=\left(l_k(x)\right)^2$, $p_{k,0}(x)=1-2l_k'(x_k)(x-x_k)$ és $p_{k,1}(x)=x-x_k$.
5. Melyek az Hermite-interpolációs bázispolinomok, ha a tartópontok: (x_0,f_0,f_0'), és (x_1,f_1,f_1')?
6. Készítsük el az Hermite-interpoláció bázispolinomjait, ha a tartópontok: (x_0,f_0,f_0''), és (x_1,f_1,f_1''). Bár hiányos a deriváltak megadása, a bázispolinomok mégis léteznek.
7. Egy függvényhez tartozó négy pont: $(1,-1)$, $(2,1)$, $(3,2)$, $(5,3)$. Inverz interpolációt választva határozzuk meg a gyök közelítését Neville-interpolációval!

8. Általánosítsuk a Neville-interpolációt Hermite-interpoláció esetére!

9. Legyen $f(x)=\sum_{j=0}^{n} a_j x^j$ n-edfokú polinom. $f[x_0,x_1,\ldots,x_n]=?$

10. Igazoljuk: $\exists\, \xi\in[x_0,x_1]$: $\frac{1}{n+1}\sum_{j=0}^{n} x_0^j x_1^{n-j}=\xi^n$, ahol $n=1$-re ξ a számtani közép.

15. Interpoláció spline (donga-) függvényekkel

15.1. Spline- vagy dongafüggvények

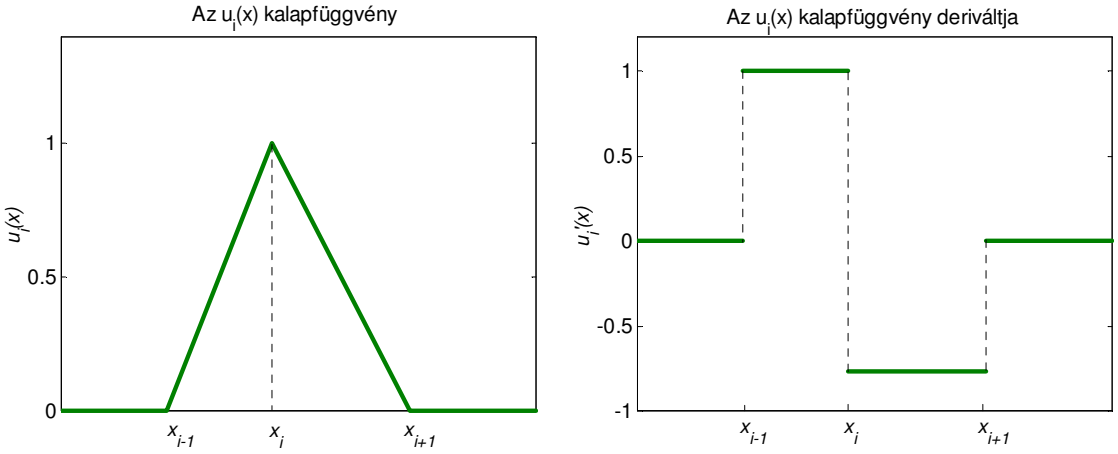
Ha az interpoláló polinomok fokszámát növeljük, gyakran tapasztalhatjuk, hogy a magasabb fokszámú polinomok erőteljes hullámzást mutatnak az alappontok környezetében. Olyannyira, hogy ránézésre sem hihető, hogy a függvényt jól közelítik.

A *spline* függvényekkel történő interpolációnál az ötlet az, hogy az egyes részintervallumokban csak alacsony fokszámú polinomokat engedünk meg, az intervallum-határokon pedig a polinomokat folytonosan illesztjük. Lehetőség szerint a folytonosságot a deriváltakra is előírjuk.

Spline [ejtsd: 'szplájn'] angolul dongát jelent, ami azokat a fabordákat jelenti, amikkel a kádár kirakja a hordó alakját. Spline-oknak hívják angol nyelvterületen azokat a hajlítható, görbíthető 'vonalzókat' is, amelyekkel görbe vonal rajzolható.

Itt most matematikailag hasonló dolog történik: az egyes részintervallumokban függvényíveket polinomokból készítünk, amelyeket „összevarrunk" az intervallum-határokon valamilyen folytonossági követelmény szerint. Szemléletesen szólva: *dongafüggvényeket* alkalmazunk.

A továbbiakban legyenek $\Theta_n = \{x_i, f_i = f(x_i)\}_{i=0}^{n}$ a tartópontok, az abszcisszák legyenek nagyság szerint rendezettek: $x_{i-1} < x_i$, $0 < i$ és $S_l(\Theta_n)$ jelölje az l-edfokú spline-ok halmazát. Ez azt jelenti, hogy $S_l(\Theta_n)$ elemei minden $[x_{i-1}, x_i]$ részintervallumon l-edfokú polinomok. A spline interpoláció egyszerűen tárgyalható az $u_i(x)$ *kalapfüggvények* segítségével:



1. ábra. A kalapfüggvény és deriváltja

$$u_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & \text{ha } x_{i-1} \le x \le x_i \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i}, & \text{ha } x_i \le x \le x_{i+1} \\ 0, & \text{egyébként} \end{cases}$$

$$u_i'(x) = \begin{cases} \dfrac{1}{x_i - x_{i-1}}, & \text{ha } x_{i-1} < x < x_i \\ \dfrac{-1}{x_{i+1} - x_i}, & \text{ha } x_i < x < x_{i+1} \\ 0, & \text{ha } x < x_{i-1},\ x_{i+1} < x \end{cases}$$

Az első derivált az alappontokban nem létezik, csak az alsó és felső határértékük. Később ennyi számunkra elég lesz. A kalapfüggvény magasabbrendű deriváltjai mind eltűnnek.

15.2. Elsőfokú spline-ok: $s(x)\in S_1(\Theta_n)$

$$s(x)=f_{i-1}\frac{x_i-x}{x_i-x_{i-1}}+f_i\frac{x-x_{i-1}}{x_i-x_{i-1}},\quad x\in[x_{i-1},x_i]. \tag{14.1}$$

Az eredmény egy törtvonal. A számítógép nagyon gyakran így rajzolja ki a Θ_n halmazzal adott függvényt. Ha a felbontás elég sűrű, akkor nem annyira szembetűnő a vonalak törése. A kalapfüggvények segítségével ez az $s(x)$ függvény Lagrange-alappolinom stílusban így írható:

$$s(x)=\sum_{i=0}^{n}f_iu_i(x) \tag{14.2}$$

15.3. Másodfokú spline-ok: $s(x)\in S_2(\Theta_n)$

Az egyszerűség kedvéért tegyük fel, hogy a kezdőpontban adott $f(a)$ és $f'(a)$. Ekkor az

$x_0=a$	x_0	x_1
$f(x_0)$	$f'(x_0)$	$f(x_1)$

pontokra készíthetünk Hermite-interpolációval egy másodfokú polinomot, amely az első intervallumhoz tartozik. E polinom x_1-ben felvett deriváltjával és $f(x_1)$-gyel folytassuk az eljárást ugyanígy az $[x_1,x_2]$ intervallumra. Általánosan $[x_i,x_{i+1}]$-ben a Newton interpoláció segítségével készített polinom táblázata:

$$\begin{array}{llll} x_i & f(x_i) & & \\ x_i & f(x_i) & f'(x_i) & \\ x_{i+1} & f(x_{i+1}) & f[x_i,x_{i+1}] & \dfrac{f[x_i,x_{i+1}]-f'(x_i)}{x_{i+1}-x_i}, \end{array}$$

és

$$s(x)=f(x_i)+f'(x_i)(x-x_i)+f[x_i,x_i,x_{i+1}](x-x_i)^2,\quad x\in[x_i,x_{i+1}],$$

ahol a függvény deriváltját mindig az előző intervallumban készített polinomból kapjuk.

Ha az induló derivált nem ismert, az első intervallum polinomját vehetjük lineárisnak. Egy másik lehetőség az első három ponton átvetett parabolával kezdeni, majd az eljárást a megismert módon folytatni.

15.4. Harmadfokú spline-ok: $s(x)\in S_3(\Theta_n)$

Most az interpolációt egy az $[x_0,x_1]$ intervallumban készítendő hiányos Hermite-interpolációval kezdjük.

Az $\{x_0,f_0,f_0''\}$, $\{x_1,f_1,f_1''\}$ tartópontokhoz tartozó harmadfokú Hermite-interpolációs alappolinomokat jelölje $l_{ki}(x)$. Az első index az abszcissza indexére utal, a második pedig a derivált rendjére. Az első polinomra teljesül: $l_{00}(x_0)=1$, $l_{00}(x_1)=0$, $l_{00}''(x_0)=0$ és $l_{00}''(x_1)=0$. A második derivált első vagy alacsonyabb fokú és mindkét pontban eltűnik, emiatt csak az azonosan 0 függvény lehet. Következésképp $l_{00}(x)$ elsőfokú és az x_0 és x_1 helyen felvett értékei teljesen meghatározzák:

$$l_{00}(x)=(x_1-x)/(x_1-x_0)=u_0(x),\quad x\in[x_0,x_1]. \tag{14.3}$$

Az $l_{02}(x)$ polinomot meghatározó összefüggések: $l_{02}(x_0)=0,\ \ l_{02}(x_1)=0,\ \ l''_{02}(x_0)=1$ és $l''_{02}(x_1)=0$. A második derivált elsőfokú és a felvett értékei alapján $l''_{02}(x)=u_0(x),\ x\in[x_0,x_1]$. Ezt kétszer integrálva

$$l'_{02}(x)=\frac{u_0^2(x)}{2u'_0(x)}+\beta,$$

$$l_{02}(x)=\frac{u_0^3(x)}{6u_0'^2(x)}+\beta\frac{u_0(x)}{u'_0(x)}+\gamma,\ \ x\in[x_0,x_1]$$

ahol kihasználtuk, hogy $u'_0(x)$ az intervallumban konstans. Mivel $u_0(x_1)=0$, emiatt $l_{02}(x_1)=0$-ból $\gamma=0$ következik. Az $l_{02}(x_0)=0$ feltételből pedig $\beta=-1/\left(6u'_0(x)\right)$ adódik, így

$$l_{02}(x)=\frac{u_0^3(x)-u_0(x)}{6u_0'^2(x)}. \tag{14.4}$$

Az $l_{10}(x)$ és $l_{12}(x)$ polinomok meghatározása teljesen hasonlóan történik, az eredmény:

$$l_{10}(x)=u_1(x),\quad l_{12}(x)=\frac{u_1^3(x)-u_1(x)}{6u_1'^2(x)},\quad x\in[x_0,x_1]. \tag{14.5}$$

Ezekkel az $[x_0,x_1]$ intervallumban interpoláló harmadfokú polinom:

$$p_{0,3}(x)=f_0l_{00}(x)+f''_0l_{02}(x)+f_1l_{10}(x)+f''_1l_{12}(x)=$$

$$=f_0u_0(x)+f''_0\frac{u_0^3(x)-u_0(x)}{6u_0'^2(x)}+f_1u_1(x)+f''_1\frac{u_1^3(x)-u_1(x)}{6u_1'^2(x)}.$$

Most tegyük fel, a függvényérték és a második derivált az $x_0,x_1,\ldots,x_n$ alappontokban adott. Akkor e formula mintájára minden intervallumban fel tudunk írni egy harmadfokú interpoláló polinomot:

$$p_{i,3}(x)=f_iu_i(x)+f''_i\frac{u_i^3(x)-u_i(x)}{6u_i'^2(x)}+f_{i+1}u_{i+1}(x)+f''_{i+1}\frac{u_{i+1}^3(x)-u_{i+1}(x)}{6u_{i+1}'^2(x)},\quad x\in[x_i,x_{i+1}]$$

de a kalapfüggvények definícióját szem előtt tartva ezt a polinom-sereget egy szummával is megadhatjuk a teljes intervallumra:

$$p_{[a,b],3}(x)=\sum_{i=0}^{n}f_iu_i(x)+f''_i\frac{u_i^3(x)-u_i(x)}{6u_i'^2(x)},\quad x\in[a,b]. \tag{14.6}$$

Így olyan függvényünk van, amely minden részintervallumban harmadfokú polinom, és az alappontokban folytonos a függvény és a második deriváltja.

A látottak alapján a Θ_n tartópontokra a harmadfokú dongafüggvényt a következőképp vesszük fel:

$$s_3(x)=\sum_{i=0}^{n}f_iu_i(x)+s_i\frac{u_i^3(x)-u_i(x)}{6u_i'^2(x)},\quad x\in[a,b],\quad s_i=s''_3(x_i). \tag{14.7}$$

Az s_i második deriváltakat abból a feltételből határozzuk meg, hogy az első deriváltak is legyenek folytonosak az alappontokban. Az első derivált

$$s'_3(x)=\sum_{i=0}^{n}f_iu'_i(x)+s_i\frac{3u_i^2(x)-1}{6u'_i(x)},\quad x\in[a,b] \tag{14.8}$$

csak akkor lesz folytonos, ha minden alappontban az alsó és felső határérték megegyezik. Jelölje $s'_3(x_i-)$ az alsó, $s'_3(x_i+)$ a felső határértéket az x_i helyen. Ekkor csak két kalapfüggvény ad nemzérus járulékot a határértékekhez:

$$s_3'(x_i-) = f_{i-1}u_{i-1}'(x_i-) + f_i u_i'(x_i-) + s_{i-1}\frac{3u_{i-1}^2(x_i)-1}{6u_{i-1}'(x_i-)} + s_i\frac{3u_i^2(x_i)-1}{6u_i'(x_i-)} =$$

$$= \frac{-f_{i-1}+f_i}{h_{i-1}} + \frac{s_{i-1}+2s_i}{6}h_{i-1}, \quad \text{ahol } h_{i-1} = x_i - x_{i-1}$$

és

$$s_3'(x_i+) = \sum_{j=i}^{i+1} f_j u_j'(x_i+) + s_j\frac{3u_j^2(x_i)-1}{6u_j'(x_i+)} = \frac{f_{i+1}-f_i}{h_i} - \frac{s_{i+1}+2s_i}{6}h_i.$$

A kettőt egyenlővé téve az x_i pontban:

$$f[x_{i-1},x_i] + \frac{2s_i+s_{i-1}}{6}h_{i-1} = f[x_i,x_{i+1}] - \frac{2s_i+s_{i+1}}{6}h_i, \quad h_i = x_{i+1}-x_i.$$

Ez tovább rendezve

$$\frac{s_{i-1}h_{i-1}}{6} + \frac{s_i(h_{i-1}+h_i)}{3} + \frac{s_{i+1}h_i}{6} = f[x_i,x_{i+1}] - f[x_{i-1},x_i],$$

majd a $\sigma_{i-1} = h_{i-1}/(h_{i-1}+h_i)$ jelölés bevezetésével az

$$s_{i-1}\sigma_{i-1} + 2s_i + s_{i+1}(1-\sigma_{i-1}) = 6f[x_{i-1},x_i,x_{i+1}], \quad i = 1,2,\ldots,n-1. \tag{14.9}$$

alakra egyszerűsödik.

A kapott háromátlójú lineáris egyenletrendszer mátrixa diagonáldomináns, ami a megoldás szempontjából kedvező. Azonban az egyenletrendszer nem határozza meg s_0 és s_n értékét, így a kezdő- és végpontban még feltételeket kell előírnunk. A gyakorlatban az alábbi három megoldás valamelyikét szokták választani:

1. Hermite-peremfeltétel: az első deriváltak $f'(a)$ és $f'(b)$ adottak.
2. A második derivált értékéről rendelkezünk a kezdő és végpontban. Ha nem ismerjük, $s_0 = s_n = 0$ egy lehetséges választás. Hagyományosan ezt nevezik természetes spline-nak. Ennél azonban jobb megoldás, ha a szélső két intervallumban a harmadik deriváltat konstansnak vesszük: az így kapott s_0 és s_n-et tartalmazó kifejezéseket hozzávéve az (14.9) rendszerhez kapjuk azt a spline-interpolációt, amely harmadfokú polinomig pontos.
3. Periodikus határfeltétel. Ha a függvény periodikus, és a teljes periódusban történik az interpoláció, akkor a függvényt és deriváltjait a két végpontban egyenlővé tesszük.

15.5. Példa

Az Hermite-féle peremfeltétel mellett hogy fog kinézni a megoldandó (14.9) egyenletrendszer első és utolsó sora?

Megoldás. (14.9)-ben vegyünk $i = 0$-t és az x_{-1} formálisan felvett alapponttal tartsunk x_0-hoz. Ekkor $\sigma_{-1} \to 0$ és így

$$2s_0 + s_1 = 6f[x_0,x_{0,}x_1] = 6\left(f[x_0,x_1] - f_0'\right)/h_0. \tag{14.10}$$

Az utolsó egyenlethez helyettesítsük (14.9)-be $i = n$-et és x_{n+1} tartson x_n-hez:

$$s_{n-1} + 2s_n = 6f[x_{n-1},x_n,x_n], \tag{14.11}$$

E két egyenlettel kiegészítve (14.9)-et már annyi egyenletünk van, mint az ismeretlenek száma.

15.6. Feladatok

1. Hogy egyszerűsödik (14.9), ha az alappontok egyenlő távolságra vannak egymástól?
2. Hogy módosul (14.9) első és utolsó sora, ha a szelső két intervallumban a harmadik deriváltat tesszük állandóvá? (Útmutatás: pl. az x_0, x_1 és x_1, x_2 pontok között képezzük differencia-hányadossal a harmadik deriváltakat és tegyük őket egyenlővé: $(s_2 - s_1)/h_1 = (s_1 - s_0)/h_0$. Az utolsó három pontnál járjunk el hasonlóan. A kapott eredményt helyettesítsük a megfelelő egyenletbe.)
3. Mutassuk meg, hogy legalább négy tartópontnál az előbbi módon készített spline harmadfokú polinomra pontos, azaz annak pontjaiból visszakapjuk magát a polinomot.
4. A harmadfokú dongafüggvényt úgy is reprezentálhatjuk, hogy egy intervallumon belül a polinomot az intervallum határain felvett függvényértékkel és az első deriválttal adjuk meg. Az Hermite-interpolációs alappolinomokkal készítsük el az interpoláló formulát az (x_0, x_1) intervallumra, ld. 14.5.5 példa. Igazoljuk, hogy (14.6)-hoz hasonlóan a következő formula nyerhető:

$$\tilde{p}_{[a,b],3}(x) = \sum_{i=0}^{n} f_i \left(-2u_i^3(x) + 3u_i^2(x)\right) + f_i' \frac{u_i^3(x) - u_i^2(x)}{u_i'(x)}, \quad x \in [a,b].$$

5. Ha a harmadfokú dongafüggvényt a részintervallumok határain felvett függvényértékkel és az első deriválttal reprezentáljuk, akkor az előző feladat alapján a következő kifejezést kapjuk:

$$s_3(x) = \sum_{i=0}^{n} f_i \left(-2u_i^3(x) + 3u_i^2(x)\right) + t_i \frac{u_i^3(x) - u_i^2(x)}{u_i'(x)}, \quad x \in [a,b],$$

   ahol $f_i = s_i(x_i)$, $t_i = s_i'(x_i)$ és $u_i(x)$ az i-edik kalapfüggvény. A részintervallumok határain a második derivált egyeztetésével származtassunk egyenletrendszert a t_i paraméterek meghatározására!
6. Szeretnénk harmadfokú spline függvénnyel közelíteni a következő differenciálegyenlet megoldását: $-v''(x) = g(x), \quad x \in [0,1], \quad v(0) = v(1) = 0,$ ahol $g(x)$ megadott függvény. Osszuk fel a $[0,1]$ intervallumot n egyenlő részre és írjuk fel (14.9) felhasználásával a közelítést meghatározó egyenleteket. A differenciálegyenletből nyerhető információ alapján alakítsuk át az egyenletrendszert, hogy megoldásul a $v(x_i), \ i = 1, \ldots, n-1$ értékek közelítéseit kapjuk!
7. Az interpolációnál tanultak alapján adjunk felső becslést az elsőfokú spline közelítés hibájára!

16. Nemlineáris egyenletek megoldása I.

Eddig lényegében lineáris egyenletrendszerek megoldásával foglalkoztunk. De sokszor felvetődik az

$$f(x)=0 \tag{16.1}$$

egyenlet egy (vagy esetleg több) gyökének keresése, ahol az $f(x)\in C[a,b]$ egyváltozós függvény. Minden olyan x^* érték, amelyre $f(x^*)=0$, a (16.1) egyenlet *gyöke* vagy $f(x)$ *zérushelye*. A gyök az x^* helyen m*-edrendű*, ha $f(x)=(x-x^*)^m g(x),\ g(x^*)\neq 0$ alakban írható. Azzal az esettel foglalkozunk, amikor a megoldás közelítése valamilyen numerikus módszer segítségével végezhető.

16.1. A gyököt tartalmazó intervallum

Ha a függvény előjelet vált: $f(a)f(b)<0$, akkor a folytonosság miatt legalább 1 gyök található $[a,b]$-ben. Ha létezik $f(x)$ első deriváltja is, és előjeltartó $[a,b]$-ben, akkor csak 1 gyök van.

Ha a függvény nem monoton, akkor $[a,b]$-t célszerű olyan részintervallumokra bontani, ahol az intervallum két végpontja között előjelváltás van. Ílymódon $f(x)$ páratlan gyökeit el tudjuk különíteni. Deriválható függvény esetén a páros gyököket kereshetjük $f'(x)$ gyökeiként, mert ekkor a párosakat páratlanná tettük, de kereshetjük $f(x)/f'(x)$ gyökeit is, amelyek mind egyszeresek.

16.2. Fixpont iteráció

Egy lehetséges eljárás, hogy megpróbáljuk az $f(x)=0$ egyenletet fixpont-egyenletté alakítani:

$$x=F(x). \tag{16.2}$$

Példa: legyen a megoldandó egyenlet: $x^2-\sin(x)=0$. Ekkor próbálkozhatunk az $x_{k+1}=\sqrt{\sin(x_k)}$ iterációval. Fixpont-egyenletet mindig tudunk készíteni, hiszen $x=x+cf(x)$ is ilyen, ahol c nemzérus állandó, de választhatunk valamely $c(x)$ függvényt is olymódon, hogy az iteráció konvergencia-tulajdonságai javuljanak. A fixpont létezéséről szól a

16.2.1 Brouwer fixponttétel[1]

Ha $F(x)$ folytonos $[a,b]$-ben és $F:[a,b]\to[a,b]$, akkor létezik fixpontja.

Bizonyítás. Legyen $g(x)=x-F(x)$, ekkor $g(a)\le 0$ és $g(b)\ge 0$, amiből $g(a)g(b)\le 0$. Ha itt egyenlőségjel érvényes, akkor már van egy gyök, ha pedig a $<$ jel érvényes, akkor a folytonosság miatt kell léteznie gyöknek $[a,b]$-ben. ■

16.2.2 Tétel, kontrakció

Ha $F:S\to\mathbb{R}$ folytonosan differenciálható az S zárt intervallumon és $|F'(x)|<1,\ \forall x\in S$, akkor F kontrakció.

Bizonyítás. A Lagrange középérték tétel alapján $x,y\in S$-re $\exists\zeta:\ F(x)-F(y)=F'(\zeta)(x-y)$. Térjünk át az abszolút értékre és használjuk fel, hogy $\exists|F'(x)|$ maximuma S-ben:

---

[1] A tétel többdimenziós megfogalmazása: ha a folytonos $F(x)$ függvény a gömböt önmagába képezi le, akkor van fixpontja.

$$|F(x)-F(y)| \le \max_{x\in S}|F'(x)|\,|x-y|, \quad x,y\in S$$

tehát F kontrakció $q=\max_{x\in S}|F'(x)|<1$ kontrakciós állandóval. ■

16.2.3 Következmény

Ha $F(x)$ kontrakció, akkor a Banach fixponttétel szerint csak egy gyök van és a kontrakciós állandó ismeretében a közelítés pontosságát is becsülni tudjuk.

Visszatérve a fenti példához: a kapott iteráció biztosan konvergens abban a tartományban, ahol $\left(\sqrt{\sin(x)}\right)' = \dfrac{\cos(x)}{2\sqrt{\sin(x)}} < 1$. Látjuk, ha $x=0$ vagy $x=\pi$, ezzel a kifejezéssel baj van, mert a formula kiértékelésekor 0-val kéne osztani. De például $x=\pi/4$ esetén a kifejezés értéke $1/\sqrt{8}$, ami már jobbnak tűnik. Ha megrajzoljuk az x^2 parabolát és a $\sin(x)$ függvény képét, látjuk, hogy két nemnegatív gyök van, az egyik a zérus, a másik pedig közel $x=\pi/4$-hez, tehát remélhető, hogy az iteráció $\pi/4$-gyel indítva konvergens. De az is látszik, hogyha nagyon kicsi pozitív értékkel indítjuk az iterációt, akkor sem kapjuk meg a zérus gyököt, mert az iteráció mindig elvisz a nagyobbik gyök irányába.

Ha azonban az $x=\arcsin(x^2)$ iterációt készítjük, könnyen meggyőződhetünk arról, hogy kis pozitív x-re zérushoz tart. Ha azonban $x=1$-gyel indítunk, akkor először a $\pi/2$ értéket kapjuk, majd komplex számokat, mivel az argumentum nagyobb 1-nél.

A tanulság: ügyelnünk kell, a kapott függvény hova képez le, és a leképezés tartományában megmaradnak-e a konvergencia tulajdonságok, illetve azt a gyököt kapjuk-e, amit szeretnénk meghatározni.

Ha a megoldandó egyenletben több helyen is szerepel x, akkor több $x=F(x)$ kifejezés is készíthető. Például szerepeljen két helyen, ekkor meg lehet mutatni: a kapott két iteráció egy adott helyen egyszerre nem lehet konvergens. Legyen ugyanis $f(x_1,x_2)=0$ a megoldandó egyenlet, ahol a két előfordulást x_1 és x_2-vel azonosítjuk. Legyen $F_i(x)$ az az iterációs függvény, amelyet x_i kifejezésével kaptunk. Ez azt jelenti, hogy

$$f(F_1(x),x)=0 \quad \text{és} \quad f(x,F_2(x))=0. \tag{16.3}$$

Legyen α egyszeres gyök: $f(\alpha,\alpha)=0$. Ha a (16.3)-ben szereplő kifejezéseket deriváljuk x szerint és helyettesítjük $x=\alpha$-t, az eredmény:

$$f_1'(\alpha,\alpha)F_1'(\alpha)+f_2'(\alpha,a)=0,$$
$$f_1'(\alpha,\alpha)+f_2'(\alpha,\alpha)F_2'(\alpha)=0,$$

ahol f alsó indexe azt jelöli, melyik hely szerint deriváltunk. Ahhoz, hogy egyszeres gyök mellett nemzérus megoldást kapjunk $f_i'(\alpha,\alpha)$-ra kell, hogy a kapott rendszer determinánsa 0 legyen:

$$\begin{vmatrix} F_1'(\alpha) & 1 \\ 1 & F_2'(\alpha) \end{vmatrix} = F_1'(\alpha)F_2'(\alpha)-1=0,$$

ahonnan

$$|F_1'(\alpha)| = 1/|F_2'(\alpha)|. \tag{16.4}$$

Hacsak nem 1 abszolút értékűek a deriváltak, a gyök közelében az egyik iterációs függvény konvergens, a másik meg divergens lesz. Hasonlóan lehet vizsgálni azt az esetet, amikor x 2-nél

többször fordul elő. De ekkor a helyzet rosszabb, az is lehetséges, hogy egyik iterációs függvény sem konvergens. Emiatt azt célszerű tennünk, hogy az x-ek előfordulását két csoportba osztjuk és x-et az egyik csoportból teljesen kifejezük. Például $3x^2 - 2x + \exp(2.2x) + 1 = 0$-nél az iterációs függvényre kereshetjük a másodfokú polinom gyökeit úgy, hogy a konstans tag helyére $\exp(2.2x) + 1$-et írunk.

16.3. A konvergencia-sebesség

Legyen az x_n sorozat konvergens, $\lim_{n\to\infty} x_n = x^*$. Jelölje $\varepsilon = x_n - x^*$ az n-edik hibát. Ekkor, ha létezik c állandó és $p \geq 1$ szám úgy, hogy

$$|\varepsilon_{n+1}| \leq c|\varepsilon_n|^p, \quad n = 0,1,\ldots, \tag{16.5}$$

akkor az x_n sorozat konvergenciája p*-edrendű*. Ha

- $p = 1$, akkor a konvergencia *lineáris* vagy *elsőrendű*,
- $1 < p < 2$, akkor a konvergencia *szuperlineáris*,
- $p = 2$, akkor *kvadratikus* vagy *másodrendű*,
- $p = 3$, akkor *köbös*, vagy *harmadrendű*.

A p szám jellemzi az iterációs módszer konvergenciájának sebességét. Ha például $p = 2$, akkor ez nagyjából azt jelenti, hogy lépésenként az értékes jegyek száma megduplázódik.

A fixpont iteráció nem rendelkezik ezzel a sebességgel. Megmutatjuk, hogy $p = 1$, azaz a konvergenciája elsőrendű, amennyiben $|F'(x^*)| \neq 0$. Ugyanis

$$|\varepsilon_{n+1}| = |x_{n+1} - x^*| = |F(x_n) - F(x^*)| \leq q|x_n - x^*| = q|\varepsilon_n|. \tag{16.6}$$

Ha $F'(x^*) = 0$, akkor a konvergencia magasabb rendű. Erre vonatkozik a következő

16.3.1 Tétel

Legyen F valós függvény: $F(S) \subset S \subset \mathbb{R}$, S zárt. Tegyük fel, $F \in C^m(S)$ és $F^{(k)}(x^*) = 0$, $k = 1,2,\cdots,m-1$. Ekkor az F által meghatározott iteráció konvergencia-sebessége $p = m$-edrendű.

Bizonyítás. Az x^* körüli Taylor-polinom m-edrendű maradéktaggal

$$F(x) = F(x^*) + F'(x^*)(x - x^*) + \ldots + \frac{F^{(m-1)}(x^*)}{(m-1)!}(x - x^*)^{m-1} + \frac{F^{(m)}(\xi_x)}{m!}(x - x^*)^m$$

ahol a feltevés szerint az első, második, … , $m-1$-edik derivált eltűnik. Helyettesítsük $x = x_n$-et, vegyük figyelembe, hogy $x^* = F(x^*)$ és $x_{n+1} = F(x_n)$, ezzel

$$x_{n+1} - x^* = \frac{F^{(m)}(\xi_x)}{m!}(x_n - x^*)^m,$$

ahonnan

$$|\varepsilon_{n+1}| = \frac{|F^{(m)}(\xi_x)|}{m!}|x_n - x^*|^m \leq \frac{M_m}{m!}|\varepsilon_n|^m, \quad n = 0,1,\ldots$$

ahol $M_k = \max_{x\in S}|F^{(k)}(x)|$. Innen látható, a konvergencia m-edrendű. ■

16.4. Newton-iteráció (Newton-Raphson módszer) és a szelőmódszer

Ha a függvény első deriváltja létezik a gyök környezetében, akkor a gyököt közelíthetjük úgy, hogy az x_n pontban a függvényhez húzott érintő metszéspontját vesszük az x tengellyel. Ez ugyanaz, mint amikor az x_n körüli elsőfokú Taylor-polinomot zérussá tesszük és x_{n+1}-re megoldjuk:

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n),$$

innen a Newton-Raphson módszer iterációs formulája:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = F(x_n). \tag{16.7}$$

A *szelőmódszert* ebből úgy nyerjük, hogy a derivált helyére az utolsó két pontra felírt osztott differenciát tesszük:

$$x_{n+1} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})} = \frac{f(x_n)x_{n-1} - f(x_{n-1})x_n}{f(x_n) - f(x_{n-1})} = F(x_{n-1}, x_n), \tag{16.8}$$

tehát ez az iterációs függvény két pontra támaszkodik. A módszer előnye a Newton-módszerrel szemben, hogy nem kell hozzá a derivált, amit néha eléggé körülményes kiszámítani. Hátránya pedig a kisebb konvergencia-sebesség.

16.4.1 Tétel, a szelőmódszer hibája

Legyen $f(x) \in C^2[x_{n-1}, x_n, x^*]$, ekkor a szelőmódszernél az $n+1$-edik iterált hibájára fennáll

$$\varepsilon_{n+1} = \varepsilon_n \varepsilon_{n-1} \frac{f''(\xi)}{2f'(\eta)}, \quad \xi, \eta \in [x^*, x_{n-1}, x_n], \tag{16.9}$$

8. *ahol x^* a zérushely és $[x^*, x_{n-1}, x_n]$ az adott pontok által lefedett intervallum.*

9. Bizonyítás. Az állítást (16.8)-ból osztott differenciák segítségével származtatjuk. Kihasználjuk, hogy $f(x^*) = 0$:

$$\varepsilon_{n+1} = x_{n+1} - x^* = x_n - x^* - (x_n - x^*)\frac{f(x_n) - f(x^*)}{x_n - x^*}\frac{1}{f[x_{n-1}, x_n]} = \varepsilon_n\left(1 - \frac{f[x^*, x_n]}{f[x_{n-1}, x_n]}\right) =$$

$$= \varepsilon_n\left(\frac{f[x_{n-1}, x_n] - f[x^*, x_n]}{f[x_{n-1}, x_n]}\frac{\varepsilon_{n-1}}{x_{n-1} - x^*}\right) = \varepsilon_n\varepsilon_{n-1}\frac{f[x^*, x_{n-1}, x_n]}{f[x_{n-1}, x_n]}$$

és innen az osztott differenciák és a deriváltak között érvényes összefüggés (14.1.1 Következmény) segítségével kapjuk az eredményt. ■

16.4.2 Következmény

A Newton-módszerre vonatkozó eredményt az $x_{n-1} \to x_n$ határátmenettel kapjuk:

$$\varepsilon_{n+1} = \varepsilon_n^2 \frac{f''(\xi)}{2f'(x_n)}, \quad \xi \in [x_n, x^*], \tag{16.10}$$

Látjuk, ha van konvergencia, akkor az másodrendű, feltéve, hogy $f'(x^*) \neq 0$.

16.4.3 Tétel, monoton konvergencia

Legyen $f \in C^2[a,b],\ f(x^*)=0,\ x^* \in [a,b]$, *az* $f'(x), f''(x)$ *deriváltak ne váltsanak előjelet* $[a,b]$*-ben, továbbá az* $x_0 \in [a,b]$ *kezdőpontra teljesüljön* $f(x_0)f''(x_0)>0$. *Ekkor a Newton-módszer konvergens és az általa készített* x_n *sorozat monoton módon tart az* x^* *zérushelyhez.*

Bizonyítás. A Newton-módszer (16.10) formulája szerint az összes iterált a gyöktől vagy jobbra, vagy balra helyezkedik el, mert f''/f' előjele állandó. A (16.7) formulából x^*-ot levonva

$$\varepsilon_{n+1} = \varepsilon_n - f(x_n)/f'(x_n) \tag{16.11}$$

következik. Az $f(x_0)f''(x_0)>0$ feltétel miatt $f''(x_0)/f'(x_0)$ és $f(x_0)/f'(x_0)$ előjele megegyezik. Emiatt, ha (16.10)-ben $\varepsilon_1>0$, akkor $f''(x_0)/f'(x_0)$ pozitív és (16.11)-ben ε_0 kissebítve van és az összes további lépésben $\varepsilon_n>0$ kisebbedik. Hasonlóan kapjuk, hogy $\varepsilon_1<0$ esetén az összes további $\varepsilon_n<0$ nagyobbodik, tehát az ε_n-ek vagy felülről vagy alulról monoton módon tartanak 0-hoz. ■

Következmény. A (16.9) formula mutatja, hogyha a szelőmódszert úgy indítjuk, hogy $x_0, x_1 \in [a,b]$, ε_0 ε_1 és $f''(x_0)/f'(x_0)$ előjele megegyezik, akkor a tétel feltételei mellett a szelőmódszer is monoton konvergens sorozatot állít elő, mert a formulájában szereplő osztott differencia mindig helyettesíthető egy intervallum-beli deriválttal, aminek az előjele $[a,b]$-ben állandó.

16.4.4 Tétel, lokális konvergencia

Legyen $f \in C^2[a,b],\ f(x^*)=0,\ f'(x)\neq 0,\ x, x^* \in [a,b]$, *és az* $x_0 \in [a,b]$ *kezdőpontra teljesüljön*

$$\left|x_0 - x^*\right| < \frac{2\min_{[a,b]}\left|f'(x)\right|}{\max_{[a,b]}\left|f''(x)\right|} = \frac{1}{M}. \tag{16.12}$$

Ilyen x_0*-ból indítva a Newton-Raphson módszer konvergál* x^**-hoz. A szelőmódszer konvergál, ha* x_0 *mellett* x_1 *is kielégíti a* (16.12) *feltételt.*

Bizonyítás. Az első lépéstől kezdve van kontrakció, ha (16.9) vagy (16.10) alapján

$$\left|\varepsilon_0 \frac{f''(\xi)}{2f'(\eta)}\right| \le \left|x_0 - x^*\right| \frac{\max_{[a,b]}\left|f''(x)\right|}{2\min_{[a,b]}\left|f'(x)\right|} < 1.$$

Az állítás innen átrendezéssel adódik. A szelőmódszernél a második lépéshez még ε_1-re is meg kell követelnünk ugyanezt a feltételt. ■

10. A fentiek alapján a Newton-Raphson módszernél megbecsüljük az $n+1$-edik hibát. Bevezetve a $d_k = M\left|\varepsilon_k\right|$ jelölést

$$d_{n+1} = M\left|\varepsilon_{n+1}\right| \le M^2\varepsilon_n^2 \quad \rightarrow \quad d_{n+1} \le d_0^{2^n} \quad \rightarrow \quad \left|\varepsilon_{n+1}\right| \le \left(M\varepsilon_0\right)^{2^n}. \tag{16.13}$$

16.4.5 Tétel, szelőmódszer konvergencia-sebessége

A 16.4.4 Tétel feltételei mellett az x_0, x_1 *kezdőpontokból indítva a szelőmódszer* $p=(1+\sqrt{5})/2 \approx 1{,}62$ *aszimptotikus sebességgel konvergál* x^**-hoz.*

Bizonyítás. Most

$$\left|\varepsilon_{n+1}\right| \le M\left|\varepsilon_n\right|\left|\varepsilon_{n-1}\right|$$

érvényes (16.9) alapján, ahol M ugyanaz, mint (16.12)-ben. Ismét a $d_k = M\left|\varepsilon_k\right|$ jelöléssel

$$d_{n+1} \le d_n d_{n-1}, \quad n = 1,2,\ldots$$

Az indításkor $\left|x_0 - x^*\right| < 1/M$ és $\left|x_1 - x^*\right| < 1/M$, ezzel $d_0, d_1 < 1$. Igaz tehát, hogy $\exists d < 1$: $d_0, d_1 \le d$, amellyel $d_2 \le d^2$, $d_3 \le d^3$, $d_4 \le d^5$, általában

$$d_n \le d^{f_n}, \quad f_0 = f_1 = 1, \quad f_{n+1} = f_{n-1} + f_n, \quad n = 1,2,\ldots$$

Itt f_n-ek a jólismert Fibonacci-sorozat tagjai, melyeknek explicit előállítása ismert:

$$f_n = \frac{1}{\sqrt{5}}\left[b_1^{n+1} - b_2^{n+1}\right], \quad b_1 = \frac{1+\sqrt{5}}{2}, \quad b_2 = \frac{1-\sqrt{5}}{2}. \tag{16.14}$$

Mivel $\left|b_2\right| < 1$, a növekvő hatványai zérushoz fognak tartani. Emiatt létezik egy K szám, hogy minden n-re $d^{s_{n+1}} \le K$, $s_{n+1} = -b_2^{n+1}/\sqrt{5}$. Tehát írható

$$d_n \le K\left(d^{b_1/\sqrt{5}}\right)^{b_1^n} = K\left(\tilde{d}\right)^{b_1^n}, \quad \tilde{d} = d^{b_1/\sqrt{5}}.$$

Kaptuk, hogy a szelőmódszerhez tartozó hibák majorálhatók egy olyan sorozattal, amelynek konvergenciarendje $b_1 = \frac{1+\sqrt{5}}{2} \approx 1,62$, azaz a módszer *szuperlineáris*. ■

16.5. Példák

1. Legyen $f \in C^3[a,b]$. Parabola interpolációval készítsünk három pontra támaszkodó iterációs módszert $f(x)$ egy $[a,b]$-beli lokális minimumának meghatározására!

Megoldás. Legyen $[a,b]$-ben három pont $(x_{i-2}, f_{i-2}), (x_{i-1}, f_{i-1}), (x_i, f_i)$. Newton-interpolációval $p_2(x) = f_{i-2} + f[x_{i-2}, x_{i-1}](x - x_{i-2}) + f[x_{i-2}, x_{i-1}, x_i](x - x_{i-2})(x - x_{i-1})$. A deriváltjának zérushelye:

$$x_{i+1} = \frac{x_{i-2} + x_{i-1}}{2} - \frac{f[x_{i-2}, x_{i-1}]}{2f[x_{i-2}, x_{i-1}, x_i]}. \tag{16.15}$$

2. Egyenletes lépésközzel haladva hogyan derítenénk fel egy minimumhelyet?

Megoldás. Legyen a lépésköz h és $x_j = a + jh$, $j = 0,1,\ldots$. Az x_{j-1}, x_j, x_{j+1} alappont-hármas megfelelő, ha $f[x_{j-1}, x_j] < 0$ és $f[x_j, x_{j+1}] > 0$. Ekkor a lokális minimumot a következő egyszerűsített formulával becsülhetjük, ha (16.15)-ben x_j-t vesszük a középső pontnak:

$$x_{\min} \approx \frac{x_{j-1} + x_{j+1}}{2} - \frac{f[x_{j-1}, x_{j+1}]}{2f[x_{j-1}, x_j, x_{j+1}]} = x_j - \frac{h}{2}\frac{f_{j+1} - f_{j-1}}{f_{j+1} - 2f_j + f_{j-1}}. \tag{16.16}$$

3. A kapott iterációs módszerre fogalmazzunk meg ahhoz hasonló tételt, mint amit a Newton-módszernél láttunk a lokális konvergenciára!

Megoldás. Az egyszerűség kedvéért tekintsük (16.15)-ben azt az esetet, amikor $i = 2$. Az osztott differenciák tulajdonságait kihasználva a hibák terjedésére próbálunk egy összefüggést származtatni. Vonjuk le mindkét oldalból a minimumhelyet adó x^*-ot és legyen $\varepsilon_i = x_i - x^*$, ezzel

$$\varepsilon_3 = \frac{\varepsilon_0 + \varepsilon_1}{2} - \frac{f[x_0, x_1]}{2f[x_0, x_1, x_2]}.$$

A számlálóban lévő osztott differenciát átalakítjuk, kihasználva, hogy $f[x^*,x^*]=0$ és az alappontok sorrendje az osztott differenciákban tetszőleges: $f[x_0,x_1]=f[x_0,x_1]-f[x_1,x^*]+f[x_1,x^*]-f[x^*,x^*]=$ $=\varepsilon_0 f[x_0,x_1,x^*]+\varepsilon_1 f[x_1,x^*,x^*]$. Beírva a fenti formulába és közös nevezőre hozva:

$$\varepsilon_3=\frac{\varepsilon_0\left(f[x_0,x_1,x_2]-f[x_0,x_1,x^*]\right)+\varepsilon_1\left(f[x_0,x_1,x_2]-f[x_1,x^*,x^*]\right)}{2f[x_0,x_1,x_2]}.$$

A számláló első két tagja továbbírva $\varepsilon_0\varepsilon_2 f[x_0,x_1,x_2,x^*]$. A utolsó két tag átalakítása kicsit hosszabb: $\varepsilon_1\left(f[x_0,x_1,x_2]-f[x_0,x_1,x^*]+f[x_0,x_1,x^*]-f[x_1,x^*,x^*]\right)=\varepsilon_1\varepsilon_2 f[x_0,x_1,x_2,x^*]+\varepsilon_0\varepsilon_1 f[x_0,x_1,x^*,x^*]$. Ezekkel

$$\varepsilon_3=\frac{(\varepsilon_0+\varepsilon_1)\varepsilon_2 f[x_0,x_1,x_2,x^*]}{2f[x_0,x_1,x_2]}+\frac{\varepsilon_0\varepsilon_1 f[x_0,x_1,x^*,x^*]}{2f[x_0,x_1,x_2]}.$$

Legyen $\delta_2=\max\{|\varepsilon_0|,|\varepsilon_1|,|\varepsilon_2|\}$ és

$$M=\frac{\max\limits_{x\in[a,b]}|f^{(3)}(x)|}{2\min\limits_{x\in[a,b]}|f''(x)|}. \tag{16.17}$$

Felhasználva, hogy az osztott differenciák kifejezhetők a nekik megfelelő rendű deriváltakkal, kapjuk:

$$|\varepsilon_3|\le\frac{3}{2}\delta_2^2\frac{2!}{3!}2M=\delta_2^2 M\,. \tag{16.18}$$

Így $|\varepsilon_3|$ biztosan kisebb a három megelőző ε abszolút maximumánál, ha $\delta_2 M<1$, vagy másképp $\delta_2<1/M$. Tehát a kapott módszer biztosan konvergens, ha a három induló pont a minimumhely $1/M$ -sugarú környezetében van.

16.6. Gyakorlatok

1. Bizonyítsuk be, hogyha $f\in C^1[a,b]$, $f(a)f(b)<0$ és $f'(x)$ nem vált előjelet $[a,b]$-ben, akkor ott az $f(x)$ függvénynek csak egy gyöke van.
2. A fixponttétel alkalmazásával mutassuk meg, hogy a $\cos x-4x+2=0,\ x\in\mathbb{R}$ egyenletnek egy zérushelye van és x-et $4x$ felől kifejezve a fixpont iteráció minden kezdőértékre konvergens!
3. Az előző feladatban a gyök milyen környezetéből konvergál biztosan a Newton-iteráció?
4. Oldjuk meg az $f(x)=1/x-a=0$ egyenletet Newton-iterációval! Milyen kezdőértékekre van konvergencia? A kapott formula érdekesége, hogy nincs benne osztás, aminek régebben külön jelentősége volt az osztás műveletével nem rendelkező gépi aritmetikákban.
5. Oldjuk meg az $f(x)=x^2-a=0,\ a>0$ egyenletet Newton-iterációval és tisztázzuk a konvergenciát!
6. Az előző feladat megoldása alapján készítsünk módszert $a^{1/k}$ meghatározására, ahol a pozitív valós szám.
7. Mutasssuk meg, hogy a 16.4.3 Tételt módosíthatjuk úgy, hogy az $f(x_0)f''(x_0)>0$ feltételt elhagyjuk és helyette azt követeljük meg, hogy az első lépés után $x_1\in[a,b]$.
8. Mutassuk meg, hogy az $F(x_n)=x_n-\dfrac{\left(f(x_n)\right)^2}{f(x_n)-f(x_n-f(x_n))}$ iteráció konvergenciája másodrendű!

9. Ellenőrízzük, hogy a Newton-módszer többszörös gyök esetén csak elsőrendben konvergál.
10. Igazoljuk, hogy r-szeres multiplicitású gyöknél a kvadratikus konvergencia megmarad, ha a Newton-módszer formuláját a következőre módosítjuk: $x_{n+1} = x_n - rf(x_n)/f'(x_n)$.
11. Adott ε pontosság elérése érdekében dolgozzuk ki annak feltételét, hogy mikor állítsuk le a Newton-módszert.
12. Mi történik a szelőmódszernél, ha a 16.4.3 Tétel feltételeitől csak annyi a különbség, hogy $\varepsilon_0 > 0$, de $\varepsilon_1 < 0$?
13. Mikor állítsuk le a szelőmódszert, hogy a megoldás előírt pontosságú legyen?

17. Nemlineáris egyenletek megoldása II.

Néhány speciális esettel folytatjuk.

17.1. Az intervallumfelezés módszere

Tegyük fel, az $[a,b]$ intervallum tartalmaz 1 db gyököt: $f(a)f(b)<0$ és a függvény folytonos $[a,b]$-ben. Az intervallumfelezés módszere szerint ekkor megfelezzük az intervallumot és a két intervallum közül megtartjuk azt, ahol az előjelváltás megmarad. Így az algoritmus:

1. $\exists f \in C[a,b],\ f(a)f(b)<0$ és adott $\in$ előírt pontosság.
2. Indulás: $[a_0,b_0]=[a,b],\ x_1=(a+b)/2$.
3. $$[a_n,b_n]=\begin{cases}[a_{n-1},x_n], & \text{ha } f(a_{n-1})f(x_n)<0,\\ [x_n,b_{n-1}], & \text{egyébként,}\end{cases}$$
   $x_{n+1}=(a_n+b_n)/2.$
4. Megállás: ha $f(x_n)=0$, vagy $|b_n-a_n|<\in$.

Ez nem túl gyors, de biztos módszer. Az előjelváltásból nem mindig következik a gyök léte. Gondoljunk az $1/x$ függvényre, amikor az algoritmust -1 és 2 között indítjuk.

17.1.1 Tétel

Az intervallumfelezéssel kapott $x_n,\ n=1,2,\ldots$ *sorozat elsőrendben konvergens és*

$$|\varepsilon_n| \le \frac{b-a}{2^n},\ n=0,1,\ldots \tag{17.1}$$

Bizonyítás. A konvergencia abból következik, hogy mindig a gyököt tartalmazó intervallumot tartjuk meg. A hibára minden lépésben teljesül:

$$|\varepsilon_{n+1}| \le \frac{1}{2}|\varepsilon_n|,$$

ez pedig elsőrendű konvergenciát jelent. ■

17.2. A húrmódszer (regula falsi)

Itt csak annyi az eltérés az intervallumfelezés módszerétől, hogy nem az intervallum közepét vesszük, hanem az $(a_n, f(a_n))$ és $(b_n, f(b_n))$ pontokra illesztett egyenes, más névvel: *húr* zérushelye a következő közlítés:

$$x_{n+1} = a_n - f(a_n)\frac{b_n-a_n}{f(b_n)-f(a_n)}. \tag{17.2}$$

Megfelelő feltételek mellett bizonyítható, hogy a húrmódszer konvergencája lineáris, így nem gyorsabb, mint az intervallumfelezés. Még az is megeshet, hogy annál lassúbb. Ez történik például olyan esetben, amikor a függvény értékei az x-tengelyhez közel vannak és az egyik végponthoz (a_n vagy b_n) nagyon közeli a gyök.

17.3. A Newton-iteráció többváltozós esetben

Legyen $f: \mathbb{R}^n \to \mathbb{R}^n$ egy n-változós leképezés, amelynek keressük azt a vektorát, amelyre $f(x)=0$. Tételezzük fel a differenciálhatóságot, ekkor az $x_k \in \mathbb{R}^n$ körüli sorfejtésből közelítve

$$f(x_k)+f'(x_k)(x-x_k)=0, \tag{17.3}$$

ahol most $f'(x)=\left[\partial f_i(x)/\partial x_j\right]\in\mathbb{R}^{n\times n}$ mátrix – a rendszer un. Jacobi-mátrixa - , amelyről feltesszük, hogy invertálható. (17.3)-et x-re megoldva a következő iterációt kapjuk:

$$x_{k+1}=x_k-\left[f'(x_k)\right]^{-1}f(x_k). \tag{17.4}$$

Ha van megoldás és elég közel vagyunk hozzá, akkor remélhetjük, hogy a többváltozós Newton-iteráció konvergens lesz.

A módszer azt követeli meg, hogy minden lépésben elkészítsük a deriváltak mátrixát és megoldjunk vele egy lineáris egyenletrendszert. Mivel ez nagyon munkaigényes lehet, szokás alkalmazni a következő egyszerűsítést: Elkészítjük az $f'(x_k)=LU$ faktorizációt és utána az egyszerűbb

$$x_{k+1}=x_k-\left(LU\right)^{-1}f(x_k) \tag{17.5}$$

iterációt alkalmazzuk. Ez 1-dimenzióban annak felel meg, hogy lépésenként a derivált értékét nem változtatjuk. Az ilyen módszereket kvázi-Newton módszereknek nevezzük.

17.4. Polinomok gyökei

A polinomok gyökeinek meghatározása talán leggyakrabban a mátrixok sajátértékeinek keresésekor jön elő, de ekkor nem érdemes a hatványösszeg alakot használni, mert a lineáris algabrai algoritmusok numerikusan előnyösebb megoldásokat kínálnak. Valóban magasabbfokú polinomok esetén a hatványösszeg reprezentáció

$$p(x)=\sum_{i=0}^{n}a_i x^i \tag{17.6}$$

nem előnyös, mert gépi számként ábrázolva az együtthatók n növekedésével egyre bizonytalanabb információt nyújtanak a gyökök pontos értékéről. Példának álljon itt Wilkinson kisérlete, aki az 1,2,...,19,20 gyökökkel rendelkező huszadfokú polinomot (17.6) alakban előállította, majd visszaszámolta a gyököket. Az eredmény annyira más volt, hogy több komplex gyökpárt is kapott. A jelenséget magyarázza a gyökök és együtthatók összefüggése: például a nulladfokú tag a gyökök szorzata: 20!, aminek a pontos ábrázolására messzi nem elegendő 15 decimális jegy. Így a gépi számábrázolás folytán sok fontos információ elvész.

A (17.6) alak összefüggésbe hozható az un. *Frobenius-féle kisérő mátrix*-szal, amellyel már találkoztunk a 7.3 szakaszban:

$$F=\begin{bmatrix} 0 & 0 & \dots & \dots & 0 & -a_0/a_n \\ 1 & 0 & \dots & \dots & 0 & -a_1/a_n \\ & 1 & \ddots & & \vdots & \vdots \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & \ddots & 0 & -a_{n-2}/a_n \\ & & & & 1 & -a_{n-1}/a_n \end{bmatrix}. \tag{17.7}$$

Az utolsó oszlopa mentén kifejtve könnyen igazolható, hogy $\det\left(\lambda I-F\right)=\frac{1}{a_n}\sum_{i=0}^{n}a_i\lambda^i$. Ennek a mátrixnak ismerete két szempontból is hasznos. Egyrészt mutatja, hogy a polinom x_k gyökei lineáris

algebrai módszerekkel kereshetők, amelyek a legstabilabbnak tekinthető módszerek közé tartoznak. Másrészt rögtön lehetőségünk van egy olyan körlemez megadására a komplex síkon, amelyben a polinom összes gyöke benne van:

$$\|F\|_\infty = \max_{0\le i<n}\left(1-\delta_{i0}+|a_i/a_n|\right) = R \ge |x_k|,$$

ahol δ_{ij} a Kronecker delta. R nagyobb vagy egyenlő F spektrál sugaránál, ami most a polinom gyökök abszolút értékeinek maximuma.

Megadhatunk egy másik kisebb körlemezt is, amelyen kívül van az összes gyök. Vezessük be az $x=1/y$ transzformációt és írjuk át a polinomot y szerint. Eredményül egy olyan polinomot kapunk, ahol az együtthatók fordított sorrendben vannak és ennek a polinomnak a gyökei az eredeti gyökök reciprokai. Az új polinomhoz tartozó Frobenius-féle mátrix sornormáját véve kapjuk: $1/|x_k| \le \max_{0<i\le n}\left(1-\delta_{in}+|a_i/a_0|\right) = 1/r$, ahol természetesen feltételeztük, hogy $a_0 \ne 0$. A két eredményt egybevetve látjuk, hogy a polinom gyökei az

$$r \le |x_k| \le R, \quad k=1,2,\ldots,n \tag{17.8}$$

körgyűrű tartományba esenek.

A (17.6)-tal adott polinomoknál előnyösen alkalmazható a Newton-módszer, mert a polinom értéke és a deriváltja egy lépésben egyszerűen számolható. Ha például a ξ helyen szeretnénk ezeket kiszámolni, nem kell mást tennünk, mint a polinomot maradékos osztással elosztani $(x-\xi)^2$-tel:

$$p(x) = q(x)(x-\xi)^2 + \alpha x + \beta. \tag{17.9}$$

Könnyen meggyőződhetünk róla, hogy a helyettesítési érték $\alpha\xi+\beta$, a derivált pedig α lesz a ξ helyen.

A többszörös gyökök kiszűrésére alkalmazhatjuk az Euklidészi algoritmust. Ekkor a két induló polinom $p_0(x)=p(x), \quad p_1(x)=-p'(x)$, az $i+1$-edik polinomot pedig úgy készítjük, hogy $p_{i-1}(x)$-et osztjuk $p_i(x)$-szel és a maradékot képezzük:

$$p_{i-1}(x) = q_i(x)p_i(x) - c_i p_{i+1}(x), \quad i=1,2,\ldots \tag{17.10}$$

A sorozatban a polinomok fokszáma csökkenő, $c_i > 0$, egyébként tetszőleges. Az algoritmus $m \le n$ lépés után befejeződik:

$$p_{m-1}(x) = q_m(x)p_m(x), \quad p_m(x) \ne 0.$$

Az utolsó polinom a két kezdő polinom legnagyobb közös osztója. Mivel a derivált polinom az 1-nél nagyobb multiplicitású gyököket tartalmazza, így ezek a gyökök megjelennek $p_m(x)$-ben.

Abban az esetben, amikor minden gyök valós és egyszeres, akkor az Eulidészi algoritmus olyan polinomsorozatot készít, amely *Sturm-sorozat tulajdonságú*. Legyen az a helyen a sorozat előjelváltásainak száma $V(a)$, a b helyen pedig $V(b)$, ekkor megmutatható, hogy az $[a,b]$ intervallumban a gyökök száma $V(a)-V(b)$.

17.5. Gyakorlatok

1. Készítsük el a (17.9) osztás algoritmusát!
2. Adjuk meg a $4x^5-3x^4+6x^3-5x^2-8x+2$ polinom gyökeit tartalmazó körgyűrűt!

18. Numerikus integrálás (kvadratúra) I.

Az integrálok kiszámításakor nem mindig ismert a primitív függvény, vagy ha igen, némely esetben nagyon bonyolult, nehezen számítható. Ilyenkor a numerikus módszerek a kívánt pontosságú eredmény előállítására egyszerűbb alternatívát kínálnak. A továbbiakban az interpolációból nyerhető kvadratúra-formulákkal fogunk foglalkozni.

Láttuk, a függvény az $[a,b]$ intervallumban a következő módon állítható elő:

$$f = L_n + r_n, \tag{18.1}$$

ahol L_n a Lagrange-interpolációs polinom és r_n a hibatag. (Feltesszük, az alappontok nagyság szerint rendezettek: $x_{i-1} < x_i$ és $x_0 = a$, $x_n = b$.) A kvadratúra-formulák származ-tatási elve:

$$\int_a^b f = \int_a^b L_n + \int_a^b r_n = \sum_{i=0}^{n} a_i f(x_i) + R_n, \tag{18.2}$$

ahol az

$$a_i = \int_a^b l_i(x)dx \tag{18.3}$$

súlyok a Lagrange alappolinomok integrálásából adódnak.

Következmény. Az így nyert formulák legfeljebb n-edrendű polinomig pontosak.

Ekvidisztáns alappontok esetén nyerjük a Newton-Cotes formulákat.

18.1. Zárt és nyilt Newton-Cotes kvadratúra formulák

18.1.1 Definició

Az alappontok halmaza legyen $\Omega_n = \{x_0,\dots,x_n\}$. *Zárt* a kvadratúra-formula, ha $a,b \in \Omega_n$, $h = (b-a)/n$, $x_k = a + k\cdot h$, $k = 0,1,\dots,n$. *Nyilt* a formula, ha $a,b \notin \Omega_n$, $h = (b-a)/(n+2)$, $x_k = a + (k+1)\cdot h$, $k = 0,1,\dots,n$, $x_{-1} = a$, $x_{n+1} = b$.

A továbbiakban rátérünk a zárt Newton-Cotes formulák együtthatóinak előállítására.

$$a_k = \int_a^b l_k(x)dx = \int_a^b \frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)}dx.$$

Vegyük észre: $x_k - x_j = (k-j)h$ és vezessünk be új változót: $t = (x-a)/h$, ahonnan $x = a + th$ és $x - x_j = (t-j)h$, s ezzel

$$\begin{aligned} a_k &= \int_0^n \frac{t(t-1)\overbrace{\dots\dots}^{(t-k)\text{ hiányzik}}(t-n)h^n}{k(k-1)\dots1(-1)(-2)\dots(-n+k)h^n}hdt = \\ &= (b-a)\frac{(-1)^{n-k}}{nk!(n-k)!}\int_0^n \frac{t(t-1)\dots(t-n)}{t-k}dt = (b-a)B_{k,n}^{\text{zárt}}, \end{aligned} \tag{18.4}$$

ahol a $B_{k,n}^{\text{zárt}}$ együtthatók az intervallumtól függetlenül egyszer s mindenkorra kiszámíthatók. Hasonló módon nyerhetjük a *nyilt Newton-Cotes* formulák együtthatóit:

$$a_k = (b-a)\frac{(-1)^{n-k}}{(n+2)k!(n-k)!}\int_0^{n+2}\frac{(t-1)(t-2)\ldots(t-n-1)}{t-k-1}dt = (b-a)B_{k,n}^{\text{ny}}, \qquad (18.5)$$

Az első néhány Newton-Cotes együttható:

Zárt					Nyílt				
1	1			Trapéz	1			Érintő formula	
1	4	1		Simpson	1	1			
1	3	3	1		2	-1	2		
7	32	12	32	7	11	1	1	11	

A táblázatban minden sort osztani kell az együtthatók összegével, mert az együtthatók összegének 1-nek kell lenni. Például az 1 4 1 súlyok az 1/6 4/6 1/6 valódi súlyokra utalnak.

18.1.2 Tétel

$$1.\ \sum_{k=0}^{n} B_{k,n} = 1, \qquad 2.\ B_{k,n} = B_{n-k,n}. \qquad (18.6)$$

Bizonyítás. Az első állítás az $f(x)\equiv 1$ függvény integrálásából adódik, kihasználva, hogy az integrál 0-adfokú polinomra pontos. A második állítást az $y = n - t$ új változóra való áttéréssel nyerjük. ■

18.2. Néhány egyszerű integráló formula

1. Az *érintőformula* (nyílt Newton-Cotes): $n = 0$, $B_{0,0}^{\text{ny}} = \frac{1}{2\cdot 1\cdot 1}\int_0^2 1\cdot dt$, tehát

$$I_0(f) = (b-a)f\left(\frac{a+b}{2}\right). \qquad (18.7)$$

Az érintőformula úgy is értelmezhető, hogy a függvényt $[a,b]$-ben a középponthoz húzott érintő egyenessel közelítjük, és az egyenes alatti területet vesszük. Ez mutatja, hogy legfeljebb elsőfokú polinomig pontos.

18.2.1 Tétel, érintő formula hibája

Legyen $c = (a+b)/2$, $f \in C^2[a,b]$, *ekkor az érintő formulával*

$$\int_a^b f(x)dx = (b-a)f(c) + \frac{(b-a)^3}{24}f''(\eta), \quad \eta \in [a,b]. \qquad (18.8)$$

Bizonyítás. A c körüli sorfejtésből

$$f(x) = f(c) + (x-c)f'(c) + \frac{(x-c)^2}{2}f''(\xi_x).$$

Integráláskor az első tag adja a közelítő formulát, a második tag eredménye zérus, így elegendő a hibatagot vizsgálni:

$$R_1(f) = \frac{1}{2}\int_a^b (x-c)^2 f''(\xi_x)dx.$$

Az integrálszámítás középértéktétele szerint

$$R_1(f)=\frac{f''(\eta)}{2}\int_a^b (x-c)^2\,dx=\frac{f''(\eta)}{2}\left[\frac{(x-c)^3}{3}\right]_a^b=\frac{(b-a)^3}{24}f''(\eta).$$ ■

A gyakorlatban ezt a formulát nem az egész $[a,b]$ intervallumra alkalmazzuk, hanem azt m részre osztjuk, és az egyes részintervallumokban az érintőformulával integrálunk. Például, ha $m=3$: $h=(b-a)/3$ és a három részintervallumra alkalmazzuk a (18.7) szabályt.

A részintervallumon nyert eredmények felösszegzésével jutunk az *érintőszabályhoz:*

$$\int_a^b f=\frac{b-a}{m}\sum_{i=1}^{m} f(a-h/2+ih)+\frac{(b-a)^3}{24m^2}f''(\eta), \tag{18.9}$$

ahol most $f''(\eta)=\frac{1}{m}(f''(\eta_1)+f''(\eta_2)+\ldots+f''(\eta_m))$, mert a Darboux-tulajdonság miatt f'' ezt az átlagértéket is felveszi valahol a teljes intervallumban.

2. A *trapézformula*. Elsőfokú polinom interpolációból nyert zárt formula: $n=1,\ B_0^z=B_1^z=1/2$:

$$I_1(f)=\frac{b-a}{2}(f(a)+f(b)). \tag{18.10}$$

18.2.2 Tétel, trapézformula hibája

Legyen $f\in C^2[a,b]$, *ekkor*

$$\int_a^b f=\frac{b-a}{2}(f(a)+f(b))-\frac{(b-a)^3}{12}f''(\eta),\quad \eta\in[a,b]. \tag{18.11}$$

Bizonyítás. Az interpoláció hibatagjának integrálja az integrálszámítás középérték tételének felhasználásával:

$$R_1(f)=\int_a^b \frac{f''(\xi_x)}{2}(x-a)(x-b)dx=-\int_a^b \frac{f''(\xi_x)}{2}\underbrace{(x-a)(b-x)}_{\geq 0}dx=-\frac{f''(\eta)}{12}(b-a)^3.$$ ■

A teljes intervallumot m részre osztva, a részintervallumok eredményét felösszegezve nyerjük a *trapéz szabályt*:

$$\int_a^b f=\frac{b-a}{2m}[f(x_0)+2f(x_1)+\ldots+2f(x_{m-1})+f(x_m)]-\frac{(b-a)^3}{12m^2}f''(\eta). \tag{18.12}$$

18.2.3 Definició

Egy kvadratúra formulát akkor mondunk k*-adrendűnek,* ha k-adfokú az a legkisebb fokszámú polinom, amelyre a formula már nem pontos.

Eszerint az érintő- és trapézformula *másodrendű.*

3. A *Simpson formula*: másodfokú polinom interpolációból nyert zárt formula, $n=2$, $B_0^z=1/6$, $B_1^z=4/6$, $B_2^z=1/6$ és

$$I_2(f)=\frac{b-a}{6}(f(a)+4f(\frac{a+b}{2})+f(b)).$$

Az interpoláció maradéktagjában $\omega_2(x)=(x-a)(x-\frac{a+b}{2})(x-b)$ szerepel, ennek az integrálja $[a,b]$-ben zérus. Ezt legegyszerűbben úgy tudjuk belátni, hogy $[a,b]$-t a $[-1,1]$ intervallumba transzformáljuk. Ekkor $\omega_2(x)$ páratlan függvény, amelynek az integrálja zérus. Emiatt a hibatételt az Hermite-interpolációból származtatjuk,

$$f(x)=H_3(x)+\frac{f^{(4)}(\xi_x)}{4!}(x-a)(x-\frac{a+b}{2})^2(x-b)\,, \tag{18.13}$$

ahol az $(a+b)/2$ középpontban az első deriváltat is interpoláljuk. Az általánosított osztott differenciák táblázatára gondolva tudjuk, hogy az interpoláló polinom a következő alakú: $H_3(x)=L_2(x)+C(x-a)(x-\frac{a+b}{2})(x-b)$. A második tag együtthatójának értéke nem fontos, mert az integrálja az előbbiek alapján zérus, s ezzel $\int_a^b H_3=\int_a^b L_2$.

18.2.4 Tétel, Simpson-formula hibája

Legyen $f\in C^4[a,b]$. *Ekkor létezik* $\eta\in[a,b]$, *amelyre*

$$\begin{aligned}\int_a^b f &= \frac{b-a}{6}\left(f(a)+4f(\frac{a+b}{2})+f(b)\right)-\frac{f^{(4)}(\eta)}{90}\left(\frac{b-a}{2}\right)^5= \\ &= I_2(f)-\frac{f^{(4)}(\eta)}{90}h^5,\end{aligned} \tag{18.14}$$

ahol $h=(b-a)/2.$

Bizonyítás. Kiindulunk az Hermite-interpoláció (18.13) alakjából, ahonnan integrálással kapjuk:

$$I(f)-I_2(f)=\int_a^b\frac{f^{(4)}(\xi_x)}{4!}(x-a)(x-\frac{a+b}{2})^2(x-b)dx.$$

Ahhoz, hogy az integrálszámítás középértéktételét alkalmazhassuk, az $f^{(4)}$ mellett álló tényező nem lehet negatív. Ezt úgy biztosíthatjuk, hogy $(x-b)$ helyett $(b-x)$-et írunk, s így

$$I(f)-I_2(f)=-\frac{f^{(4)}(\eta)}{4!}\int_a^b(x-a)(x-\frac{a+b}{2})^2(b-x)dx=-\frac{f^{(4)}(\eta)}{90}\left(\frac{b-a}{2}\right)^5. \qquad \blacksquare$$

Simpson-szabály. A teljes $b-a$ intervallumot páros számú m részintervallumra osztva és a Simpson-formulát a szomszédos intervallumpárokra alkalmazva kapjuk a Simpson-szabályt, mint összetett formulát. Ekkor három pontonként fogjuk össze a formulákat és az összetett formula:

$$\begin{aligned}\int_a^b f &= \sum_{k=1,3,\ldots}\left(\frac{2h}{6}(f(x_{k-1})+4f(x_k)+f(x_{k+1}))-\frac{f^{(4)}(\eta_k)}{90}h^5\right)= \\ &= \frac{h}{3}\left(f(x_0)+2\sum_{\substack{k\ \text{páros}\\ \text{belső pont}}}f(x_k)+4\sum_{k\ \text{ptlan}}f(x_k)+f(x_m)\right)-\frac{h^5}{90}\sum_{k\ \text{ptlan}}f^{(4)}(\eta_k).\end{aligned} \tag{18.15}$$

A hibatag még tovább írható:

$$-\frac{h^5}{90}\sum_{k\ \text{ptlan}} f^{(4)}(\eta_k) = -\frac{(b-a)^5}{180m^4}\left(\frac{\sum f^{(4)}(\eta_k)}{m/2}\right) = -\frac{(b-a)^5}{180m^4} f^{(4)}(\eta), \tag{18.16}$$

mivel a Darboux-tulajdonság miatt van egy η, amelyre a negyedik derivált az átlagértéket felveszi.

18.3. Példák

1) Az $\int_{-1}^{1} \frac{dx}{2+x}$ integrált az érintőszabállyal közelítjük. Hány osztópontot kell választanunk, hogy az integrált 10^{-2}-nél kisebb hibával kapjuk?

Megoldás. Azt kell biztosítani, hogy $\frac{(b-a)^3}{24m^2} M_2 \le 10^{-2}$ teljesüljön, ahol $b-a=2$ és $M_2 = 2 \max_{x\in[-1,1]} \left|(2+x)^{-3}\right| = 2$. A számokat helyettesítve: $200/3 \le m^2$, így $m=9$ megfelel.

2) Határozzuk meg az A_0, A_1, A_2 paramétereket úgy, hogy a $\int_0^2 \sqrt{x} f(x)dx \approx$ $\approx I_2(f) = A_0 f(0) + A_1 f(1) + A_2 f(2)$ kvadratúra legfeljebb másodfokú polinomokra pontos legyen!

Megoldás. Két megoldás is létezik. Az egyik, hogy kiszámítjuk a kijelölt integrálokat a Lagrange-alappolinomokkal: $A_i = \int_0^2 l_i(x)\sqrt{x}dx$, ahol $\sqrt{x}$-et súlyfüggvénynek tekintjük. A másik módszer szerint felírjuk azt a lineáris egyenletrendszert, ami a pontossági követeléseket tartalmazza. A következő egyenletrendszer első sora azt fejezi ki, hogy a kvadratúra az 1 polinomra pontos, a második sor szerint az x polinomra pontos, a harmadik sor szerint pedig az x^2 polinomra pontos:

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 1 & 4 \end{bmatrix}\begin{bmatrix} A_0 \\ A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} \int_0^2 x^{1/2}dx = 4\sqrt{2}/3 \\ \int_0^2 x\cdot x^{1/2}dx = 8\sqrt{2}/5 \\ \int_0^2 x^2 x^{1/2}dx = 16\sqrt{2}/7 \end{bmatrix}.$$

Ennek a megoldása: $A_0 = \frac{8\sqrt{2}}{105}$, $A_1 = \frac{32\sqrt{2}}{35}$, $A_2 = \frac{12\sqrt{2}}{35}$.

19. Numerikus integrálás, Gauss-kvadratúrák II.

Az eddigi, interpolációból származtatott kvadratúra-formulák legalább 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. Ismét szükségünk lesz az ortogonális polinomokra.

19.1. Tétel, ortogonális polinom gyökei

Legyen $\{p_k(x)\}$ *egy ortogonális polinom rendszer. Ekkor bármely* n*-re* $p_{n+1}(x)$ *gyökei valósak, egyszeresek és az* $[a,b]$ *intervallumban vannak, ahol* $[a,b]$ *a skalárszorzat integrálási tartománya.*

Bizonyítás. Legyenek $x_0, x_1, \ldots, x_k$ $p_{n+1}(x)$ páratlan multiplicitású gyökei $[a,b]$-ben, azaz ott $p_{n+1}(x)$ előjelet vált. Ha $k = n$, akkor a tétel állítása rendben van. Ha nem, akkor indirekt úton feltesszük: $k < n$ és megmutatjuk, hogy az állítás ellentmondásra vezet. Ehhez tekintsük a $q(x) = (x - x_0)(x - x_1)\ldots(x - x_k)$ $k+1$-edfokú polinomot. Mivel $k+1 < n+1$, az ortogonalitás miatt $(p_{n+1}, q) = 0$. De ezzel ellentmondásra jutunk, mert $p_{n+1}(x)q(x)$ nem vált előjelet $[a,b]$-ben, mivel p_{n+1} minden előjelválátását $q(x)$ megszünteti és így $\int p_{n+1} q\alpha \neq 0$ volna. Vegyük észre, a gondolatmenet akkor is jó, ha egyetlen páratlan multiplicitású gyök sincs, mert ekkor $q(x)$ 0-adfokú. ■

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:

$$\int_a^b f\alpha = \int_a^b L_n\alpha + \int_a^b r_n\alpha = \sum_{i=0}^{n} a_i f(x_i) + R_n, \quad a_i = \int_a^b l_i(x)\alpha(x)dx. \tag{19.1}$$

19.2. Tétel, Gauss-kvadratúra pontossága

Legyenek a $p_{n+1}(x)$ *ortogonális polinom gyökei* $x_0, x_1, \ldots, x_n$, $a_i = \int l_i\alpha$, *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:* $f \in \mathcal{P}_{2n+1} \to \int f\alpha = G_n(f)$.

Bizonyítás. Az interpolációból való származatatás miatt $G_n(f)$ biztosan pontos a legfeljebb n-edfokú polinomokra. Tegyük fel, $f \in \mathcal{P}_{2n+1}$, $f = p_{n+1} \cdot q + r$, $q, r \in \mathcal{P}_n$, így

$$\begin{aligned}
G_n(f) &= \sum_{i=0}^{n} a_i f(x_i) = \sum_{i=0}^{n} a_i [\underbrace{p_{n+1}(x_i)}_{=0 \text{ minden } i-\text{re}} \cdot q(x_i) + r(x_i)] = \\
&= \sum_{i=0}^{n} a_i r(x_i) = G_n(r) = \int r\alpha = \qquad (\text{mert } n\text{-edfokig pontos}) \\
&= \int (p_{n+1} \cdot q + r)\alpha = \qquad (\text{mert } q \in \mathcal{P}_n, \text{ így ortogonális } p_{n+1}\text{-re}) \\
&= \int f\alpha.
\end{aligned}$$

19.2.1 Következmény

Az a_i együtthatók pozitívak.

Bizonyítás. Tudjuk, $l_i(x_j) = l_i^2(x_j) = \delta_{ij}$, ahol δ_{ij} a Kronecker-delta, $l_i^2(x) \geq 0$ és $l_i^2(x) \in \mathcal{P}_{2n}$, következésképp a Gauss-kvadratúra pontos :

$$0 < \int l_i^2 \alpha = G_n(l_i^2) = \sum_{j=0}^{n} a_i l_i^2(x_j) = a_i.$$

Az $f(x) = 1$ függvény integrálásával most a következőt kapjuk az együtthatók összegére:

$$\sum_{i=0}^{n} a_i = \int \alpha = \mu_0, \quad \mu_i = \int x^i \alpha,$$

ahol μ_0 a nulladik momentum. Vegyük észre, ez egyenlő $b-a$-val, ha a súlyfüggvény 1.

19.2.2 Tétel, Gauss-kvadratúra hibaformulája

Legyen $f \in C^{2n+2}[a,b]$ és $G_n(f) = \sum_{k=0}^{n} a_k f(x_k)$, ahol az alappontok $p_{n+1}(x)$ gyökei. Akkor

$$I(f) - G_n(f) = \frac{f^{(2n+2)}(\eta)}{(2n+2)!}\left(p_{n+1}, p_{n+1}\right), \tag{19.2}$$

itt $p_{n+1}(x)$ 1-főegyütthatós ortogonális polinom.

Bizonyítás. Hermite-Fejér interpolációból (amikor az interpolációban a függvényértékek és az első deriváltak vesznek részt) kapjuk a következő hiba-előállítást:

$$f(x) = H_{2n+1}(x) + \frac{f^{(2n+2)}(\xi_x)}{(2n+2)!} \underbrace{(x-x_0)^2(x-x_1)^2 \ldots (x-x_n)^2}_{=p_{n+1}^2(x)}.$$

Innen az integrálás középérték-tételének alkalmazásával

$$I(f) - G_n(f) = \int_a^b \frac{f^{(2n+2)}(\xi_x)}{(2n+2)!} \underbrace{p_{n+1}^2(x)}_{\geq 0} \alpha(x) dx$$

nyerjük az állítást, mert $H_{2n+1}(x)$-re a Gauss-kvadratúra pontos. ■

Megadunk néhány 1-főegyütthatós ortogonális polinomot:

Név	$[a,b]$	$\alpha(x)$	μ_0	α_{n+1}	β_n	p_0	p_1	p_2
Legendre	$[-1,1]$	1	2	0	$n^2/(4n^2-1)$	1	x	$x^2 - 1/3$
Csebisev	$[-1,1]$	$\frac{1}{\sqrt{1-x^2}}$	π	0	1/4, de $\beta_1 = 1/2$	1	x	$x^2 - 1/2$
Laguerre	$[0,\infty]$	e^{-x}	1	$2n+1$	n^2	1	$x-1$	$x^2 - 4x + 2$
Hermite	$[-\infty,\infty]$	e^{-x^2}	$\sqrt{\pi}$	0	$n/2$	1	x	$x^2 - 1/2$

19.3. Példák

1. Három-pontos Gauss-Csebisev kvadratúrával közelítjük a $\int_{-1}^{1}\left(1-x^2\right)^{-1/2}e^{-x}dx$ integrált. Becsüljük meg a hibát!

Megoldás. A három-pontos kvadratúránál $n=2$, ezt alkalmazzuk a (19.2) formulánál: $M_6=e$ és mivel 1-főegyütthatós polinomoknak kell szerepelni, emiatt $p_3(x)=T_3(x)/4$. Így a hiba kisebb, mint $e\cdot(T_3,T_3)/(16\cdot 6!)=e\pi/(32\cdot 720)$, mert $(T_3,T_3)=\pi/2$, (lásd a 7.4 feladatot).

2. Készítsük el a két-pontos Gauss-Hermite kvadratúra súlyait! Ellenőrízzük, hogy az így kapott kvadratúra legfeljebb harmadfokú polinomokra pontos!

Megoldás. A két-pontos kvadratúránál a másodfokú Hermite-polinom gyökei $-x_0=x_1=2^{-1/2}$. Így $a_0=\int_{-\infty}^{\infty}\frac{(x-x_1)}{(x_0-x_1)}\exp(-x^2)dx=\frac{x_1\mu_0}{2x_1}=\mu_0/2$, mert az elsőfokú tag integrálja zérus, lévén az integrandus páratlan függvény. Hasonlóan adódik, hogy $a_1=a_0$. A kapott kvadaratúra pontos az 1 függvényre, mert az eredmény μ_0. Az x és x^3 függvényre is pontos, mert a két tag a gyökhelyeken a páratlanság miatt kiejti egymást. Így már csak azt kell igazolunk, hogy a pontosság a másodfokú x^2 polinomra is teljesül. Ennek az integrálja a (9.8) formula alapján nem más mint $(p_1,p_1)=\beta_1(p_0,p_0)$, ami az előző oldalon látható táblázat szerint egyenlő $\mu_0/2$-vel. A Gauss-Hermite kvadratúra eredménye pedig $\frac{\mu_0}{2}(\frac{1}{2}+\frac{1}{2})$, tehát igaz az egyezés.

3. Határozzuk meg a következő integrál értékét:

$$\int_{-1}^{1}\frac{(2x^2+x)dx}{\sqrt{1-x^2}}.$$

Megoldás. A számlálóban szereplő polinomot előállítjuk az első három Csebisev polinom lineáris kombinációjaként: $2x^2+x=c_0T_0+c_1T_1+c_2T_2$. Ezt felhasználva az integrálunk így is írható: $(T_0,c_0T_0+c_1T_1+c_2T_2)=c_0(T_0,T_0)=c_0\pi$. Az első három Csebisev polinom együtthatóival a következő lineáris egyenletrendszert írhatjuk fel (9.5) alapján:

$$\begin{bmatrix}1 & 0 & -1\\ & 1 & 0\\ & & 2\end{bmatrix}\begin{bmatrix}c_0\\ c_1\\ c_2\end{bmatrix}=\begin{bmatrix}0\\ 1\\ 2\end{bmatrix},$$

amiből $c_0=1$, tehát az integrál értéke π.

20. Közönséges differenciálegyenletek

20.1. Alapfogalmak

Az

$$F(x, y, y', y'', \ldots, y^{(m)}) = 0 \tag{20.1}$$

alakú egyenletet *közönséges differenciálegyenletnek* nevezzük, ahol keresendő az $y = y(x)$ függvény, mely a fenti kifejezésben deriváltjaival együtt szerepel. Ha y m-szer differenciálható $[0,1]$-ben és kielégíti (20.1)-et, akkor y egy *megoldás*. Ismeretes, a differenciálegyenleteknek sok megoldásuk van. A megoldást úgy tudjuk *egyértelművé* tenni, hogy a peremen még feltételeket teszünk a függvény vagy deriváltjainak értékére. Ha ezen feltételek mindegyike a kezdőpontban (általánosabban fogalmazva: csak egy pontban) van megadva, akkor *kezdetiérték feladatról* beszélünk, ha pedig több ponthoz kapcsolódnak a feltételek, akkor *peremérték feladatunk* van.

A differenciálegyenlet *lineáris,* ha y és deriváltjainak lineáris kombinációja szerepel (20.1)-ben és a differenciálegyenlet *explicit* alakú, ha a legmagasabb derivált explicit módon kifejezhető:

$$y^{(m)} = f(x, y, y', \ldots). \tag{20.2}$$

Világos, minden lineáris differenciálegyenlet ebbe a kategóriába tartozik, és a gyakorlatban előforduló nemlineáris differenciálegyenletek zöme is ilyen. A továbbiakban explicit elsőrendű differenciálegyenletek megoldásának numerikus közelítéseivel fogunk foglalkozni, amikor kezdetiérték feladatról van szó. Ekkor a kezdetiérték feladat

$$y(0) = y_0, \quad y' = f(x, y), \quad x \in [0,1], \tag{20.3}$$

ahol $f : \mathbb{R}^2 \to \mathbb{R}$, x_0 és y_0 adottak, és keressük azt az $y : [0,1] \to \mathbb{R}$ függvényt, amely felveszi x_0-ban az y_0 értéket és az egyenletet kielégíti. Az egyszerűség kedvéért szoritkozunk a $[0,1]$ intervallumra, hiszen tudjuk, az $[a,b]$ intervallum lineáris transzformációval ide átvihető. A megoldás létezésére és egyértelműségére vonatkozik a következő tétel, melyet bizonyítás nélkül idézünk.

20.1.1 Tétel, a kezdetiérték feladat egyértelműsége

Ha f folytonos egy $(x, y) \in [0,1] \times [c,d]$ téglán és f a második változója szerint eleget tesz a Lipschitz-feltételnek, azaz létezik L:

$$|f(x, y_1) - f(x, y_2)| \le L|y_1 - y_2|, \quad \forall x \in [0,1], \quad y_1, y_2 \in [c,d], \tag{20.4}$$

akkor a kezdetiérték feladatnak egyértelmű megoldása van.

A numerikus eljárás során a megoldást az $x_n = nh$ osztópontokban közelítjük, ahol $h = 1/N$ és a numerikus közelítést x_n-ben y_n-nel jelöljük. Természetesen azt szeretnénk, hogy y_n minél közelebb legyen a pontos értékhez, $y(x_n)$-hez.

20.1.2 Az Euler-módszer

Ez a legegyszerűbb módszer, amit a differencia-hányadosból származtatunk a következő közelítő érték előállítására:

$$\frac{y_{n+1} - y_n}{h} = f(x_n, y_n) \quad \to \quad y_{n+1} = y_n + hf(x_n, y_n). \tag{20.5}$$

Példa: $y(0)=1,\ y'=y$. Ennek a megoldása $y=e^x$ és (20.5)-ből $y_n=(1+h)^n$. Ha $n\to\infty$, akkor az analízisből tudjuk, hogy $y_n=(1+h)^n=(1+x_n/n)^n \ \to\ e^{x_n}$.

20.1.3 Definíció, konvergencia

Egy numerikus módszer *lokálisan konvergens* az $x\in[0,1]$ pontban, ha $h=x/n,\ x=x_n$ mellett

$$\lim_{n\to\infty} y_n = y(x)$$

teljesül. Ha a módszer konvergens $\forall x\in[0,1]$ -re, akkor azt mondjuk, a *módszer konvergens.*

20.1.4 Definíció, konvergencia-sebesség

Egy konvergens numerikus módszer sebessége p -edrendű, $(1\le p)$, ha $h=x/n,\ x=x_n$ mellett $\exists M$ úgy, hogy az

$$\left|y(x)-y_n\right|\le Mh^p \tag{20.6}$$

hibabecslés teljesül, ahol M független h -tól és n -től.

20.1.5 Definíció, lokális hiba v. képlethiba

Ez annak a képletnek a hibája, amellyel a következő függvényértéket közelítjük. Ilyenkor a képletbe mindenütt a pontos értéket írjuk. Például az Euler-módszernél a

$$g_{i-1}=y(x_i)-y(x_{i-1})-hf(x_{i-1},y(x_{i-1})),\ \ i=1,2,\ldots,N \tag{20.7}$$

mennyiségek a *lokális hibák vagy képlethibák.*

20.1.6 Definíció, konzisztencia

Ha $\exists p\ge 1$ és M konstans úgy, hogy

$$\left|g_i\right|\le Mh^{p+1},\ \ i=1,2,\ldots,N \tag{20.8}$$

akkor a módszer p *-edrendben konzisztens.*

A definició alapján világos, nagyobb p -re pontosabb megoldás várható.

20.1.7 Definició, globális hiba

Jelölje a pontos és numerikus megoldás eltérését

$$e_i=y(x_i)-y_i,\ \ i=1,2,\ldots,N,\ \ h=1/N,\ \ x_i=ih. \tag{20.9}$$

Az e_i mennyiségeket *globális hibának* nevezzük.

20.1.8 Definició, stabilitás

A numerikus módszer stabil, ha van olyan K konstans, amellyel

$$\left|e_i\right|\le K\left(\left|e_0\right|+\sum_{j=1}^{i}\left|g_j\right|\right),\ \ i=1,2\ldots,N \tag{20.10}$$

teljesül, vagyis a globális hiba felülről becsülhető a lokális hibák abszolút összegével.

20.1.9 Tétel, numerikus módszer konvergenciája

Ha egy numerikus módszer stabil és p-edrendben konzisztens, akkor p-edrendben konvergens.

Bizonyítás. Ha $x=0$, akkor a tétel igaz. Ha $x\in(0,1]$, legyen $h=x/n$, $x_n=x$ $\forall n$-re. Először a stabilitás, majd a konzisztencia definícióját felhasználva

$$|e_i|\le K\left(|e_0|+\sum_{j=1}^{n}|g_j|\right)\le K\sum_{j=1}^{n}ch^{p+1}\le$$
$$\le Kcnh^{p+1}\le Kc(nh)h^p\le(Kc)h^p,$$

ahol $e_0=y(x_0)-y_0=0$. Ezzel p-edrendű konvergencia-sebességre jutottunk. ■

20.2. Az Euler-módszer vizsgálata

Látjuk, az Euler-módszerre a konzisztenciát és stabilitást kéne megmutatni ahhoz, hogy a konvergenciát igazoljuk.

20.2.1 Tétel, konzisztencia

Az Euler-módszer elsőrendben konzisztens, vagyis $g_i=O(h^2)$ teljesül, ha a megoldás kétszer folytonosan differenciálható.

Bizonyítás. Tekintsük a lokális hibát az i-edik pontban és $y(x_{i+1})$-et fejtsük sorba az x_i hely körül:

$$g_i=y(x_{i+1})-y(x_i)-hf(x_i,y(x_i))=$$
$$=y(x_i)+hy'(x_i)+\frac{h^2}{2!}y''(\xi_i)-y(x_i)-hy'(x_i),$$

emiatt

$$|g_i|=\frac{h^2}{2!}|y''(\xi_i)|\le\|y''\|_\infty\frac{h^2}{2!},$$

tehát $p+1=2$, $p=1$, elsőrendű a konzisztencia. ■

20.2.2 Tétel

Az Euler-módszer stabil, ha f eleget tesz a második változója szerint a Lipschitz-feltételnek.

Bizonyítás. Vegyük az i-edik pontban a lokális hiba képletét, a módszer képletét és vonjuk ki őket egymásból:

$$y(x_{i+1})=y(x_i)+hf(x_i,y(x_i))+g_i \qquad /(+)$$
$$y_{i+1}=y_i+hf(x_i,y_i) \qquad /(-)$$
$$e_{i+1}=e_i+h[f(x_i,y(x_i))-f(x_i,y_i)]+g_i$$

Mivel f eleget tesz a Lipschitz-feltételnek:

$$|e_{i+1}|\le|e_i|+h|f(x_i,y(x_i))-f(x_i,y_i)|+|g_i|\le|e_i|(1+hL)+|g_i|.$$

Fejtsük vissza a rekurziót e_0-ig:

$$|e_{i+1}| \le |e_i|(1+hL) + |g_i| \le \big(|e_{i-1}|(1+hL) + |g_{i-1}|\big)(1+hL) + |g_i| \le$$
$$\le (1+hL)^2 |e_{i-1}| + (1+hL)|g_{i-1}| + |g_i| \le (1+hL)^{i+1}|e_0| + \sum_{k=0}^{i} (1+hL)^{i-k} |g_k|,$$

ebből megkapjuk a kívánt becslést, ha felhasználjuk az $(1+hL)^j \le e^{jhL} = e^{x_j L} \le e^L$ relációt:

$$|e_{i+1}| \le e^L |e_0| + \sum_{k=0}^{i} e^L |g_k| = e^L \left(|e_0| + \sum_{k=0}^{i} |g_k| \right). \qquad \blacksquare$$

A továbbiakban rátérünk néhány fontosabb módszercsalád rövid ismertetésére.

20.3. Taylor-polinomos módszerek

Az Euler-módszer nem egyéb, minthogy vesszük a függvény elsőfokú Taylor-polinomját x_n körül és annak segítségével lépünk a következő pontba. Ebből az ötletből kiindulva készíthetünk magasabbrendű módszereket is. A következő módszert m-edrendű Taylor-polinomos módszernek nevezzük:

$$y_{n+1} = y_n + y'(x_n) + \ldots + y^{(m)}(x_n) h^m / m!, \quad n = 0,1,\ldots,N. \tag{20.11}$$

Fontos kérdés, egyáltalán kiszámíthatók-e ezek a deriváltak. Ha f elegendően sokszor differenciálható, ennek elvi akadálya nincs. Sokszor azonban súlyos gyakorlati nehézség, hogy a deriváltak nagyon hosszú, nehezen kezelhető formulákat eredményeznek. A konstrukcióból látható, m-edrendben konzisztens módszerre vezet a fenti eljárás.

20.4. Runge-Kutta módszerek

Ezek a Taylor-polinomos módszerek fenti nehézségét küszöbölik ki: nem kell magasrendű deriváltakat számolni, *a magasabb konzisztencia-rend rekurzív függvényhívásokkal is elérhető.* Az általános alak:

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + ha_2, y_n + hb_{21}k_1), \\ &\vdots \\ k_j &= f(x_n + ha_j, y_n + h\sum_{l=1}^{j-1} b_{jl}k_l), \quad j = 1,2,\ldots,s, \\ y_{n+1} &= y_n + h\sum_{j=1}^{s} c_j k_j. \end{aligned} \tag{20.12}$$

Az előre kiszámolt a_i, b_{ij}, c_i paraméterek határozzák meg a konkrét módszert. Az s szám a rekurzió mélysége, más szóval: az egy lépés megtételéhez szükséges függvényhívások száma. A következő, ún. *módosított Euler-módszer* Runge-től származik:

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + h/2, y_n + (h/2)k_1), \\ y_{n+1} &= y_n + hk_2. \end{aligned} \tag{20.13}$$

Deriválással megmutatható, hogy másodrendben konzisztens. Hasonlóképp másodrendű a következő Runge-Kutta módszer, amelyet az egyszerűsége miatt egy sorban írunk fel:

$$y_{n+1} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_n + hf(x_n, y_n))]. \tag{20.14}$$

Az igen népszerű negyedrendű Runge-Kutta módszer negyedrendben konzisztens és stabil, vagyis negyedrendű konvergens módszer:

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + h/2, y_n + hk_1/2), \\ k_3 &= f(x_n + h/2, y_n + hk_2/2), \\ k_4 &= f(x_n + h, y_n + hk_3), \\ y_{n+1} &= y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4). \end{aligned} \tag{20.15}$$

A Runge-Kutta módszerek lokális hibája:

$$g_n = y(x_n) - y(x_{n-1}) - h\sum_{j=1}^{s} c_j k_j(x_{n-1}, y(x_{n-1}), h), \quad n = 1, \ldots, N, \tag{20.16}$$

ahol $k_j(x_{n-1}, y(x_{n-1}), h)$ azt jelenti, hogy a k_j sorozatot a pontos $y(x_{n-1})$ értékkel indítva számoljuk végig. Egy adott rend eléréséhez a lokális hibát sorbafejtjük és a paramétereket úgy választjuk, hogy a tagok minél magasabb rendig eltűnjenek.

20.5. Lineáris többlépéses módszerek

Ezek is az Euler-módszer általánosításának tekinthetők. Az s -lépéses módszer általános alakja:

$$\sum_{k=0}^{s} \alpha_k y_{i+k} = h\sum_{k=0}^{s} \beta_k f_{i+k}, \quad f_{i+k} = f(x_{i+k}, y_{i+k}) \tag{20.17}$$

ahol α_k, β_k, $k = 0,1,\ldots,s$ adottak, $\alpha_s = 1$ valamint $|\alpha_0| + |\beta_0 \neq 0|$ teljesül. Például az Euler-módszerre $s = 1,\ \alpha_0 = -1,\ \alpha_1 = 1,\ \beta_0 = 1,\ \beta_1 = 0.$ A módszer indításához az y_0 értékén túl kellenek még az $y_1, y_2, \ldots, y_{s-1}$ értékek. Ha $\beta_s = 0$, akkor a módszer *explicit* és könnyen számolható. Ha $\beta_s \neq 0$, akkor a módszer implicit és a következő y_{i+s} érték az adódó nemlineáris egyenletből számítandó. Vegyük észre, az explicit módszernél minden lépésben egy új $f(x, y)$ típusú függvény-kiértékelés kell, (mivel a többi már korábbról megvan), ugyanakkor az implicit módszernél még ügyes esetben is legalább 2-3 kiértékelésre szükség van. A mondottakat két egyszerű példán szemléltejük.

Középpontszabály. Explicit, 2-lépéses módszer, a centrális differenciahányadosból származtatható:

$$y_{n+1} = y_{n-1} + 2hf_n, \quad n = 1,2,\ldots,N \tag{20.18}$$

tehát $\alpha_0 = -1,\ \alpha_1 = 0,\ \alpha_2 = 1,\ \beta_0 = 0,\ \beta_1 = -2,\ \beta_2 = 0.$ Bebizonyítható, hogy a módszer másodrendben konzisztens és stabil, ha f a második változójában eleget tesz a Lipschitz-feltételnek. A középpontszabály indításához legalább másodrendűen pontos y_1 értéket kell előállítani, amit megtehetünk valamely Runge-Kutta vagy Taylor-polinomos módszerrel, különben a másodrendű konvergencia nem lesz igaz.

Trapézszabály. Ez implicit módszer. Úgy származtatható, hogy a differenciálegyenletet átírjuk integrál alakba és a jobb oldalon keletkező integrált a trapézmódszerrel közelítjük:

$$y_{n+1} = y_n + \frac{h}{2}\left(f_n + f_{n+1}\right), \quad n = 0,1,\ldots,N-1. \tag{20.19}$$

így $s = 1,\ \alpha_0 = -1,\ \alpha_1 = 1,\ \beta_0 = \beta_1 = 1/2.$ Itt a következő pont előállításához meg kell oldanunk y -ra az

$$y = y_n + \frac{h}{2}\left(f_n + f(x_n + h, y)\right) = F(y)$$

fixpont egyenletet. $F(y)$ az általános feltételünk alapján eleget tesz a Lipschitz-feltételnek $hL/2$ állandóval, így $F(y)$ kontrakció, ha $h<2/L$. Célszerű formája az iterációnak:

$$\begin{aligned} y_{n+1}^{0} &= y_n + hf_n, \qquad \text{kezdőérték az Euler-módszerből} \\ y_{n+1}^{k+1} &= y_n + \frac{h}{2}\left(f_n + f(x_{n+1}, y_{n+1}^{k}\right), \quad k \geq 0. \end{aligned} \tag{20.20}$$

Leállás: ha $\left|y_{n+1}^{k+1} - y_{n+1}^{k}\right| < \varepsilon(1-q)$, ahol q a konvergencia-tényező és ε a kívánt hibakorlát.

A trapézmódszer másodrendben konzisztens és stabil, ha $f \in C^2([0,1]\times\mathbb{R})$ és $h<1/L$, tehát ekkor másodrendű konvergenciára számíthatunk.

További többlépéses módszereket a zárt vagy a jobbról nyilt kvadratúra-formulák segítségével lehet származtatni.

20.6. Aszimptotikus stabilitás

A gyakorlati számítások szempontjából nem elegendő, ha egy módszer konzisztens és stabil. A kezdeti hiba (- ha y_0 is hibával terhelt, -) továbbterjedése szempontjából fontos egy további stabilitási tulajdonság. Egy differenciálegyenlet-megoldó numerikus módszer *aszimptotikusan stabil,* ha az

$$y(0)=1, \;\; y'(x) = q\,y(x), \;\; q<0 \tag{20.21}$$

tesztfeladatra alkalmazva a kapott numerikus közelítések sorozatára minden $h>0$ lépésköz esetén fennáll $y_n \to 0$, ha $n \to \infty$. Ezt a stabilitási fogalmat a szakirodalom gyakran A_0-stabilitásnak nevezi.

A tesztfeladat megoldása x növekedésével zérushoz tart: $y(x) = e^{qx} \to 0, \;\; x \to \infty$, mivel q negatív. Így a definició azt követeli a módszertől, hogy a lépésköztől függetlenül a numerikus megoldás is tartsa meg ezt a lecsengő jelleget. Kimutatható, az Euler-módszer nem aszimptotikusan stabil, de a trapézmódszer az.