BEVEZETÉS A NUMERIKUS ANALÍZISBE

Hartung Ferenc

Pannon Egyetem

2020

Tartalomjegyzék

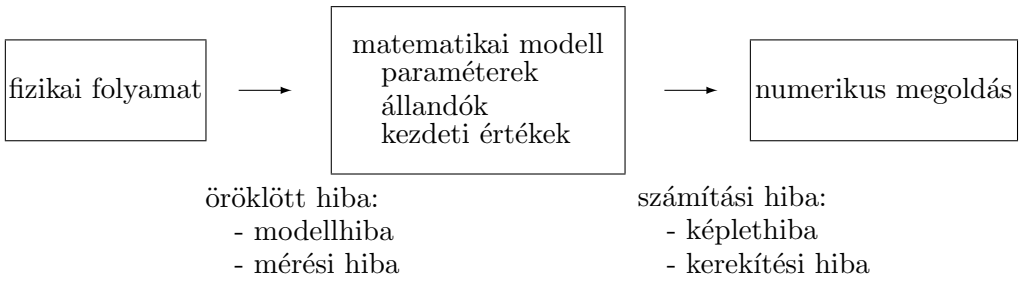
1. fejezet

Bevezetés

Ebben a fejezetben először a numerikus analízis feladatát ismertetjük, majd alapfogalmakat vizsgálunk. A matematikai számítások közben felmerülő hibák több fajtáját definiáljuk, bevezetjük egy matematikai feladat illetve egy numerikus algoritmus stabilitásának, műveletigényének, tárolási igényének fogalmát. Ezután az egész és valós számok számítógépen történő tárolásának különféle szabványait ismertetjük, és a véges sok számjegyen történő tárolás miatt fellépő problémákat vizsgáljuk.

1.1. A numerikus analízis feladata, alapfogalmak

A fizikai valóság folyamatainak leírására és a vizsgált fizikai változók jelen ill. jövőbeli értékének meghatározására szolgáló matematikai számítások vázlatos menetét az 1.1. ábrával lehet szemléltetni.



1.1. ábra.

Az első lépés a vizsgált folyamat leírása, matematikai modellezése. Ez a lépés a megfelelő tudományág (fizika, kémia, biológia, közgazdaságtan stb.) feladata. A kapott modellben sokszor szerepelnek paraméterek, állandók, kezdeti feltételek, amelyeket általában megfigyeléssel, méréssel állapíthatunk meg. Ha a modell és az abban szereplő paraméterek ismertek, akkor használhatjuk a matematikai modellt a fizikai rendszerre vonatkozó kérdések megválaszolására. A fizikai rendszerre, ill. az azt leíró matematikai modellre feltehetünk kvalitatív kérdéseket (pl. létezike egyértelmű megoldása a matematikai feladatnak, van-e határértéke a vizsgált változónak, periodikus-e a folyamat stb.) vagy kvantitatív kérdéseket (mi a vizsgált fizikai változó értéke egy adott időpontban, mi a pontos vagy közelítő megoldása a matematikai modellnek stb.). A kvalitatív kérdésekre az adott modellhez tartozó matematikai szakterület keresi a választ, a kvantitatív kérdésekkel pedig a numerikus analízis foglalkozik. A numerikus analízis feladata matematikai feladatok numerikus eredményének aritmetikai műveletekkel (osztás, szorzás, összeadás, kivonás) való pontos vagy közelítő megoldása.

Az 1.1. ábrán leírt folyamattal számolt fizikai változó értéke általában nem egyezik meg pontosan a változó tényleges értékével. Az elkövetett hibát két nagyobb kategóriára bontjuk:

öröklött hiba és *számítási hiba*. Az öröklött hiba az első lépésben, azaz a fizikai folyamat matematikai modellel való helyettesítésekor elkövetett hiba. Ezt is két részre oszthatjuk: *modellhiba* és *mérési hiba*. A modellhiba abból adódik, hogy a matematikai modellek levezetésére használt törvények idealizáltak, csak „közelítései" a valóságnak. A mérési hiba az a hiba, amit azáltal kapunk, hogy a matematikai modellben a valódi paramétereknek, kezdeti feltételeknek csak mért, így közelítő értékét használjuk a tényleges értékek helyett.

A számítási hibát is két részre bontjuk, *képlethibára* és *kerekítési hibára*. A képlethiba az a hiba, amit akkor követünk el, amikor egy matematikai kifejezés pontos értéke helyett annak közelítő értékét használjuk.

1.1. példa. Tegyük fel, hogy az $f(x) = \sin x$ függvény értékét kell kiszámítanunk egy megadott x pontban. Az $f(x)$ függvényérték helyett kiszámíthatjuk pl. az f függvény ötödrendű Taylor-polinomját: $T_5(x) = x - x^3/3! + x^5/5!$. A Taylor-tétel (2.5. tétel) szerint ha $f(x)$-et $T_5(x)$-szel helyettesítjük, akkor a közelítés hibája $\frac{f^{(6)}(\xi)}{6!}x^6 = -\frac{\sin\xi}{6!}x^6$ alakban írható fel. Ez a hiba (azaz a módszerünk képlethibája) kicsi, ha x közel van 0-hoz. □

A kerekítési hiba abból adódik, hogy a számítógépen egy valós számot csak véges sok tizedesjegy pontossággal tudunk tárolni, így általában már a valós számok tárolásakor követünk el hibát. Kerekítési hiba lép fel az aritmetikai műveletek végzése közben is: a számítógép az egyes aritmetikai műveletek eredményeit adott számú tizedesjegyre kerekítés után tárolja/használja tovább. A kerekítési hibával bővebben az 1.2.–1.4. szakaszokban foglalkozunk.

Egy numerikus módszer levezetésekor az első kérdés amit vizsgálnunk kell, a módszer képlethibája, hiszen egy numerikus közelítő érték csak akkor hasznos, ha azt is tudjuk róla, hogy mekkora hibával közelíti a pontos értéket. A következő fogalom, ami egy numerikus módszerrel kapcsolatban felmerül, a *stabilitás*. Ezt a fogalmat kétféle értelemben is használjuk. Beszélhetünk egy matematikai modell vagy feladat stabilitásáról, vagy egy numerikus módszer stabilitásáról. Kezdjük egy példával.

1.2. példa. Tekintsük a

$$\begin{aligned} 8x + 917y &= 1794 \\ 7x + 802y &= 1569. \end{aligned}$$

lineáris egyenletrendszert. Ennek megoldása $x = -5$ és $y = 2$. Ha viszont a második egyenletben x együtthatóját 7.01-re[1] változtatjuk, akkor a

$$\begin{aligned} 8x + 917y &= 1794 \\ 7.01x + 802y &= 1569 \end{aligned}$$

egyenletrendszer megoldása $x = -1.232562589$ és $y = 1.967132499$ lesz (9 tizedesjegy pontossággal). Azt tapasztaljuk, hogy 0.14%-os változás az együtthatóban a megoldás 75.3%-os ill. 1.6%-os változását eredményezte. □

Azt mondjuk, hogy egy matematikai feladat *korrekt* vagy *stabil*, ha „kis" változás a feladat paramétereiben a megoldás „kis" változását idézi csak elő. Ellenkező esetben *inkorrekt* vagy *instabil feladatról* beszélünk. Az előző példában vizsgált egyenletrendszer tehát e szerint a terminológia szerint egy inkorrekt feladat.

[1]Ebben a jegyzetben, hogy a számítógépek és programozási nyelvek által használt jelöléssel összhangban legyünk, a törtszámoknál tizedespontot használunk tizedesvessző helyett.

Egy numerikus algoritmust a kerekítési hibákra nézve *stabilnak* nevezünk, ha a kerekítési hibák nem befolyásolják jelentősen a számított végeredményt. Ha a kerekítési hibák miatt a számított végeredmény jelentősen eltér a számítandó értéktől, akkor az algoritmust *instabilnak* nevezzük. A következő példában egy instabil algoritmust mutatunk be.

1.3. példa. Tekintsük a következő három, rekurzív definícióval megadott sorozatot:

$$\begin{aligned} x_n &= \frac{1}{3}x_{n-1}, \qquad x_0 = 1, \\ y_n &= 2y_{n-1} - \frac{5}{9}y_{n-2}, \qquad y_0 = 1, \quad y_1 = \frac{1}{3}, \\ z_n &= \frac{13}{3}z_{n-1} - \frac{4}{3}z_{n-2}, \qquad z_0 = 1, \quad z_1 = \frac{1}{3}. \end{aligned} \tag{1.1}$$

Könnyen látható, hogy mindhárom képlet az $x_n = y_n = z_n = \frac{1}{3^n}$ számsorozatot definiálja, azaz a három sorozat algebrailag ekvivalens. A gyakorlatban viszont észrevehető különbség van a három képlet között. Az 1.1. táblázatban kinyomtattuk az (1.1) képlettel számított sorozatok első 18 tagjait. A számításokat egyszeres pontossággal végeztük csak, hogy a kerekítési hibák hatását jobban láthassuk. Azt tapasztaljuk, hogy x_n tényleg az $1/3^n$ értékeket állítja elő, viszont az y_n és z_n számolt értékeiben kerekítési hibából származó eltérést láthatunk. Mindkét sorozatnál fellép a hiba, de a z_n esetében a hiba igen gyorsan nő, a 18. tagnál már 100-as nagyságrendű. Azt tapasztaltuk tehát, hogy az x_n képlete egy stabil, a z_n képlete pedig egy instabil módszer $1/3^n$ generálására.

Arról, hogy az előbb vizsgált hibák tényleg a kerekítési hibák rovására írhatók, meggyőződhetünk úgy, hogy megismételjük a számításokat, de most dupla pontosságot használva a számok tárolásához. Az egész listát nem, csak a 18. tag hibáját közöljük: $|y_{18} - 1/3^{18}| = -2.5104e-13$ és $|z_{18} - 1/3^{18}| = 2.3804e-07$. Látható, hogy ebben az esetben a számolás hibája sokkal kisebb. □

1.1. táblázat.

n	x_n	y_n	$\lvert y_n - 1/3^n\rvert$	z_n	$\lvert z_n - 1/3^n\rvert$
2	0.111111	0.111111	2.2352e-08	0.111111	4.4703e-08
3	0.037037	0.037037	4.0978e-08	0.037037	1.8254e-07
4	0.012346	0.012346	6.9849e-08	0.012346	7.3109e-07
5	0.004115	0.004115	1.1688e-07	0.004118	2.9248e-06
6	0.001372	0.001372	1.9465e-07	0.001383	1.1699e-05
7	0.000457	0.000458	3.2442e-07	0.000504	4.6795e-05
8	0.000152	0.000153	5.4071e-07	0.000340	1.8718e-04
9	0.000051	0.000052	9.0117e-07	0.000800	7.4872e-04
10	0.000017	0.000018	1.5019e-06	0.003012	2.9949e-03
11	0.000006	0.000008	2.5032e-06	0.011985	1.1980e-02
12	0.000002	0.000006	4.1721e-06	0.047920	4.7918e-02
13	0.000001	0.000008	6.9535e-06	0.191674	1.9167e-01
14	0.000000	0.000012	1.1589e-05	0.766693	7.6669e-01
15	0.000000	0.000019	1.9315e-05	3.066773	3.0668e+00
16	0.000000	0.000032	3.2192e-05	12.267091	1.2267e+01
17	0.000000	0.000054	5.3653e-05	49.068363	4.9068e+01
18	0.000000	0.000089	8.9422e-05	196.273453	1.9627e+02

A következő fogalom, amit egy (véges sok lépésből álló) numerikus módszernél vizsgálni szoktunk, az algoritmus *műveletigénye* vagy *műveletszáma*. Tekintsünk először egy példát:

1.4. példa. Számítsuk ki a $p(x) = p(x) = 5x^4 - 8x^3 + 2x^2 + 4x - 10$ negyedfokú polinom értékét egy megadott x pontban! Természetesen ezt könnyen megtehetjük p képletét és aritmetikai műveleteket használva. A képletben 4 összeadás/kivonás, 4 szorzás és 3 hatványozás szerepel. A hatványozások tulajdonképpen 3+2+1=6 szorzást jelentenek, azaz összesen 10 szorzásra van szükség a képlet alkalmazásához. Megtehetjük viszont, hogy átalakítjuk p képletét:

$$p(x) = 5x^4 - 8x^3 + 2x^2 + 4x - 10 = (((5x - 8)x + 2)x + 4)x - 10.$$

A p polinomnak ezt az alakját használva már csak 4 összeadás ill. kivonás valamint 4 szorzás kell a képlet kiértékeléséhez. □

Az előző példában bemutatott eljárást megismételhetjük általános n-edfokú polinomokra:

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = ((\cdots((a_n x + a_{n-1})x + a_{n-2})x + \cdots)x + a_1)x + a_0$$

Ebben a képletben összesen csak n összeadás/kivonás és n szorzás szerepel. Ezt a polinomok kiértékelésére vonatkozó módszert *Horner-eljárásnak* nevezzük. A módszert az 1.5. algoritmussal írhatjuk le.

1.5. algoritmus. Horner-eljárás

INPUT: n - a polinom fokszáma
$a_n, a_{n-1}, \ldots, a_0$ - a polinom együtthatói
x - ahol a polinomot kiértékeljük
OUTPUT: p - a polinom értéke az x pontban

$p \leftarrow a_n$
for $i = n-1, \ldots, 0$ **do**
  $p \leftarrow a_i + px$
end do
output(p)

A számítógépeken egy szorzás ill. osztás elvégzése jelentősen tovább tart, mint egy összeadás vagy kivonás. Ezért egy algoritmus műveletigényén általában a benne szereplő osztások/szorzások számát szokás érteni.

Egy algoritmusra jellemző tulajdonság még az *adattárolási igénye*. Egy 10×10-es lineáris egyenletrendszer megoldására használt algoritmus esetében az adatok tárolása nem jelenthet problémát, de ugyanez 10000×10000-es rendszerre már gond lehet. Ilyen mennyiségű adat kezelésekor előnyben részesítünk olyan algoritmusokat, amelyeknek minél kisebb az adattárolási igénye. Például, ha tudjuk, hogy az együtthatómátrixban csak a főátlóban és az alatta ill. felette levő néhány átlóban vannak csak nem nulla elemek (ún. szalagmátrix), akkor mindenképpen célszerű olyan algoritmust használni, amely kihasználja az adatok speciális szerkezetét, és nem tárolja számolás közben a felesleges nullákat. Ilyen módszerre látunk majd példát a 3.5. szakaszban.

1.2. Egész és valós számok tárolása

Legyen I egy b-alapú számrendszerben felírt m jegyű pozitív egész szám:

$$I = (a_{m-1}a_{m-2}\ldots a_1 a_0)_b, \qquad \text{ahol} \quad a_i \in \{0, 1, \ldots, b-1\}.$$

Ennek értéke:

$$I = a_{m-1}b^{m-1} + a_{m-2}b^{m-2} + \cdots + a_1 b + a_0.$$

m jegyen tárolható legnagyobb egész szám tehát az az $I_{\max}$ szám, amelynek minden számjegye $b-1$. Ennek értéke

$$I_{\max} = (b-1)(b^{m-1} + b^{m-2} + \cdots + b + 1) = b^m - 1.$$

m jegyen tehát a 0-tól b^m-1-ig terjedő (b^m db) nemnegatív számokat tudjuk ábrázolni (tárolni). A számítógépen a kettes (bináris) számrendszerben tároljuk az egész számokat. m biten tehát 2^m db számot tudunk ábrázolni. Negatív egész számok tárolására két módszert ismertetünk. Az első az ún. *direkt* vagy *egyenes kód*. Ebben a kódolásban egy bitet lefoglalunk az előjelnek, (ezt hívjuk előjelbitnek), és a maradék $m-1$ biten tudjuk a szám abszolút értékét tárolni. Ekkor $I_{\max}=2^{m-1}-1$, és a legkisebb tárolható egész szám $I_{\min}=-I_{\max}$. Ebben a kódolásban a 0-t kétféleképpen tárolhatjuk: a csupa 0, ill. az $100\ldots0$ bitsorozattal.

1.6. példa. Az 1.2. táblázatban az $m=3$ biten, direkt kóddal ábrázolható számokat soroltuk fel. □

1.2. táblázat. Egyenes kód, $m=3$

I	a tárolt bináris kód
0	000
1	001
2	010
3	011
0	100
-1	101
-2	110
-3	111

A gyakorlatban általában az ún. *kettes komplemens kódot* szokták használni negatív számok tárolására. Legyen I egy egész szám, amit m biten szeretnénk tárolni. Az I helyett a C szám bináris alakját tároljuk, ahol

$$C=\begin{cases} I, & \text{ha } 0\le I\le 2^{m-1}-1,\\ 2^m+I, & \text{ha } -2^{m-1}\le I<0.\end{cases}$$

Ennél a tárolásnál tehát a legnagyobb és a legkisebb ábrázolható szám $I_{\max}=2^{m-1}-1$ ill. $I_{\min}=-2^{m-1}$. A feltételek szerint ha $0\le I\le 2^{m-1}-1$, akkor $C<2^{m-1}$, azaz C első bitje 0. Ha viszont $-2^{m-1}\le I<0$, akkor könnyen ellenőrizhető, hogy $2^{m-1}\le C\le 2^m-1$, azaz C első bitje 1.

A kettes komplemens kód egyik fontos előnye, hogy segítségével a kivonás visszavezethető összeadásra. (Lásd a 4. feladatot!)

1.7. példa. Az 1.3. táblázat az $m=3$ biten, kettes komplemens kóddal ábrázolható számokat tartalmazza. □

1.3. táblázat. Kettes komplemens kód, $m=3$

I (decimálisan)	I (binárisan)	C, a tárolt bináris kód
0	000	000
1	001	001
2	010	010
3	011	011
-1	-001	111
-2	-010	110
-3	-011	101
-4	-100	100

A továbbiakban a valós számok tárolását vizsgáljuk. Emlékeztetünk arra, hogy a b alapú számrendszerben felírt

$$x=(x_{m-1}x_{m-2}\cdots x_0.x_{-1}x_{-2}\cdots)_b,\qquad x_i\in\{0,1,\ldots,b-1\},$$

valós szám értéke

$$x = x_{m-1}b^{m-1} + x_{m-2}b^{m-2} + \cdots x_1 b + x_0 + \frac{x_{-1}}{b} + \frac{x_{-2}}{b^2} + \cdots = \sum_{i=-\infty}^{m-1} x_i b^i.$$

Tekintsük a 126.42 valós számot. Ennek normál alakján vagy az $1.2642 \cdot 10^2$ vagy pedig a $0.12642 \cdot 10^3$ alakot szokás érteni. Mi ebben a jegyzetben az első alakot fogjuk használni. Ennek megfelelően egy $x \neq 0$ valós szám b alapra vonatkozó *normál alakján* az $x = \pm m \cdot b^k$ alakját hívjuk, ahol $1 \leq m < b$. m-et a szám *mantisszájának*, k-t pedig kitevőjének nevezzük. Valós számok, más szóval *lebegőpontos számok* tárolásához a számot felírjuk (valamely b alapot használva) normál alakban, és az előjeles mantisszát, valamint a kitevőt tároljuk. Különböző számítógépek eltérő alapot és bithosszúságot használnak egy valós szám tárolására. Mi most egy IEEE szabványt[2] ismertetünk valós számok 32 biten (ún. *egyszeres pontosságú*), ill. 64 biten történő (ún. *dupla pontosságú*) tárolására bináris alapot használva. Ezt a kódolást használják az IBM-kompatibilis személyi számítógépek is. Vegyük a szám $x = (-1)^s m \cdot 2^k$ bináris normál alakját, ahol $s \in \{0, 1\}$ és $m = 1.m_1m_2m_3\ldots$. Az s értékét az 1. biten tároljuk. A k kitevő helyett annak eltolt értékét, az $e = k + 127$ nemnegatív számot a 2.–9. biteken tároljuk, a mantissza törtrészének első 23 bitre kerekített értékét pedig a 10.–32. biten tároljuk. (A nemnulla szám mantisszájának egész része bináris normál alakban mindig 1-gyel egyenlő, ezt az 1-est nem tároljuk!) A fent említett IEEE szabvány külön definiálja a 0 tárolását, és bevezet két speciális szimbólumot is, az `Inf` (infinity, azaz végtelen) és `NaN` (not-a-number, azaz nem szám) szimbólumokat:

tárolandó szám	s	e (2.–9. bitek)	mantissza bitek (10.–32. bitek)
$+0$	0	00000000	minden mantissza bit=0
-0	1	00000000	minden mantissza bit=0
`+Inf`	0	11111111	legalább az egyik mantissza bit=0
`-Inf`	1	11111111	legalább az egyik mantissza bit=0
`+NaN`	0	11111111	minden mantissza bit=1
`+NaN`	1	11111111	minden mantissza bit=1

Az `Inf` szimbólumot a programok használhatják olyan matematikai művelet eredményének tárolására, amelynek értéke végtelen, a `NaN` szimbólumot pedig olyan művelet „eredményének” tárolására, amely nem definiált (pl. nullával való osztás eredménye vagy negatív szám négyzetgyöke valós számok körében). Mindkét szimbólumnak lehet pozitív vagy negatív előjele. A szabvány definíciójából következik, hogy az $e = (11111111)_2 = 255$ azaz a $k = 128$ kitevő az `Inf` és `NaN` speciális szimbólumoknak van fenntartva. A véges valós számok esetén $0 \leq e \leq 254$, így k lehetséges értékei $-127 \leq k \leq 127$. A legkisebb pozitív valós szám tehát a $k = -127$ kitevőhöz és az $(1.00\ldots01)_2$ mantisszához tartozik. Ennek értéke $(1 + 1/2^{23})2^{-127} \approx 10^{-38}$. A legnagyobb (véges) valós szám pedig $x_{\max} = (1.11\ldots1)_2 2^{127} = (2 - 2^{-23})2^{127} \approx 10^{38}$.

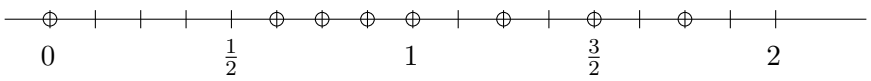
64 biten történő tárolás az előzőekhez hasonlóan történik: az $e = k + 1023$ eltolt kitevőt a 2.–12. biten, a mantissza törtrészét pedig a 13.–64. biten tároljuk. Ekkor a tárolható pozitív számok tartománya körülbelül 10^{-308} – 10^{308} lesz.

1.8. példa. Tegyük fel, hogy 4 biten szeretnénk valós számokat tárolni, bináris normál alak segítségével. Ezt megtehetjük pl. úgy, hogy az 1. biten a szám előjelét, a 2. biten a bináris normál alak eltolt kitevőjét, $e = k + 1$-et, a 3.–4. biten pedig a mantissza tört részének első két bitjét tároljuk. (Az `Inf` és `NaN` szimbólumokat nem definiáljuk most.) A fenti szabály szerint négy biten ábrázolható nemnegatív valós számokat az 1.4. táblázat ill. az 1.2. ábra tartalmazza. □

[2]IEEE Binary Floating Point Arithmetic Standard, 754-1985.

1.4. táblázat. Nemnegatív valós számok 4 biten

s	e	m	x
0	0	00	0
0	0	01	$(1.01)_2 \cdot 2^{-1} = (1+\frac{1}{4})\frac{1}{2} = \frac{5}{8}$
0	0	10	$(1.10)_2 \cdot 2^{-1} = (1+\frac{1}{2})\frac{1}{2} = \frac{3}{4} = \frac{6}{8}$
0	0	11	$(1.11)_2 \cdot 2^{-1} = (1++\frac{1}{2}+\frac{1}{4})\frac{1}{2} = \frac{7}{8}$
0	1	00	$(1.00)_2 \cdot 2^{0} = 1 = \frac{8}{8}$
0	1	01	$(1.01)_2 \cdot 2^{0} = 1+\frac{1}{4} = \frac{10}{8}$
0	1	10	$(1.10)_2 \cdot 2^{0} = 1+\frac{1}{2} = \frac{12}{8}$
0	1	11	$(1.11)_2 \cdot 2^{0} = 1+\frac{1}{2}+\frac{1}{4} = \frac{7}{4} = \frac{14}{8}$



1.2. ábra. Nemnegatív gépi számok 4 bites tárolás esetében

Láthatjuk, hogy bármely tárolási módot használjuk, csak véges sok valós számot tudunk a számítógépen tárolni. Azokat a számokat, amelyeket pontosan, azaz tárolási hiba nélkül tudunk tárolni, *gépi számoknak* nevezzük. Azt a gépi számot, amelyet egy adott x valós szám helyett tárolunk a számítógépen, $\mathrm{fl}(x)$-szel jelöljük. Ha $|x|$ kisebb, mint a legkisebb ábrázolható pozitív szám, akkor definíció szerint $\mathrm{fl}(x) = 0$, ha pedig $|x|$ nagyobb, mint a legnagyobb gépi szám, akkor legyen $\mathrm{fl}(x) =$ `Inf`. Az első esetben *alácsordulásról*, a másodikban *túlcsordulásról* beszélünk. Hogy definiálhatjuk $\mathrm{fl}(x)$-et a többi esetben? Két alapvető megközelítés lehetséges. Az egyik esetben vesszük az x szám bináris normál alakját, és annak mantisszájából annyi bitet tárolunk, amennyit az adott tárolási rendszerben tudunk, a többit elhagyjuk. Az egyszeres pontosságú számábrázolás esetében tehát az első 23 törtbitet tároljuk. Ezt a stratégiát *levágásnak* nevezzük. A másik, gyakrabban használt megközelítés *kerekítést* használ. Ebben az esetben $\mathrm{fl}(x)$-et definiáljuk úgy, hogy legyen az x-hez legközelebbi gépi szám. Amikor x két egymás után következő gépi szám számtani közepe, akkor az előbbi definíció még nem határozza meg pontosan $\mathrm{fl}(x)$-et, mert ekkor kerekíthetünk felfelé és lefelé is. A már említett IEEE szabvány ebben az esetben is egyértelműen definiálja a kerekítést. A kerekítési szabályt az egyszeres pontosságú tárolásra fogalmazzuk meg. Vezessük be a következő jelöléseket: legyen az x pozitív valós szám bináris normál alakja $x = m2^k$, ahol $m = 1.m_1m_2\ldots m_{23}m_{24}\ldots$. Legyen $x' = (1.m_1m_2\ldots m_{23})_2 2^k$ és $x'' = \big((1.m_1m_2\ldots m_{23})_2 + 2^{-23}\big)2^k$. Ekkor x' és x'' egymás utáni gépi számok, és $x' \le x \le x''$, valamint $x'' - x' = 2^{k-23}$. A szabvány szerint legyen

$$\mathrm{fl}(x) = \begin{cases} x', & \text{ha } |x-x'| < \frac{1}{2}|x''-x'|, \\ x'', & \text{ha } |x-x''| < \frac{1}{2}|x''-x'|, \\ x', & \text{ha } |x-x'| = \frac{1}{2}|x''-x'| \text{ és } m_{23} = 0, \\ x'', & \text{ha } |x-x'| = \frac{1}{2}|x''-x'| \text{ és } m_{23} = 1. \end{cases}$$

A határesetben, azaz ha $|x-x'| = \frac{1}{2}|x''-x'|$ körülbelül az esetek felerészében fogunk így felfelé, és felerészben lefelé kerekíteni. A másik indoka ennek a definíciónak az, hogy ekkor a határesetben a kerekítéskor a mantissza utolsó bitje mindig 0 lesz, azaz a kerekített számon a 2-vel való osztás hiba nélkül végrehajtható. Kerekítést használva tehát az elkövetett hiba

$$|x - \mathrm{fl}(x)| \le \frac{1}{2}|x''-x'| = \frac{1}{2}2^{-23}2^k.$$

Vizsgáljuk most a kerekítési hibát a pontos értékhez viszonyítva:

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \leq \frac{|x - \mathrm{fl}(x)|}{(1.m_1m_2\ldots)_2 \cdot 2^k} \leq \frac{1}{2}2^{-23}.$$

Könnyen látható, hogy az 1 gépi szám után következő első gépi szám $1 + 2^{-23}$ az előbb vizsgált 32 bites számábrázolási rendszerben. Ezt általánosítva jelölje $\varepsilon_{\text{gépi}}$ az adott számábrázolási rendszerben az első 1-nél nagyobb gépi szám és 1 különbségét. Ezt a számot *gépi epszilonnak* nevezzük. Eszerint $\varepsilon_{\text{gépi}}$ a legkisebb olyan 2 hatvány, amelyre a számítógépen az $1 + \varepsilon_{\text{gépi}} > 1$ egyenlőtlenség ellenőrizhető. Könnyen igazolható a következő tétel, amelyet az előbb a 32 bites bináris alapú tárolási rendszerre beláttunk:

1.9. tétel. *Legyen* $0 < \mathrm{fl}(x) < \texttt{Inf}$*, és tegyük fel, hogy a valós számokat kerekítve tároljuk. Ekkor*

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \leq \frac{1}{2}\varepsilon_{gépi}.$$

A következő tétel bizonyítását az 5. feladatra hagyjuk.

1.10. tétel. *Legyen b a valós számok ábrázolásakor használt számrendszer alapja, és t a mantissza tárolására használt bitek száma. Ekkor*

$$\varepsilon_{gépi} = \begin{cases} 2^{-t}, & ha\ b = 2, \\ b^{1-t}, & ha\ b \neq 2. \end{cases}$$

Most definiáljuk a közelítés hibájának fogalmát, és egyéb, ehhez kapcsolódó fogalmakat. Legyen x egy valós szám, és tekintsük az $\tilde{x}$ valós számot x közelítésének. Ekkor a *közelítés hibáján* az $|x - \tilde{x}|$ számot értjük. Gyakran maga a hiba a számok nagyságrendjének ismerete nélkül nem mond túl sokat. Pl. az 10000 számnak az 10000.1 közelítését elég pontosnak érezzük, az 1-nek viszont az 1.1 nem túl jó közelítése, pedig mindkét esetben a közelítés hibája 0.1. Több információt jelent, ha a pontos értékhez viszonyítjuk a hibát. A *közelítés relatív hibáján* az

$$\frac{|x - \tilde{x}|}{|x|} \qquad (x \neq 0)$$

számot értjük. Azt mondjuk, hogy a b alapú számrendszerben az $\tilde{x}$ *közelítés n számjegyben pontos*, ha

$$\frac{|x - \tilde{x}|}{|x|} \leq \frac{1}{2}b^{1-n}.$$

Látható, hogy minél kisebb a közelítés relatív hibája, annál nagyobb lesz a közelítésben szereplő pontos számjegyek száma. Ha $b = 10$, akkor úgy is megfogalmazhatjuk ezt a kapcsolatot a relatív hiba és a pontos számjegyek száma között, hogy egy nagyságrendi csökkenés (növekedés) a relatív hibában a pontos számjegyek számának egy jeggyel való növekedését (csökkenését) eredményezi.

1.11. példa. Legyen $x = 1657.3$ és $\tilde{x} = 1656.2$. Ekkor a közelítés hibája $|x - \tilde{x}| = 1.1$, relatív hibája $|x - \tilde{x}|/x = 0.0006637$. Mivel $|x - \tilde{x}|/x = 0.0006637 < 0.5 \cdot 10^{-2}$, ezért a közelítés az előbbi definíció értelmében 3 számjegyben pontos. Ha viszont x-et az $\tilde{x} = 1656.9$ számmal közelítjük, akkor ebben az esetben $|x - \tilde{x}|/x = 0.0002413 < 0.5 \cdot 10^{-3}$, azaz definíciónk szerint a közelítés 4 számjegyben pontos. □

Az előbbi definíció és az 1.9. tétel szerint egyszeres pontosságú számábrázolás esetén az x valós szám helyett tárolt $\mathrm{fl}(x)$ gépi szám 24 bináris számjegyben pontos. Minket általában a

tízes számrendszerben felírt alakban érdekel a pontos számjegyek száma. Az egyszeres pontosság esetében ezt megkapjuk, ha az

$$\frac{1}{2}2^{-23} \leq \frac{1}{2}10^{1-n}$$

egyenlőtlenséget teljesítő legnagyobb n egész számot megkeressük. Könnyen kiszámolható, hogy ez $n = 7$, azaz egyszeres számábrázolás esetében a tárolt gépi szám legalább 7 számjegyben pontos a tízes számrendszerben.

1.12. példa. Tekintsük az $x = 12.4$ valós számot. Írjuk fel először ennek bináris alakját. Könnyű ellenőrizni, hogy $12 = (1100)_2$. Keressük tehát a törtrésznek, 0.4-nek a bináris alakját:

$$0.4 = (0.x_1x_2x_3\ldots)_2 = \frac{x_1}{2} + \frac{x_2}{2^2} + \frac{x_3}{2^3} + \cdots .$$

Ha tehát 0.4-nek vesszük a 2-szeresét, akkor annak egész része x_1-et fogja megadni. $0.4 \cdot 2 = 0.8$, azaz $x_1 = 0$. Vesszük a szorzat törtrészét, 0.8-at, és megismételjük az eljárást. $0.8 \cdot 2 = 1.6$, tehát $x_2 = 1$. A szorzat törtrésze 0.6, amivel folytatjuk: $0.6 \cdot 2 = 1.2$, így $x_3 = 1$. Ennek a szorzatnak a törtrésze 0.2. $0.2 \cdot 2 = 0.4$, ezért $x_4 = 0$, és 0.4-gyel folytatjuk az eljárást. Látható, hogy ezután az eddigi jegyek, 0011 ismétlődnek ciklikusan végtelen sokszor, azaz $0.4 = (0.01100110011001100110011001100110011\ldots)_2$. Az x szám bináris normál alakja tehát

$$x = 12.4 = (1.10001100110011001100110011001100110011\ldots)_2 \cdot 2^3.$$

x mantisszáját 23 bitre kerekítve (lefelé) kapjuk, hogy

$$\mathrm{fl}(x) = (1.10001100110011001100110)_2 \cdot 2^3.$$

Ennek értéke a tízes számrendszerben felírva: $\mathrm{fl}(x) = 12.3999996185302734375$. □

A számítógép által elvégzett gépi aritmetikai műveleteket a következőképpen lehet formálisan definiálni:

$$\begin{aligned} x \oplus y &:= \mathrm{fl}(\mathrm{fl}(x) + \mathrm{fl}(y)), \\ x \ominus y &:= \mathrm{fl}(\mathrm{fl}(x) - \mathrm{fl}(y)), \\ x \odot y &:= \mathrm{fl}(\mathrm{fl}(x) \cdot \mathrm{fl}(y)), \\ x \oslash y &:= \mathrm{fl}(\mathrm{fl}(x) / \mathrm{fl}(y)). \end{aligned}$$

Eszerint vesszük az adott műveletben szereplő tényezők gépi számra kerekített értékeit, azon elvégezzük a műveletet, majd a művelet eredményét kerekítjük a legközelebbi gépi számra.

Későbbi példáinkban gyakran fogunk hivatkozni az ún. *négyjegyű aritmetikára*. Ezen azt értjük, hogy olyan számábrázolási rendszert használunk, amely tízes alapú, és 4 mantissza jegyet tárol (és feltesszük, hogy elegendően sok helyünk van a számolás közben fellépő számok kitevőinek tárolásához). Ez azt jelenti, hogy minden egyes részletszámolás eredményét az első nem nulla számjegytől számított 4 jegyre, azaz az első 4 *értékes számjegyre* kerekítjük, és ezt használjuk tovább a számolás során. Négyjegyű aritmetikát használva a kerekítési hibák hatását fel tudjuk erősíteni a vizsgált példákban.

1.13. példa. Négyjegyű aritmetikát használva $1.043 + 32.25 = 33.29$, és hasonlóan $1.043 \cdot 32.25 = 33.64$ (kerekítés után). Viszont $1.043 + 20340 = 20340$ lesz, mivel négy értékes számjegyre kell kerekítenünk a 20341.043 pontos értéket. □

Feladatok

1. Váltsa át bináris alakra a következő tízes számrendszerben felírt számokat:
$$57, \quad -243, \quad 0.25, \quad 35.27$$
2. Írja fel a következő bináris számokat tízes alapú számrendszerben:
$$(101101)_2, \quad (0.10011)_2, \quad (1010.01101)_2$$
3. Mutassa meg, hogy a kettes komplemens kódot negatív szám esetén megkaphatjuk a következőképpen: Vegyük a tárolandó negatív szám abszolút értékének bináris alakját. Cseréljünk minden 0-t 1-re és 1-et 0-ra, majd adjunk 1-et a kapott számhoz.
4. Legyen I_1 és I_2 két m biten tárolt pozitív egész szám. Mutassa meg, hogy az $I_1 - I_2$ különbség kiszámítható úgy, hogy vesszük I_2 kettes komplemens kódját, C_2-t, és ehhez hozzáadjuk I_1-et, majd vesszük az összeg utolsó m bitjét!
5. Bizonyítsa be az 1.10. tételt!
6. Írjon egy olyan programot, amely kiszámítja az adott számítógéphez és a használt számábrázolási rendszerhez tartozó gépi epszilon értékét!
7. Számítsa ki, hogy kétszeres pontosságú számábrázolás esetén hány számjegyben pontos a tárolt gépi szám!
8. Legyen $x = (x_0.x_1x_2\ldots x_mx_{m+1}x_{m+2}\ldots)\cdot 10^k$, $\tilde{x} = (x_0.x_1x_2\ldots x_m\tilde{x}_{m+1}\tilde{x}_{m+2}\ldots)\cdot 10^k$, azaz x és $\tilde{x}$ azonos nagyságrendűek, és az első $m+1$ db számjegyük megegyezik. Lássa be, hogy ekkor $\tilde{x}$ legalább m számjegy pontosságú közelítése x-nek!

1.3. Hibaanalízis

Legyen x és y pozitív, és tekintsük az $\tilde{x}$ és $\tilde{y}$ számokat x és y közelítésének. Legyen $|x-\tilde{x}| \le \Delta_x$ valamint $|y-\tilde{y}| \le \Delta_y$ a közelítések hibakorlátja. A megfelelő relatív hibakorlátokat $\delta_x = \Delta_x/x$ és $\delta_y = \Delta_y/y$ jelölik. Ebben a szakaszban azzal a kérdéssel foglalkozunk, hogy ha az x és y számokon egy aritmetrikai műveletet (összeadás, kivonás, szorzás, osztás) kell elvégeznünk, és ahelyett az $\tilde{x}$ és $\tilde{y}$ számokon végezzük el a műveletet, és annak (pontos) eredményével közelítjük az eredeti művelet eredményét, mekkora lehet a közelítés hibája ill. relatív hibája.

Először tekintsük az összeadás műveletét. Keresünk tehát olyan Δ_{x+y} és δ_{x+y} számokat, hogy
$$|x+y-(\tilde{x}+\tilde{y})| \le \Delta_{x+y} \quad \text{és} \quad \frac{|x+y-(\tilde{x}+\tilde{y})|}{x+y} \le \delta_{x+y}.$$

1.14. tétel. *A*
$$\Delta_{x+y} := \Delta_x + \Delta_y \qquad \text{és} \qquad \delta_{x+y} := \max\{\delta_x, \delta_y\}$$
számok az összeadás hiba- ill. relatív hibakorlátai.

Bizonyítás. A háromszög-egyenlőtlenséget és Δ_x és Δ_y definícióját alkalmazva
$$|x+y-(\tilde{x}+\tilde{y})| \le |x-\tilde{x}| + |y-\tilde{y}| \le \Delta_x + \Delta_y.$$
Ebből kapjuk hogy $\Delta_x + \Delta_y$ egy hibakorlátja lesz az összeadásnak.

Az előbbi összefüggést felhasználva
$$\frac{|x+y-(\tilde{x}+\tilde{y})|}{x+y} \le \frac{\Delta_x+\Delta_y}{x+y} = \frac{x}{x+y}\delta_x + \frac{y}{x+y}\delta_y \le \max\{\delta_x,\delta_y\}.$$
Tehát $\max\{\delta_x, \delta_y\}$ egy relatív hibakorlátja az összeadásnak. □

A tételt nyilván lehet általánosítani több szám összeadására: a tagok hibái összeadódhatnak, az összeg relatív hibája nem nagyobb, mint a tagok relatív hibái közül a legnagyobb. Az állítást megfogalmazhatjuk úgy is, hogy a közelítő összeg pontos jegyeinek száma nem kevesebb, mint az egyes tagok közelítéseiben szereplő pontos jegyek számai közül a legkisebb szám. Természetesen a tétel a legrosszabb esetre vonatkozik. A gyakorlatban a hibák kiegyenlíthetik egymást. Pl. legyen $x = 1$, $y = 2$, $\tilde{x} = 1.1$, $\tilde{y} = 1.8$. Ekkor $x + y = 3$, $\tilde{x} + \tilde{y} = 2.9$. Azaz az összeg hibája csak 0.1, kisebb, mint az egyes tagok hibáinak összege, 0.3.

1.15. tétel. *Legyen $x > y > 0$. A*

$$\Delta_{x-y} := \Delta_x + \Delta_y \qquad \text{és} \qquad \delta_{x-y} := \frac{x}{x-y}\delta_x + \frac{y}{x-y}\delta_y$$

számok a kivonás hiba- ill. relatív hibakorlátai.

Bizonyítás. Az

$$|x - y - (\tilde{x} - \tilde{y})| \le |x - \tilde{x}| + |y - \tilde{y}| \le \Delta_x + \Delta_y$$

egyenlőtlenségekből következik az első állítás. Tekintsük az

$$\frac{|x - y - (\tilde{x} - \tilde{y})|}{x + y} \le \frac{\Delta_x + \Delta_y}{x - y} = \frac{x}{x-y}\delta_x + \frac{y}{x-y}\delta_y,$$

becsléseket, amiből a második állítást kapjuk. □

Látható, hogy ha egymáshoz közeli számokat vonunk ki egymásból, akkor a relatív hiba megsokszorozódhat, azaz a pontos számjegyek száma jelentősen csökkenhet. Ezt a jelenséget hívjuk *értékes számjegyek végzetes elvesztésének.*

1.16. példa. Legyen $x = 12.47531$, $\tilde{x} = 12.47534$, $y = 12.47326$, $\tilde{y} = 12.47325$, akkor $\delta_x = 2.4 \cdot 10^{-6}$ és $\delta_y = 8 \cdot 10^{-7}$. Viszont $x - y = 0.00205$, $\tilde{x} - \tilde{y} = 0.00209$, így $\delta_{x-y} = 0.0195$. Ellenőrizhetjük, hogy $\tilde{x}$ és $\tilde{y}$ 6 pontos számjegyet, $\tilde{x} - \tilde{y}$ viszont csak 2 pontos számjegyet tartalmaz. □

1.17. tétel. *Legyen $x, y > 0$. A*

$$\Delta_{x\cdot y} := x\Delta_y + y\Delta_x + \Delta_x\Delta_y, \qquad \text{és} \qquad \delta_{x\cdot y} := \delta_x + \delta_y + \delta_x\delta_y$$

számok a szorzás hiba- ill. relatív hibakorlátai.

Bizonyítás. A háromszög-egyenlőtlenség szerint

$$\begin{aligned} |xy - \tilde{x}\tilde{y}| &= |xy - x\tilde{y} + x\tilde{y} - \tilde{x}\tilde{y}| \\ &\le x|y - \tilde{y}| + |\tilde{y}||x - \tilde{x}| \\ &\le x\Delta_y + |\tilde{y}|\Delta_x \\ &= x\Delta_y + |y + \tilde{y} - y|\Delta_x \\ &\le x\Delta_y + y\Delta_x + \Delta_x\Delta_y. \end{aligned}$$

Az első állítás szerint a szorzat relatív hibája

$$\frac{|xy - \tilde{x}\tilde{y}|}{xy} \le \frac{x\Delta_y + y\Delta_x + \Delta_x\Delta_y}{xy} = \delta_x + \delta_y + \delta_x\delta_y,$$

amiből kapjuk a második állítást. □

Mivel Δ_x és Δ_y általában sokkal kisebb mint x és y, és így $\Delta_x\Delta_y$ elhanyagolható $x\Delta_y$ és $y\Delta_x$-hez képest, ezért $x\Delta_y + y\Delta_x$ egy jó becslés a szorzat hibájára. Hasonlóan, $\delta_x + \delta_y$ jó közelítése a szorzat relatív hibakorlátjának.

1.18. tétel. *Tegyük fel, hogy $x, y > 0$ és $\delta_y < 1$. Ekkor a*

$$\Delta_{x/y} := \frac{x\Delta_y + y\Delta_x}{y(y-\Delta_y)} \qquad \text{és} \qquad \delta_{x/y} := \frac{\delta_x + \delta_y}{1-\delta_y}$$

számok az osztás hiba- ill. relatív hibakorlátai.

Bizonyítás. Elemi átalakításokat használva kapjuk

$$\left|\frac{x}{y} - \frac{\tilde{x}}{\tilde{y}}\right| = \frac{|x\tilde{y} - xy + xy - \tilde{x}y|}{y|\tilde{y}|} \le \frac{x\Delta_y + y\Delta_x}{y|\tilde{y}|} = \frac{x\Delta_y + y\Delta_x}{y|y-(y-\tilde{y})|}.$$

A $\delta_y < 1$ feltételből következik, hogy $|y-\tilde{y}| \le \Delta_y < y$, ezért az $|y-(y-\tilde{y})| \ge y - |y-\tilde{y}| \ge y - \Delta_y > 0$ egyentőtlenség felhasználásával következik a tétel első állítása.

A második állítás igazolásához tekintsük

$$\frac{\left|\frac{x}{y} - \frac{\tilde{x}}{\tilde{y}}\right|}{\frac{x}{y}} = \frac{|x(\tilde{y}-y) - y(\tilde{x}-x)|}{x|\tilde{y}|} = \frac{\left|\frac{\tilde{y}-y}{y} - \frac{\tilde{x}-x}{x}\right|}{\left|1 - \frac{y-\tilde{y}}{y}\right|} \le \frac{\delta_x + \delta_y}{1-\delta_y}.$$

□

Ha δ_y kicsi, akkor az osztás relatív hibakorlátját jól közelíti $\delta_x + \delta_y$. Hasonlóan, ha Δ_y y-hoz képest elhanyagolható, akkor $\frac{1}{y}\Delta_x + \frac{x}{y^2}\Delta_y$ jó becslése a $\Delta_{x/y}$ hibakorlátnak. Ha y sokkal kisebb, mint x, illetve ha y közel van 0-hoz, akkor Δ_y ill. Δ_x együtthatója nagy, azaz a hiba a tényezők hibáinak többszöröse lehet.

Feladatok

1. Legyen $x = 3.50$, $y = 10.00$, $\tilde{x} = 3.47$, $\tilde{y} = 10.02$. Adjon egy becslést a
$$3x + 7y, \quad \frac{1}{y}, \quad x^2, \quad y^3, \quad \frac{4xy}{x+y}$$
műveletek eredményének hibájára és relatív hibájára (a számítások elvégzése nélkül), ha azokban x és y helyett $\tilde{x}$ és $\tilde{y}$-t használunk! Ezután számítsa ki a tényleges értékeket, hibákat és relatív hibákat, és hasonlítsa össze a kapott becslésekkel!
2. Legyen $\tilde{x}$ az x szám egy közelítése, és $|x-\tilde{x}| \le \Delta_x$. Legyen $f : \mathbb{R} \to \mathbb{R}$ egy differenciálható függvény, amelyre $|f'(x)| \le M$ minden $x \in \mathbb{R}$-re. Legyen $y = f(x)$ és tekintsük az $\tilde{y} = f(\tilde{x})$ számot y közelítésének. Adjon becslést a közelítés hibájára! (Használja a Lagrange-féle középérték tételt!)

1.4. A véges számábrázolás következményei

1.19. példa. Oldjuk meg az
$$x^2 - 83.5x + 1.5 = 0$$
másodfokú egyenletet négyjegyű aritmetikát használva!

A másodfokú egyenlet megoldóképlete szerint és négyjegyű aritmetikát használva a numerikus megoldás
$$\tilde{x} = \frac{83.5 \pm \sqrt{83.5^2 - 4\cdot 1.5}}{2} = \frac{83.5 \pm \sqrt{6972 - 6.000}}{2} = \frac{83.5 \pm 83.46}{2},$$

azaz

$$\tilde{x}_1 = \frac{167.0}{2} = 83.50, \quad \text{és} \quad \tilde{x}_2 = \frac{0.040}{2} = 0.020.$$

Az egyenlet pontos megoldása $x_1 = 83.482032$ ill. $x_2 = 0.0179679$. Ha kiszámoljuk a két gyök közelítésének relatív hibáit, a $\delta_1 = 0.0002152$ és $\delta_2 = 0.113096$ értékeket kapjuk. Az első numerikus gyök tehát 4 számjegy, a második viszont csak 1 számjegy pontosságú közelítés, azaz a két gyök pontossága között 3 nagyságrendi különbség van. Mi ennek az oka? A második gyök kiszámításakor a gyökképletben két egymáshoz közeli számot kellett kivonni egymásból. Az előző szakaszból tudjuk, hogy ez járhat a pontosság elvesztésével, és ezt tapasztalhattuk a jelenlegi számolásban is. □

Tekintsük az $ax^2 + bx + c = 0$ másodfokú egyenlet két gyöke közül az

$$x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \tag{1.2}$$

gyököt. Amikor b negatív, és $4ac$ sokkal kisebb, mint b^2, két egymáshoz közeli számot vonunk ki egymásból a számlálóban, azaz fellép az értékes számjegyek végzetes elvesztésének jelensége. (Ezt az esetet vizsgáltuk az 1.19. példában.) Ennek kiküszöbölésére gyöktelenítsük a számlálót:

$$x_2 = \frac{b^2 - (b^2 - 4ac)}{2a(-b + \sqrt{b^2 - 4ac})} = \frac{2c}{-b + \sqrt{b^2 - 4ac}}. \tag{1.3}$$

Ez a képlet algebrailag ekvivalens az (1.2) formulával. A különbség viszont az, hogy ebben nem szerepel kivonás (a nevezőben két pozitív számot adunk össze). Ha b pozitív, akkor a másik gyökképlettel ismételhetjük meg ugyanezt a trükköt, és kaphatjuk az

$$x_1 = \frac{2c}{-b - \sqrt{b^2 - 4ac}} \tag{1.4}$$

formulát.

1.20. példa. Számítsuk ki az 1.19. példa második gyökét újra, négyjegyű aritmetikát és a gyökképlet (1.4) alakját használva!

$$\tilde{x}_2 = \frac{2 \cdot 1.5}{83.5 + \sqrt{83.5^2 - 4 \cdot 1.5}} = \frac{3}{83.5 + 83.46} = \frac{3}{167.0} = 0.01796.$$

Ennek a numerikus gyöknek a relatív hibája $\delta_2 = 0.00044$, azaz négy számjegy pontosságú a közelítés. □

1.21. példa. Tegyük fel, hogy a $\cos^2 x - \sin^2 x$ kifejezést kell kiértékelnünk. Ha $x = \frac{\pi}{4}$, akkor a kifejezés pontos értéke 0, azaz ha x $\frac{\pi}{4}$-hez közel van, akkor a kifejezés két egymáshoz közeli szám különbsége lesz, ahol fellép a pontosság elvesztése. Ezt könnyen kikerülhetjük, ha az eredeti kifejezés helyett az azzal algebrailag ekvivalens $\cos 2x$ alakot használjuk. □

Az eddigi példáinkban algebrai azonosságot használva tudtuk a pontosság elvesztését megakadályozni. A következő példákban ugyanezt más módszerrel tesszük meg.

1.22. példa. Tekintsük az $f(x) = e^x - 1$ függvényt. Az $x = 0$ közelében ismét két közel azonos számot kell egymásból kivonni, viszont most nincs olyan azonosság, amellyel ezt el lehetne kerülni. Ha e^x Taylor-sorát vesszük, akkor az 1-gyel való kivonással tudunk egyszerűsíteni:

$$f(x) = x + \frac{x}{2} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots.$$

Tehát f-et érdemes ennek a végtelen sornak egy véges közelítő összege segítségével kiértékelni. □

Egy más jellegű problémát vet fel a következő példa.

1.23. példa. Számítsuk ki az $y = 20^{50}/50!$ szám értékét! A probléma a következő: ha a képlet alapján először a számlálót és a nevezőt külön akarjuk kiszámolni, rögtön beleütközünk a számábrázolás szabta korlátokba, egyszeres pontosság használata esetén már túlcsordul a számolás. Másrészt tudjuk, hogy $a^n/n! \to 0$, ha $n \to \infty$, így a számolás végeredménye várhatóan kis szám lesz. Rendezzük úgy a számítást, hogy minden részeredmény benne maradjon az ábrázolható számok tartományában:

$$\frac{20^{50}}{50!} = \frac{20}{50} \cdot \frac{20}{49} \cdot \frac{20}{48} \cdots \frac{20}{1}.$$

Ezt a képletet a számítógépen egy egyszerű **for** ciklussal kiszámolhatjuk:

$y \leftarrow 20$
for $i = 2, \ldots, 50$ **do**
  $y \leftarrow y \cdot \frac{20}{i}$
end do
output(y)

A számolás eredménye: 3.701902 (6 tizedesjegy pontossággal). □

1.24. példa. Számítsuk ki az

$$A = 10.00 + 0.002 + 0.002 + \cdots + 0.002 = 10.00 + \sum_{i=1}^{10} 0.002$$

összeget, négyjegyű aritmetikát használva! Balról jobbra értékeljük ki az összeadásokat, így először az $10.00 + 0.002$ összeget kell kiszámítanunk. Négyjegyű aritmetika szerint $10.00 + 0.002 = 10.002 = 10.00$ kerekítés után. Ehhez hozzáadva a következő számot, a 4 jegyre kerekítés miatt, újra $10.00 + 0.002 + 0.002 = 10.00$ lesz. Látható, hogy $A = 10.00$ lesz a számolás eredménye.

Nézzük most újra az előbbi összeadást, de más sorrendben:

$$B = 0.002 + 0.002 + \cdots + 0.002 + 10.00 = \sum_{i=1}^{10} 0.002 + 10.00.$$

Most először a $0.002 + 0.002 = 0.004$ összeget számítjuk ki. Ezt a négyjegyű aritmetikában is pontosan tudjuk számolni! Ezután sorra számolható: $0.002 + 0.002 + 0.002 = 0.006$ stb, végül $\sum_{i=1}^{10} 0.002 = 0.02$. Az eredmény tehát ebben a sorrendben $B = 10.02$. Ezen összeadások egyikében sem lépett fel kerekítési hiba, mivel minden részeredményt pontosan tárolhattunk a négyjegyű artimetikában.

Ez a példánk mutatja azt is, hogy a kettőnél több tagú összeadás nem kommutatív művelet a számítógépeken. □

Az előző példa tanulsága az, hogy, amikor lehet, az összeadásokat érdemes a tagok növekvő sorrendjében végezni, mert ekkor van a legnagyobb esély arra, hogy a számolás során kapott részösszeg azonos nagyságrendű legyen a következő összeadandóval, és így minél kisebb legyen a kerekítési hiba.

Feladatok

1. Vizsgálja meg, hogy a következő kifejezésekben mikor lép fel az értékes számjegyek elvesztésének jelensége! Hogyan tudjuk kiküszöbölni a pontosság csökkenését?
   (a) $\ln x - 1$,
   (b) $\sqrt{x+9} - 3$,
   (c) $\sin x - x$,

(d) $1 - \cos x$,

(e) $(1 - \cos x)/\sin x$,

(f) $(\cos x - e^{-x})/x$,

2. Négyjegyű aritmetikát használva számítsa ki az $2.274 + 12.04 + 0.4233 + 0.1202 + 0.2204$ összeget, majd rendezze a tagokat növekvő sorrendbe, és úgy is számítsa ki az összeget!

2. fejezet

Nemlineáris egyenletek, egyenletrendszerek

Ebben a fejezetben nemlineáris egyenletek és egyenletrendszerek numerikus megoldásának legismertebb módszereit tárgyaljuk (intervallumfelezés módszere, húr-, szelő-, Newton-, kvázi-Newton módszerek stb.). Megismerkedünk az iterációs sorozatok általános elméletével, a fixpont, a konvergencia sebessége fogalmával, iterációs eljárások megállási feltételeivel. Definiáljuk a vektor- és mátrixnorma fogalmát, és ennek segítségével a vektor- és mátrixsorozatok konvergenciáját.

2.1. Analízis előismeretek

Ebben a szakaszban összefoglaljuk azokat az analízisből ismert fogalmakat, tételeket, amelyekre a későbbiekben gyakran fogunk hivatkozni.

Az $[a,b]$ intervallumon értelmezett valós értékű folytonos függvények halmazát $C[a,b]$-vel jelöljük. Azon $f : [a,b] \to \mathbb{R}$ függvények halmazát, amelyek $[a,b]$-n folytonosak és (a,b)-n m-szer folytonosan differenciálhatók, $C^m[a,b]$-vel jelöljük.

2.1. tétel. *Legyen $f \in C[a,b]$. Ekkor f felveszi maximumát és minimumát $[a,b]$-n, azaz létezik olyan $c,d \in [a,b]$, hogy*

$$f(c) = \max_{x\in[a,b]} f(x) \qquad \text{és} \qquad f(d) = \min_{x\in[a,b]} f(x).$$

Az a és b számok által generált nyílt intervallumot $\langle a,b\rangle$-vel jelöljük, azaz

$$\langle a,b\rangle := (\min\{a,b\}, \max\{a,b\}),$$

ill. ennek általánosításaként $\langle a_1,a_2,\ldots,a_n\rangle$ jelöli az a_1, a_2, …, a_n számok által generált nyílt intervallumot, azaz

$$\langle a_1,a_2,\ldots,a_n\rangle := (\min\{a_1,a_2,\ldots,a_n\}, \max\{a_1,a_2,\ldots,a_n\}).$$

A következő eredmény szerint egy folytonos függvény bármely két értéke közti minden értéket felvesz.

2.2. tétel. *Legyen $f \in C[a,b]$, $f(a) \neq f(b)$, és legyen $d \in \langle f(a), f(b)\rangle$. Ekkor létezik olyan $c \in (a,b)$, hogy $f(c) = d$.*

2.3. tétel (Rolle). *Legyen az $f : [a,b] \to \mathbb{R}$ folytonos függvény differenciálható az (a,b) intervallumon, és $f(a) = f(b)$. Ekkor létezik olyan $\xi \in (a,b)$ szám, hogy $f'(\xi) = 0$.*

2.4. tétel (Lagrange-féle középértéktétel). *Legyen $f : [a,b] \to \mathbb{R}$ folytonos az $[a,b]$ intervallumon és differenciálható az (a,b) intervallumon. Ekkor létezik olyan $\xi \in (a,b)$ szám, hogy $f(b) - f(a) = f'(\xi)(b-a)$.*

2.5. tétel (Taylor-tétel). *Legyen $f \in C^{n+1}[a,b]$, és legyen $x_0 \in (a,b)$. Ekkor minden $x \in (a,b)$-hez létezik olyan $\xi = \xi(x) \in \langle x, x_0 \rangle$, hogy*

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \\ &\quad + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}. \end{aligned}$$

A következő tételt integrálokra vonatkozó középértéktételnek is nevezik.

2.6. tétel. *Legyen $f\colon [a,b] \to \mathbb{R}$ folytonos függvény, $g\colon [a,b] \to \mathbb{R}$ integrálható függvény amely nem vált előjelet $[a,b]$-n (azaz $g(x) \geq 0$ vagy $g(x) \leq 0$ teljesül minden $x \in [a,b]$-re). Ekkor létezik egy olyan $\xi \in (a,b)$ szám, hogy*

$$\int_a^b f(x)g(x)\,dx = f(\xi)\int_a^b g(x)\,dx.$$

A következő eredményt úgy szokás röviden megfogalmazni, hogy egymásba skatulyázott zárt intervallumoknak létezik egy közös pontja, ha az intervallumok hossza nullához tart.

2.7. tétel. *Legyen $[a_n, b_n]$ $(n = 1, 2, \ldots)$ korlátos zárt intervallumoknak egy sorozata, amelyre $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$ teljesül minden n-re, és $(b_n - a_n) \to 0$ ha $n \to \infty$. Ekkor létezik olyan $c \in [a_1, b_1]$ szám, hogy $a_n \to c$ és $b_n \to c$, ha $n \to \infty$.*

2.8. tétel. *Monoton és korlátos számsorozatnak létezik határértéke.*

Zárjuk ezt a szakaszt az algebra alaptételének nevezett eredmény felidézésével, amelyet a következő alakban fogalmazunk meg:

2.9. tétel (Az algebra alaptétele). *Egy*

$$p(x) = a_n x^n + \cdots + a_1 x + a_0, \qquad a_j \in \mathbb{C}\ (j = 0, \ldots, n), \quad a_n \neq 0$$

polinomnak pontosan n db komplex gyöke van multiplicitásokkal számolva.

A tételnek általában arra a következményére lesz szükségünk, hogy ha egy $p(x) = a_n x^n + \cdots + a_1 x + a_0$ polinomnak van $n+1$ db gyöke, akkor az a $p \equiv 0$ (azonosan nulla) polinom.

2.2. Fixpont iteráció

A numerikus módszerek jelentős része egy végtelen sorozatot generál, amelynek határértéke adja a vizsgált probléma pontos megoldását. A numerikus analízisben szereplő sorozatokat gyakran *rekurzív definícióval*, más néven *iterációval* adjuk meg. Egy $p_{k+1} = h(p_k, p_{k-1}, \ldots, p_{k-m+1})$ $(k \geq m-1)$ rekurzív definícióval megadott iterációs módszert *m-lépéses iterációnak* nevezünk. Egy m-lépéses iterációs sorozatot m kezdeti érték, p_0, p_1, …, p_{m-1} határoz meg egyértelműen.

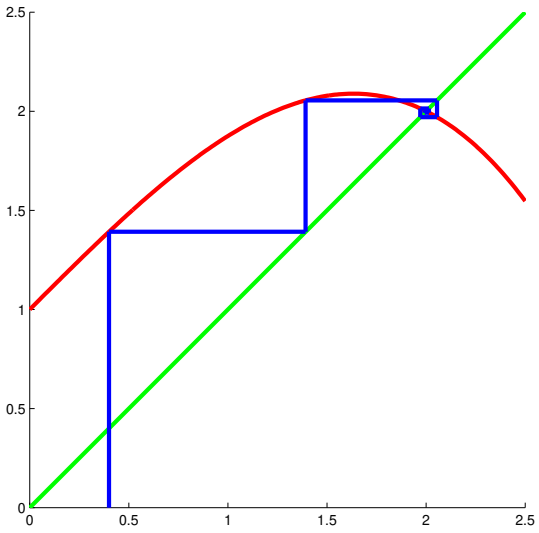
Ebben a szakaszban a leggyakoribb esettel, az egylépéses iterációval, más néven *fixpont iterációval* foglalkozunk részletesebben.

Adott egy $g\colon I \to \mathbb{R}$ függvény, ahol $I \subset \mathbb{R}$. A $p_{k+1} = g(p_k)$ képlettel definiált (és valamely $p_0 \in I$ kezdeti értékhez tartozó) sorozatot *fixpont iterációs sorozatnak*, vagy röviden *fixpont iterációnak* nevezünk.

2.10. példa. Tekintsük a $g(x) = -\frac{1}{8}x^3 + x + 1$ függvényt! A 2.1. táblázatban kiszámítottuk a fixpont iterációval generált sorozat első néhány tagját a $p_0 = 0.4$ kezdőértékből kiindulva. A sorozat tagjait a 2.1. ábrán látható ún. *lépcsős diagrammal* szokás ábrázolni. A kiindulási $(p_0, 0)$ pontból rajzolunk egy függőleges egyenest a g függvény grafikonjáig. A kimetszett pont y-koordinátája adja a sorozat p_1 elemét. A (p_0, p_1) pontból egy vízszintes szakaszt rajzolunk az $y = x$ egyenes (p_1, p_1) pontjáig. A sorozat $p_2 = g(p_1)$ elemét tehát úgy kapjuk geometriailag, ha ebből a pontból egy függőleges szakasz mentén elmegyünk a g grafikonjáig, és a kimetszett pont y-koordinátája lesz p_2. Ezt az eljárást folytatva kapjuk az ábrán látható töröttvonalat. A töröttvonal ennél a példánál spirálisan ráhúzódik az $y = x$ egyenes és az $y = g(x)$ görbe metszéspontjára. A metszéspont koordinátái $(2, 2)$. A 2.1. táblázatból látható, hogy a p_k sorozat a 2 értékhez konvergál. □

2.1. táblázat. Fixpont iteráció, $g(x) = -\frac{1}{8}x^3 + x + 1$

k	p_k
0	0.40000000
1	1.39200000
2	2.05484646
3	1.97030004
4	2.01419169
5	1.99275275
6	2.00358428
7	1.99819822
8	2.00089846
9	1.99955017
10	2.00022477
11	1.99988758
12	2.00005620
13	1.99997190
14	2.00001405
15	1.99999297



2.1. ábra. Fixpont iteráció

Az előbbi példában azt tapasztaltuk, hogy a fixpont sorozat az $y = x$ egyenes és az $y = g(x)$

grafikon metszéspontjának x-koordinátájához konvergál. Ennek a pontnak az x-koordinátája (és persze az y-koordinátája is) teljesíti a $g(x) = x$ egyenletet. A p számot a g függvény *fixpontjának* nevezzük, ha

$$g(p) = p.$$

Eszerint a terminológia szerint az előbbi példában a fixpont sorozat a g függvény egy fixpontjához konvergált. A következő tételben belátjuk, hogy ez minden konvergens fixpont sorozatra jellemző.

2.11. tétel. *Legyen $g\colon [a,b] \to [a,b]$ (vagy $\mathbb{R} \to \mathbb{R}$) folytonos függvény, $p_0 \in [a,b]$ rögzített, és tekintsük a $p_{k+1} = g(p_k)$ fixpont iterációs sorozatot. Ha p_k konvergens és $p_k \to p$, akkor $p = g(p)$.*

Bizonyítás. Mivel $p_{k+1} = g(p_k)$ és a feltételek szerint $p_{k+1} \to p$ és $g(p_k) \to g(p)$, ha $k \to \infty$, így az állítás következik. □

Egy fixpont sorozat természetesen nem feltétlenül konvergens, ill. a határérték lehet végtelen. Ehhez elég a $g(x) = 2x$ függvényt és a $p_0 = 1$ kezdőértéket tekinteni. Ekkor $p_k = 2^k$, ami a végtelenbe tart. Ha pedig a $g(x) = -x$ függvényt vesszük, akkor a fixpont iteráció a $p_k = (-1)^k$ sorozatot generálja.

A következő tétel elégséges feltételt ad a fixpont létezésére és egyértelműségére.

2.12. tétel. *Legyen $g\colon [a,b] \to [a,b]$ folytonos. Ekkor g-nek létezik legalább egy fixpontja az $[a,b]$ intervallumon. Ha ezenkívül feltesszük azt is, hogy g differenciálható (a,b)-n, és létezik olyan $0 \le c < 1$ szám, hogy $|g'(x)| \le c$ minden $x \in (a,b)$-re, akkor a fixpont egyértelmű.*

Bizonyítás. Tekintsük az $f(x) = g(x) - x$ függvényt. Ha $f(a) = 0$ vagy $f(b) = 0$, akkor a ill. b a g függvény fixpontja. Ellenkező esetben $f(a) > 0$ és $f(b) < 0$. De ekkor f folytonossága miatt létezik olyan $p \in (a,b)$ szám, hogy $f(p) = 0$, azaz $p = g(p)$.

A tétel második felének bizonyításához tegyük fel, hogy g-nek két fixpontja is van, p és q. Ekkor használva a Lagrange-féle középértéktételt, létezik olyan $\xi \in (a,b)$ szám, hogy

$$|p - q| = |g(p) - g(q)| = |g'(\xi)||p - q| \le c|p - q|,$$

amiből következik, hogy $p = q$, azaz a fixpont egyértelmű. □

2.13. tétel (fixpont tétel). *Legyen $g\colon [a,b] \to [a,b]$ folytonos függvény, g differenciálható (a,b)-n, és tegyük fel hogy létezik olyan $0 \le c < 1$ szám, hogy $|g'(x)| \le c$ minden $x \in (a,b)$-re. Legyen $p_0 \in [a,b]$ tetszőleges, és $p_{k+1} = g(p_k)$ $(k \ge 0)$. Ekkor a p_k sorozat konvergál a g függvény (egyértelmű) p fixpontjához,*

$$|p_k - p| \le c^k |p_0 - p|, \tag{2.1}$$

valamint

$$|p_k - p| \le \frac{c^k}{1-c} |p_1 - p_0|. \tag{2.2}$$

Bizonyítás. A 2.12. tétel szerint g-nek létezik egyértelmű fixpontja, p. Mivel $0 \le c < 1$ a feltételek szerint, ezért ha belátjuk (2.1)-et, abból $p_k \to p$ következik. A feltételek és a Lagrange-féle középértéktétel szerint

$$|p_k - p| = |g(p_{k-1}) - g(p)| = |g'(\xi)||p_{k-1} - p| \le c|p_{k-1} - p|.$$

Ebből (teljes indukcióval) könnyen látható a (2.1) egyenlőtlenség.

(2.2) igazolásához legyen $m > k$ tetszőleges. A háromszög-egyenlőtlenséget, középértéktételt és a feltételeket alkalmazva

$$\begin{aligned}
|p_k - p_m| &\le |p_k - p_{k+1}| + |p_{k+1} - p_{k+2}| + \cdots + |p_{m-1} - p_m| \\
&\le |g(p_{k-1}) - g(p_k)| + |g(p_k) - g(p_{k+1})| + \cdots + |g(p_{m-2}) - g(p_{m-1})| \\
&\le c|p_{k-1} - p_k| + c|p_k - p_{k+1}| + \cdots + c|p_{m-2} - p_{m-1}| \\
&\le (c^k + c^{k+1} + \cdots + c^{m-1})|p_0 - p_1| \\
&= c^k(1 + c + \cdots + c^{m-k-1})|p_1 - p_0| \\
&\le c^k \sum_{i=0}^{\infty} c^i |p_1 - p_0|.
\end{aligned}$$

Így $|p_k - p_m| \le \frac{c^k}{1-c}|p_1 - p_0|$ minden $m > k$-ra. Ha k rögzített és m tart a végtelenbe, kapjuk a (2.2) egyenlőtlenséget. □

Vegyük észre, hogy az előbbi két tétel bizonyításában g differenciálhatóságát és a derivált korlátosságát csak arra használtuk, hogy a

$$|g(x) - g(y)| \le c|x - y| \tag{2.3}$$

becslést kapjuk g-re. Azt mondjuk, hogy a g függvény *Lipschitz-tulajdonságú* az I intervallumon, ha létezik olyan $c \ge 0$ konstans, hogy (2.3) teljesül minden $x, y \in I$-re. Az egyenlőtlenségben szereplő c számot a g függvény *Lipschitz-konstansának* nevezzük. A Lagrange-féle középértéktétel szerint ha $g \in C^1[a,b]$, akkor g Lipschitz-tulajdonságú $[a,b]$-n a $c := \max\{|g'(x)| : x \in [a,b]\}$ Lipschitz-konstanssal. g Lipschitz-tulajdonságú akkor is, ha csak szakaszonként folytonosan differenciálható. Példa erre a $g(x) = |x|$ függvény. (Lásd még a 8. feladatot.) Ha g Lipschitz-tulajdonságú egy $0 \le c < 1$ Lipschitz-konstanssal, akkor g-t *kontrakciónak* nevezzük. A 2.13. tételt kimondhatjuk tehát a következő, általánosabb alakban is:

2.14. tétel (kontrakciós elv, fixpont tétel). *Legyen $g : [a,b] \to [a,b]$ folytonos függvény kontrakció, $p_0 \in [a,b]$ tetszőleges, és $p_{k+1} = g(p_k)$ $(k \ge 0)$. Ekkor a p_k sorozat konvergál a g függvény (egyértelmű) p fixpontjához, és teljesülnek a (2.1) és (2.2) becslések.*

Gyakran találkozunk olyan numerikus iterációs módszerekkel, amelyek konvergálnak, feltéve, hogy a sorozat kezdeti értékei közel vannak a feladat pontos megoldásához, azaz a sorozat határértékéhez. Azt mondjuk, hogy egy $p_{k+1} = h(p_k, p_{k-1}, \ldots, p_{k-m+1})$ iterációs módszer *lokálisan konvergál* p-hez, ha létezik olyan $\delta > 0$, hogy minden $p_0, p_1, \ldots, p_{m-1} \in (p-\delta, p+\delta)$ kezdeti értékekhez tartozó p_k sorozat p-hez konvergál. Ha a p_k sorozat tetszőleges kezdeti értékre konvergál p-hez, akkor az iterációs módszert *globálisan konvergensnek* nevezzük.

2.15. tétel. *Legyen $g \in C^1[a,b]$, és legyen $p \in (a,b)$ a g függvény egy fixpontja. Tegyük fel, hogy $|g'(p)| < 1$. Ekkor a fixpont iteráció lokálisan konvergál p-hez, azaz létezik olyan $\delta > 0$, hogy a $p_{k+1} = g(p_k)$ sorozat minden $p_0 \in (p-\delta, p+\delta)$-ra konvergál p-hez.*

Bizonyítás. Mivel a feltételek szerint g' folytonos és $|g'(p)| < 1$, ezért létezik olyan $\delta > 0$, hogy $[p-\delta, p+\delta] \subset (a,b)$ és $|g'(x)| < 1$ minden $x \in [p-\delta, p+\delta]$-re. Legyen $c = \max\{|g'(x)| : x \in [p-\delta, p+\delta]\}$. Ekkor $0 \le c < 1$.

Belátjuk, hogy g a $[p-\delta, p+\delta]$ intervallumot önmagába képezi. Legyen $p_0 \in [p-\delta, p+\delta]$. A Lagrange-féle középértéktételt és c definícióját használva

$$|g(p_0) - p| = |g(p_0) - g(p)| \le c|p_0 - p| < |p_0 - p| < \delta,$$

azaz $g(p_0)$ a $[p-\delta, p+\delta]$ intervallumba esik. Ezért a 2.13. tétel alkalmazható a g függvény $[p-\delta, p+\delta]$ intervallumra vett megszorítására, amiből következik az állítás. □

Feladatok

1. Legyen $g(x) = mx$, ahol $m \in \mathbb{R}$. Ábrázolja a g-hez (és valamely kezdőértékhez) tartozó fixpont iterációs sorozatokat $m = 0.5, 1, 1.5, -0.5, -1, -1.5$-re!
2. Alakítsa át a következő egyenleteket fixpont egyenletté, majd fixpont iteráció segítségével adja meg az egyenletek olyan közelítő megoldását, amely 4 jegyre pontos:

   (a) $(x-2)^3 = x+1$, (b) $\frac{\cos x}{x} = 2$,

   (c) $x^3 + x - 1 = 0$, (d) $2x \sin x = 4 - 3x$.

3. Tekintsük az $x^3 + x^2 + 3x - 5 = 0$ egyenletet. Mutassa meg, hogy az egyenlet bal oldalát leíró polinom monoton növekvő, és 0 és 2 között metszi a grafikonja az x-tengelyt! (Természetesen könnyű észrevenni, hogy az egyenlet pontos gyöke $x = 1$.) Ellenőrizze, hogy az egyenlet ekvivalens a következő fixpont feladatokkal:

   (a) $x = x^3 + x^2 + 4x - 5$, (b) $x = \sqrt[3]{5 - x^2 - 3x}$,

   (c) $x = \frac{5}{x^2+x+3}$, (d) $x = \frac{5-x^3}{x+3}$,

   (e) $x = \frac{2x^3+x^2+5}{3x^2+2x+3}$, (f) $x = \frac{5+7x-x^2-x^3}{10}$.

   Számítsa ki a fixpont iteráció első néhány tagját az összes egyenletre a $p_0 = 0.5$ kezdőértéket használva, és állapítsa meg, hogy melyik esetben kapunk konvergens sorozatot! Hasonlítsa össze a konvergencia/divergencia gyorsasága szempontjából a sorozatokat! Indokolja a tapasztaltakat!
4. Lássa be, hogy a $p_k = \frac{1}{2}p_{k-1} + \frac{1}{p_{k-1}}$ sorozat $\sqrt{2}$-höz konvergál, ha $p_0 > \sqrt{2}$! Mi történik, ha $0 < p_0 < \sqrt{2}$? És mit tapasztal, ha $p_0 < 0$?
5. Mutassa meg, hogy a $p_k = \frac{1}{2}p_{k-1} + \frac{A}{2p_{k-1}}$ sorozat $\sqrt{A}$-hoz konvergál, ha $p_0 > 0$! Mi történik $p_0 < 0$-ra?
6. Legyen $g \in C^1[a,b]$, és legyen $p \in (a,b)$ egy fixpontja g-nek, és $|g'(p)| > 1$. Mutassa meg, hogy a fixpont iteráció nem konvergál p-hez, ha $p_0 \neq p$!
7. Tekintsük a $g(x) = \sqrt{1+x^2}$ függvényt! Mutassa meg, hogy $|g'(x)| < 1$ minden $x \in \mathbb{R}$-ra, de a fixpont sorozat nem konvergál egyetlen kezdeti értékre sem!
8. Legyen $f\colon [a,b] \to \mathbb{R}$ folytonos, és legyenek $a = x_0 < x_1 < \cdots < x_n = b$ olyan osztópontok, hogy f lineáris minden $[x_i, x_{i+1}]$ intervallumon ($i = 0, \ldots, n-1$). Lássa be, hogy f Lipschitz-tulajdonságú!

2.3. Intervallumfelezés módszere

Ebben és a következő néhány szakaszban az $f(x) = 0$ nemlineáris egyenlet numerikus megoldását keressük. Erre a legegyszerűbb algoritmus az ún. *intervallumfelezés módszere*. Ezt ismertetjük ebben a szakaszban.

Tegyük fel, hogy $f\colon [a,b] \to \mathbb{R}$ folytonos függvény, amely ellentétes előjelű az intervallum végpontjaiban, azaz $f(a)f(b) < 0$. Ekkor tudjuk, hogy f-nek legalább egy gyöke van az $[a,b]$ intervallumon. Definiáljuk itervallumoknak egy sorozatát: Legyen $[a_0, b_0] = [a, b]$, és legyen p_0 az intervallum felezőpontja, azaz $p_0 = (a_0 + b_0)/2$. Ekkor vagy $f(p_0) = 0$, vagy az $[a_0, p_0]$ és $[p_0, b_0]$ intervallumok közül az egyiknek a végpontjaiban ellentétes előjelű az f függvény. Ha az $[a_0, p_0]$ intervallumon vált előjelet, akkor $[a_1, b_1] = [a_0, p_0]$, egyébként $[a_1, b_1] = [p_0, b_0]$ a következő intervallum definíciója. Ezt az eljárást folytatva vagy véges sok lépés után az egyik p_k felezőpont gyöke lesz az f függvénynek, vagy pedig zárt intervallumoknak egy egymásba skatulyázott sorozatát kapjuk, amelyek mindegyike tartalmazza az f függvény egy gyökét. Mivel a k-adik

intervallum hossza $(b-a)/2^k$, ezért az intervallumoknak pontosan egy p közös pontjuk van, ami az f függvény gyöke. Az intervallumok bármely pontja, így speciálisan pl. a felezőpontok sorozata, p_k, tart p-hez. Tegyük fel a meghatározottság kedvéért, hogy $f(a) < 0$ és $f(b) > 0$ (a másik eset hasonlóan kezelhető). Ekkor minden k-ra $f(a_k) < 0$ és $f(b_k) > 0$ az iteráció során. Mivel $a_k \to p$ és $b_k \to p$, ezért az f folytonossága miatt $f(p) \le 0$ és $f(p) \ge 0$ kell legyen, azaz $f(p) = 0$. Mivel $a_k \le p \le b_k$ minden k-ra, ezért $|p_k - p| \le (b_k - a_k)/2 = (b-a)/2^{k+1}$. Ezzel beláttuk a következő tételt:

2.16. tétel. *Legyen $f \in C[a,b]$ és $f(a)f(b) < 0$. Ekkor az intervallumfelezés módszerével kapott p_k sorozat konvergál az f függvény egy p gyökéhez, és*

$$|p_k - p| \le \frac{b-a}{2^{k+1}}. \tag{2.4}$$

A (2.4) becslésből következik, hogy ha egy előre megadott $\varepsilon > 0$ hibakorlátot (tolerancia értéket) szeretnénk elérni a közelítéssel, akkor olyan p_k tagot kell használni p közelítésére, amelynek indexe

$$k \ge \log_2 \frac{b-a}{\varepsilon} - 1. \tag{2.5}$$

2.17. példa. Tekintsük az $f(x) = e^x - 2\cos x$ függvényt. $f(0) = -1$ és $f(1) > 0$, tehát f-nek van gyöke a $[0,1]$ intervallumon. (Könnyű belátni, hogy f szigorúan monoton növő, így pontosan egy gyöke van.) A 2.2. táblázat tartalmazza az intervallumfelezéses módszer numerikus eredményét. Az $\varepsilon = 10^{-5}$ tolerancia eléréséhez a (2.5) formula szerint $k \ge \log_2 10^5 - 1 \approx 15.61$ lépés elegendő. □

2.2. táblázat. Intervallumfelezés módszere, $f(x) = e^x - 2\cos x$, $[0,10]$, $TOL = 10^{-5}$

k	a_k	b_k	p_k	$f(p_k)$
0	0.00000000	1.00000000	0.50000000	-1.0644e-01
1	0.50000000	1.00000000	0.75000000	6.5362e-01
2	0.50000000	0.75000000	0.62500000	2.4632e-01
3	0.50000000	0.62500000	0.56250000	6.3206e-02
4	0.50000000	0.56250000	0.53125000	-2.3292e-02
5	0.53125000	0.56250000	0.54687500	1.9538e-02
6	0.53125000	0.54687500	0.53906250	-1.9818e-03
7	0.53906250	0.54687500	0.54296875	8.7517e-03
8	0.53906250	0.54296875	0.54101563	3.3784e-03
9	0.53906250	0.54101563	0.54003906	6.9670e-04
10	0.53906250	0.54003906	0.53955078	-6.4294e-04
11	0.53955078	0.54003906	0.53979492	2.6780e-05
12	0.53955078	0.53979492	0.53967285	-3.0810e-04
13	0.53967285	0.53979492	0.53973389	-1.4067e-04
14	0.53973389	0.53979492	0.53976440	-5.6946e-05
15	0.53976440	0.53979492	0.53977966	-1.5083e-05
16	0.53977966	0.53979492	0.53978729	5.8483e-06

Feladatok

1. Lássa be, hogy az

   (a) $x^3 - 6x - 1 = 0$, $[a,b] = [-1,1]$, (b) $x = e^{-2x}$, $[a,b] = [-1,2]$,

   (c) $\tan x = x + 1$, $[a,b] = [-1,1.5]$, (d) $e^{-\sin x} = x^2 - 1$, $[a,b] = [0,2]$

   egyenleteknek létezik gyöke az $[a,b]$ intervallumon! Az intervallumfelezés módszerével, az $\varepsilon = 10^{-5}$ tolerancia értéket használva adja meg a gyök közelítését!

2. Alkalmazza az intervallumfelezés módszerét az $f(x) = \frac{1}{x}$ függvényre a $[-0.5, 3]$ kezdeti intervallumot használva! Mit tapasztal?

2.4. Húrmódszer

Az intervallumfelezéses módszer előnye, hogy előre lehet tudni, hogy egy adott pontosságú közelítés eléréséhez hány lépésre van szükség. A módszer hátránya viszont az, hogy nem veszi figyelembe a függvény alakját az intervallumok képzésekor. Ezt a hiányosságot próbálja kiküszöbölni a *húrmódszer*.

A kiindulás ugyanaz, mint az előző módszernél. Feltesszük, hogy $f : [a, b] \to \mathbb{R}$ folytonos függvény, amely ellentétes előjelű az intervallum végpontjaiban, azaz $f(a)f(b) < 0$. Ennél a módszernél is $[a_k, b_k]$ intervallumoknak és azokat osztó p_k pontoknak egy sorozatát képezzük. Kiindulásul legyen $[a_0, b_0] = [a, b]$. Az k-adik lépésben p_k-t az f függvény a_k és b_k pontjaihoz tartozó húrja (azaz az $(a_k, f(a_k))$ és $(b_k, f(b_k))$ pontokat összekötő szakasz) és az x-tengely metszeteként definiáljuk. Kis számolással kapjuk, hogy

$$p_k = a_k - f(a_k)\frac{a_k - b_k}{f(a_k) - f(b_k)}. \tag{2.6}$$

Ezután a következő lépés $[a_{k+1}, b_{k+1}]$ intervallumának az $[a_k, p_k]$ és $[p_k, b_k]$ intervallumok közül azt választjuk, ahol a függvény szintén előjelet vált. A módszert a 2.18. algoritmussal írjuk le pontosabban.

2.18. algoritmus. Húrmódszer

INPUT: f - függvény,
$[a, b]$ intervallum, ahol $f(a)f(b) < 0$
TOL - tolerancia,
$MAXIT$ - maximális iterációszám,
OUTPUT: p - közelítő gyök.

$i \leftarrow 1$ *(lépésszám)*
$q \leftarrow a$
while $i < MAXIT$ **do**
  $p \leftarrow a - f(a)(a - b)/(f(a) - f(b))$
  if $|p - q| < TOL$ **do**
    output(p)
    stop
  end do
  if $f(p)f(b) < 0$ **do**
    $a \leftarrow p$
  else if $f(a)f(p) < 0$ **do**
    $b \leftarrow p$
  else
    output(p)
    stop
  end do
  $i \leftarrow i + 1$

$q \leftarrow p$
end do
output(Maximális iterációszám túllépve)

Az előbbi algoritmus programozásakor természetesen p definiálása előtt célszerű megvizsgálni, hogy $f(a)$ egyenlő-e $f(b)$-vel, nehogy nullával való osztás miatt futási hibával álljon le a program. Ha $f(a) = f(b)$, akkor a program adjon egy figyelmeztető üzenetet, hogy nem alkalmazható a módszer, és szakítsuk meg szabályosan a program futását. Az ilyen jellegű ellenőrzéseket az egyszerűség kedvéért nem építettük be ebbe és a későbbi algoritmusokba sem, de természetesen ezekről gondoskodnia kell a programozónak az algoritmus számítógépen történő implementációjánál.

A húrmódszer konvergenciáját csak arra a speciális esetre bizonyítjuk be, amikor f konvex vagy konkáv.

2.19. tétel. *Tegyük fel, hogy az $f \in C[a,b]$ függvény konvex vagy konkáv $[a,b]$-n és $f(a)f(b) < 0$. Ekkor a húrmódszer konvergál az f függvény (egyértelmű) p gyökéhez.*

Bizonyítás. Tegyük fel, hogy f konvex és $f(a) > 0$, $f(b) < 0$. A többi eset hasonlóan igazolható. Ekkor minden lépésben a bal oldali részintervallum fogja tartalmazni f gyökét, azaz $a_{k+1} = a$ és $b_{k+1} = p_k$ minden k-ra. Mivel a p_k sorozat monoton csökkenő és az a szám egy alsó korlátja, ezért konvergál egy $p \geq a$ számhoz. $f(p_k) < 0$ minden k-ra, ezért $f(p) \leq 0$. Mivel $f(a) > 0$, ezért $p > a$. A (2.6) egyenletből kapjuk a $k \to \infty$ határértéket véve, hogy

$$p = a - f(a)\frac{a-p}{f(a)-f(p)},$$

amiből $f(p) = 0$ következik. □

2.20. példa. A húrmódszert alkalmazva a 2.17. példa feladatára a 2.3. táblázatban felsorolt értékeket kapjuk. A 2.17. példához hasonlóan most is a $[0,10]$ kezdeti intervallumot és $TOL = 10^{-5}$ értéket használtunk. Látható, hogy ezen a feladaton a húrmódszer sokkal gyorsabban konvergál mint az intervallumfelezés módszere. □

2.3. táblázat. Húrmódszer, $f(x) = e^x - 2\cos x$, $[0,10]$, $TOL = 10^{-5}$

k	a_k	b_k	p_k	$f(p_k)$
0	0.00000000	1.00000000	0.37912145	-3.9698e-01
1	0.37912145	1.00000000	0.50026042	-1.0576e-01
2	0.50026042	1.00000000	0.53057677	-2.5118e-02
3	0.53057677	1.00000000	0.53766789	-5.8011e-03
4	0.53766789	1.00000000	0.53929982	-1.3311e-03
5	0.53929982	1.00000000	0.53967399	-3.0499e-04
6	0.53967399	1.00000000	0.53975970	-6.9856e-05
7	0.53975970	1.00000000	0.53977933	-1.5999e-05
8	0.53977933	1.00000000	0.53978383	-3.6640e-06

2.21. példa. Alkalmazzuk újra a húrmódszert a 2.17. példa feladatára, de most a $[0,4]$ intervallumból kiindulva! A 2.4. táblázatban látható az eredmény. (Csak az első és utolsó néhány tagot listáztuk.) Most az előző példához képest sokkal lassabb a konvergencia. (Ez még tovább lassul, ha az intervallum bal oldali végpontját tovább csökkentjük.) Ha viszont az intervallumfelezés módszerét indítjuk a $[0,4]$ kezdeti intervallummal, akkor a lépésszám csak kettővel nő, mivel $\log_2 4/10^{-5} - 1 \approx 17.61$. □

2.4. táblázat. Húrmódszer, $f(x) = e^x - 2\cos x$, $[0, 4]$, $TOL = 10^{-5}$

k	a_k	b_k	p_k	$f(p_k)$
0	0.00000000	4.00000000	0.07029205	-9.2224e-01
1	0.07029205	4.00000000	0.13406612	-8.3858e-01
2	0.13406612	4.00000000	0.19119837	-7.5285e-01
3	0.19119837	4.00000000	0.24180834	-6.6826e-01
4	0.24180834	4.00000000	0.28620106	-5.8729e-01
⋮	⋮	⋮	⋮	⋮
47	0.53966897	4.00000000	0.53968870	-2.6464e-04
48	0.53968870	4.00000000	0.53970508	-2.1970e-04
49	0.53970508	4.00000000	0.53971868	-1.8240e-04
50	0.53971868	4.00000000	0.53972996	-1.5143e-04
51	0.53972996	4.00000000	0.53973934	-1.2572e-04

Feladatok

1. Alkalmazza a húrmódszert a 2.3. szakasz 1. feladatában felsorolt egyenletekre!
2. Legyen
$$f(x) = \begin{cases} \delta, & x \le 0.5 \\ 4(1+\delta)(x - x^2) - 1, & x \ge 0.5 \end{cases}$$
Alkalmazza az intervallumfelezés módszerét és a húrmódszert a $[0, 1]$ intervallumon az f függvény gyökének meghatározására, ha
(a) $\delta = 2$, (b) $\delta = 0.5$, (c) $\delta = 0.09$.
3. Dolgozza ki a 2.19. tétel bizonyítását a többi esetre is!

2.5. Newton-módszer

A numerikus analízisben gyakran használjuk a következő „módszert”: helyettesítsük a problémát egy „egyszerűbb” problémával, ami „közel van” az eredeti problémához, és tekintsük az „egyszerűbb” probléma megoldását az eredeti közelítésének. Az $f(x) = 0$ nemlineáris egyenlet megoldásakor tekintsük az f függvény egy közelítését: Rögzítsünk egy p_0 pontot, vegyük f elsőrendű Taylor-polinomját p_0 körül, és keressük meg annak a gyökét. Geometriailag ez azt jelenti, hogy vesszük az f függvény p_0 pontjához húzott érintő metszéspontját az x-tengellyel. A metszéspontot az $f(p_0) + f'(p_0)(x - p_0) = 0$ lineáris egyenlet megoldása adja, $x = p_0 - f(p_0)/f'(p_0)$ (feltéve, hogy $f'(p_0) \neq 0$). Ezt a számot jelöljük p_1-gyel, és ismételjük meg az eljárást. Így kapjuk a

$$p_{k+1} = p_k - \frac{f(p_k)}{f'(p_k)} \tag{2.7}$$

rekurzív képlettel definiált sorozatot. A (2.7) iterációt *Newton–Raphson módszernek* vagy röviden *Newton-módszernek* ill. *érintőmódszernek* nevezzük.

2.22. példa. A Newton-módszert alkalmazva a 2.17. példa feladatára a 2.5. táblázatban felsorolt értékeket kapjuk. A sorozat nagyon gyorsan megközelítette a függvény gyökét.

□

A Newton-módszer egy egylépéses iterációs módszer, azaz fixpont iteráció a

$$g(x) := x - \frac{f(x)}{f'(x)} \tag{2.8}$$

2.5. táblázat. Newton-módszer, $f(x) = e^x - 2\cos x$, $p_0 = 0$, $TOL = 10^{-5}$

k	p_k	$f(p_k)$
0	0.1000000000	-8.8484e-01
1	0.7781206411	7.5291e-01
2	0.5678850726	7.8450e-02
3	0.5402639121	1.3139e-03
4	0.5397853041	3.9302e-07
5	0.5397851608	3.5207e-14

iterációs függvénnyel. g-t differenciálva kapjuk

$$g'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}. \tag{2.9}$$

Legyen p az f függvény olyan gyöke, amelyre $f'(p) \neq 0$. Ekkor $g'(p) = 0$, így a 2.15. tételből rögtön következik:

2.23. tétel. *Legyen $f \in C^2[a,b]$, és legyen $p \in (a,b)$ olyan, hogy $f(p) = 0$ és $f'(p) \neq 0$. Ekkor a Newton-módszer lokálisan konvergál p-hez.*

2.24. példa. Tekintsük az $f(x) = 0.5 \operatorname{arctg} x$ függvényt. Ennek egyetlen gyöke $p = 0$. $f'(0) = 0.5$, így a Newton-módszer lokálisan konvergál $p = 0$-hoz, azaz, ha p_0 elég kicsi, akkor a Newton-sorozat 0-hoz tart. A 2.6. táblázatban a $p_0 = 1.4$ kezdeti értékhez tartozó sorozat első néhány tagját nyomtattuk ki. (A 15. lépésben a program hibaüzenettel leállt, mert $f'(p_{14}) = 0$ a számítógépen.) Látható, hogy p_k ebben az esetben nem tart 0-hoz.

□

2.6. táblázat. Newton-módszer, $f(x) = 0.5 \operatorname{arctg} x$, $p_0 = 1.4$

k	p_k	$f(p_k)$
0	1.4000000e+00	0.4752734
1	-1.4136186e+00	-0.4775591
2	1.4501293e+00	0.4835443
3	-1.5506260e+00	-0.4990071
4	1.8470541e+00	0.5372889
5	-2.8935624e+00	-0.6190257
6	8.7103258e+00	0.7282453
7	-1.0324977e+02	-0.7805557
8	1.6540564e+04	0.7853679
9	-4.2972148e+08	-0.7853982
10	2.9006412e+17	0.7853982
11	-1.3216239e+35	-0.7853982
12	2.7436939e+70	0.7853982
13	-1.1824729e+141	-0.7853982
14	2.1963537e+282	0.7853982

Feladatok

1. Alkalmazza a Newton-módszert a 2.3. szakasz 1. feladatában felsorolt egyenletek megoldására!
2. Legyen $f(x) = 0.5 \operatorname{arctg} x$. f-nek nyilván $x = 0$ az egyetlen gyöke. Legyen a p_k a Newton-iterációval generált sorozat. Mutassa meg, hogy létezik olyan p^* szám, hogy
   (a) ha $|p_0| < p^*$, akkor $p_k \to 0$,
   (b) ha $|p_0| = p^*$, akkor a p_k sorozat váltakozva a p_0, $-p_0$ értékeket veszi fel (azaz nem konvergens),

(c) ha $|p_0| > p^*$, akkor p_k váltakozó előjelű, és $|p_k| \to \infty$.

3. Vezessen le egy iterációs módszert $\sqrt[k]{a}$ kiszámítására!

2.6. Szelőmódszer

A Newton-módszer képletében szerepel az f függvény deriváltja. A gyakorlatban viszont f' sokszor nem ismert (pl. f nem egy képlettel van megadva, hanem egy numerikus eljárás generálja a függvény értékét egy megadott pontban), vagy a derivált képletének kiértékelése túl sok gépi számolást igényel, így „nem éri meg" a használata. A derivált használatát küszöböli ki a *szelőmódszer*. Legyen p_0 és p_1 két egymástól különböző, általunk választott kezdeti érték. Tekintsük az f függvény grafikonjának p_0 és p_1 pontjaihoz tartozó szelőt, azaz a $(p_0, f(p_0))$ és $(p_1, f(p_1))$ pontokon átmenő egyenest. Ennek egyenlete:

$$y = f(p_1) + \frac{f(p_1) - f(p_0)}{p_1 - p_0}(x - p_1).$$

A szelő az x-tengelyt az $x = p_1 - \frac{p_1 - p_0}{f(p_1) - f(p_0)} f(p_1)$ pontban metszi. Ezt a pontot p_2-vel jelöljuk. Ezután tekintsük a p_1 és p_2 pontokhoz tartozó szelőt, és annak az x-tengellyel vett metszéspontját jelöljük p_3-mal. Ezt az eljárást folytatva kapjuk a

$$p_{k+1} = p_k - \frac{p_k - p_{k-1}}{f(p_k) - f(p_{k-1})} f(p_k) \tag{2.10}$$

sorozatot. A (2.10) képlettel definiált kétlépéses iterációs módszert *szelőmódszernek* nevezzük.

2.25. példa. A szelőmódszert alkalmazva a 2.17. példa feladatára a 2.7. táblázatban felsorolt értékeket kapjuk. Összehasonlítva a 2.5. táblázatban látható eredménnyel, látható, hogy a szelőmódszer valamivel lassabban konvergál, mint a Newton-módszer. □

2.7. táblázat. Szelőmódszer, $f(x) = e^x - 2\cos x$, $p_0 = 0$, $p_1 = 10$, $TOL = 10^{-5}$

k	p_k	$f(p_k)$
0	0.0000000000	-1.0000e+00
1	1.0000000000	1.6377e+00
2	0.3791214458	-3.9698e-01
3	0.5002604213	-1.0576e-01
4	0.5442561500	1.2301e-02
5	0.5396724494	-3.0921e-04
6	0.5397848464	-8.6246e-07
7	0.5397851608	6.0793e-11

A szelőmódszer konvergenciájának igazolásához szükségünk lesz a következő eredményre.

2.26. tétel. *Legyen $f \in C^2[a, b]$, és legyen $p \in (a, b)$ olyan, hogy $f(p) = 0$ és $f'(p) \neq 0$. Legyen p_k a szelőmódszerrel generált sorozat. Ekkor minden k-ra léteznek olyan $\xi_k \in \langle p_k, p_{k-1}, p\rangle$ és $\eta_k \in \langle p_k, p_{k-1}\rangle$ számok, hogy*

$$p_{k+1} - p = \frac{1}{2}\frac{f''(\xi_k)}{f'(\eta_k)}(p_k - p)(p_{k-1} - p). \tag{2.11}$$

Bizonyítás. Kis számolással belátható

$$\begin{aligned} p_{k+1}-p &= p_k - p - \frac{p_k - p_{k-1}}{f(p_k)-f(p_{k-1})} f(p_k) \\ &= \frac{(p_{k-1}-p)f(p_k) - (p_k-p)f(p_{k-1})}{f(p_k)-f(p_{k-1})} \\ &= \frac{(p_k-p)(p_{k-1}-p)}{f(p_k)-f(p_{k-1})}\left(\frac{f(p_k)}{p_k-p} - \frac{f(p_{k-1})}{p_{k-1}-p}\right) \\ &= (p_k-p)(p_{k-1}-p)\frac{p_k-p_{k-1}}{f(p_k)-f(p_{k-1})} \frac{\frac{f(p_k)-f(p)}{p_k-p} - \frac{f(p_{k-1})-f(p)}{p_{k-1}-p}}{p_k-p_{k-1}} \end{aligned}$$

A Lagrange-féle középérték tétel szerint létezik olyan $\eta_k \in \langle p_k, p_{k-1}\rangle$ szám, hogy

$$\frac{f(p_k)-f(p_{k-1})}{p_k-p_{k-1}} = f'(\eta_k).$$

A tétel bizonyításának befejezéséhez azt kell megmutatnunk, hogy létezik olyan $\xi_k \in \langle p_k, p_{k-1}, p\rangle$, hogy

$$\frac{\frac{f(p_k)-f(p)}{p_k-p} - \frac{f(p_{k-1})-f(p)}{p_{k-1}-p}}{p_k-p_{k-1}} = \frac{f''(\xi_k)}{2}. \tag{2.12}$$

Ennek direkt bizonyítását a 2. feladatra hagyjuk. Itt most a 6. fejezetben bevezetendő fogalmakra és eredményekre hivatkozva látjuk be a (2.12) relációt. Eszerint (2.12) bal oldala nem más, mint az f függvény p_{k-1}, p és p_k pontokra felírt másodrendű osztott differenciája, $f[p_{k-1}, p, p_k]$ (lásd a 6.2. szakaszt). A 6.17. következmény szerint létezik olyan $\xi_k \in \langle p_k, p_{k-1}, p\rangle$ szám, hogy $f[p_{k-1}, p, p_k] = f''(\xi_k)/2$. □

2.27. tétel. *Legyen $f \in C^2[a,b]$, és legyen $p \in (a,b)$ olyan, hogy $f(p) = 0$ és $f'(p) \neq 0$. Ekkor a szelőmódszer lokálisan konvergál p-hez.*

Bizonyítás. Legyen δ^* olyan, hogy $f'(x) \neq 0$ ha $x \in [p-\delta^*, p+\delta^*]$. Ilyen δ^* létezik, mivel $f'(p) \neq 0$ és f' folytonos. Legyen

$$M := \frac{\max\{|f''(x)|\colon x \in [p-\delta^*, p+\delta^*]\}}{2\min\{|f'(x)|\colon x \in [p-\delta^*, p+\delta^*]\}}.$$

Válasszuk δ-t úgy, hogy $\delta < \min\{\delta^*, \frac{1}{M}\}$ legyen, és legyen $\varepsilon := M\delta$. Ekkor a feltételek szerint $0 < \varepsilon < 1$. Legyen $p_0, p_1 \in (p-\delta, p+\delta)$ tetszőleges, de különböző számok. (2.11) és M definíciója szerint $|p_{k+1}-p| \le M|p_k-p||p_{k-1}-p|$, és ezért

$$M|p_{k+1}-p| \le M|p_k-p|M|p_{k-1}-p| \tag{2.13}$$

minden k-ra. Ezt $k = 1$-re alkalmazva $M|p_2-p| \le M|p_1-p|M|p_0-p| \le (M\delta)^2 = \varepsilon^2 < \varepsilon$. Ebből kapjuk, hogy $|p_2-p| \le \varepsilon/M = \delta$. Ez azt jelenti, hogy $p_2 \in (p-\delta, p+\delta)$. Hasonlóan belátható, hogy $p_k \in (p-\delta, p+\delta)$ minden k-ra.

ε definíciójából következik, hogy $M|p_0-p| < \varepsilon$ és $M|p_1-p| < \varepsilon$. Most keresünk egy olyan q_k sorozatot, amelyre $M|p_k-p| \le \varepsilon^{q_k}$ teljesül minden k-ra. Az előbbiek szerint használhatjuk a $q_0 = 1$ és $q_1 = 1$ értékeket. Tegyük fel, hogy már definiáltuk a q_k sorozat első k tagját. A (2.13) egyenlőtlenség szerint ekkor az $M|p_{k+1}-p| \le \varepsilon^{q_k}\varepsilon^{q_{k-1}}$ egyenlőtlenség kell, hogy teljesüljön. Ezért a $M|p_{k+1}-p| \le \varepsilon^{q_{k+1}}$ becslés teljesülni fog, ha q_{k+1}-et úgy választjuk, hogy

$$q_{k+1} = q_k + q_{k-1}, \quad k \ge 1, \quad q_0 = 1, \quad q_1 = 1 \tag{2.14}$$

legyen. A (2.14) rekurzív képlettel definiált sorozatot *Fibonacci-sorozatnak* nevezzük. Belátható (3. feladat), hogy q_k általános képlete

$$q_k = \frac{1}{\sqrt{5}}(r_0^{k+1} - r_1^{k+1}), \qquad k \geq 0, \tag{2.15}$$

ahol

$$r_0 = \frac{1+\sqrt{5}}{2} \approx 1.618, \quad \text{és} \quad r_1 = \frac{1-\sqrt{5}}{2} \approx -0.618.$$

Ebből következik, hogy $q_k \to \infty$ ha $k \to \infty$. Ebből viszont kapjuk hogy $p_k \to p$, hiszen

$$|p_k - p| \leq \frac{1}{M}\varepsilon^{q_k} \to 0, \qquad \text{ha } k \to \infty.$$

□

Feladatok

1. Alkalmazza a szelőmódszert a 2.3. szakasz 1. feladatában felsorolt egyenletekre!
2. Lássa be a (2.12) relációt! (Útmutatás: igazolja, hogy a
$$f[a,b,c] = \frac{\frac{f(c)-f(b)}{c-b} - \frac{f(b)-f(a)}{b-a}}{c-a}$$
kifejezés értéke független az a, b, c számok sorrendjétől! Ezért feltehetjük, hogy $a < b < c$. Vegye az f függvény b-körüli elsőrendű Taylor-közelítését a másodrendű hibataggal együtt! Ennek segítségével fejezze ki a jobb oldalon álló kifejezés számlálóját! Végül használja a 2.2. tételt annak igazolására, hogy $f[a,b,c] = f''(\xi)/2$ valamely $\xi \in (a,c)$-re!)
3. Igazolja a (2.15) képletet!

2.7. Konvergencia rendje

Az eddig vizsgált iterációs módszerekel alkalmazva tapasztaltuk, hogy a különböző módszerek eltérő sebességgel konvergálnak. A konvergencia gyorsaságának jellemzésére bevezetjük a konvergencia rendjének fogalmát.

Legyen p_k egy konvergens sorozat, melynek határértéke p. A p_k sorozat *konvergencia rendje* α, ha $\alpha \geq 1$ és létezik olyan $c \geq 0$ szám, hogy

$$|p_{k+1} - p| \leq c|p_k - p|^\alpha \qquad \text{minden } k \geq 0\text{-ra}, \tag{2.16}$$

és $\alpha = 1$ esetén még azt is kikötjük, hogy $c < 1$ legyen.

Ha pontosabban akarunk fogalmazni, akkor a (2.16) egyenlőtlenséget teljesítő p_k sorozatra azt mondhatjuk, hogy a konvergencia rendje *legalább* α, hiszen elképzelhető, hogy a (2.16) egyenlőtlenséget α-nál nagyobb kitevővel is teljesíti. Mi a „legalább" jelzőt elhagyjuk, de a konvergencia rend fogalmát ebben az értelemben használjuk. Ha azt szeretnénk hangsúlyozni, hogy a p_k sorozat a (2.16) egyenlőtlenséget teljesíti valamely α-ra, de azt nem teljesíti egy α-nál nagyobb kitevőre sem, akkor azt mondjuk, hogy a konvergencia rendje *pontosan* α.

Ha egy p_k sorozat konvergencia rendje $\alpha = 1$, akkor *lineáris*, ha $\alpha = 2$, akkor *kvadratikus* konvergenciáról beszélünk.

Ha egy p_k sorozat lineárisan konvergál p-hez, akkor könnyen látható, hogy teljesíti a

$$|p_k - p| \leq c^k |p_0 - p| \tag{2.17}$$

becslést. Néhány módszer esetében nem könnyű a (2.16) típusú egyenlőtlenséget belátni az $\alpha = 1$ esetben, viszont könnyebb a (2.17) egyenlőtlenséget igazolni. Ezért a lineáris konvergencia előbbi általános definícióját kibővítjük úgy, hogy ha egy p_k sorozat teljesíti a (2.17) egyenlőtlenséget egy $0 \le c < 1$ konstanssal, akkor is lineáris konvergenciáról beszélünk.

Tegyük fel, hogy $p_k \to p$, és a konvergencia rendje α. A

$$\lim_{k\to\infty} \frac{p_{k+1} - p}{(p_k - p)^\alpha} \tag{2.18}$$

véges határértéket, ha létezik, a p_k sorozat *aszimptotikus hibakonstansának* nevezzük. Könnyen látható, hogy ha a (2.18) határérték létezik és véges, akkor a p_k sorozat konvergencia rendje α. Ha egy p_k sorozat konvergencia rendje $\alpha = 1$ és az aszimptotikus hibakonstansa 0, akkor *szuperlineáris* konvergenciáról beszélünk.

2.28. tétel. *Tegyük fel, hogy a p_k sorozat α rendben konvergál p-hez a $\lambda \neq 0$ aszimptotikus hibakonstanssal. Ekkor*

1. $\lim\limits_{k\to\infty} \dfrac{p_{k+1} - p}{(p_k - p)^\beta} = 0$ *minden $\beta < \alpha$-ra, és*

2. $\lim\limits_{k\to\infty} \dfrac{|p_{k+1} - p|}{|p_k - p|^\beta} = \infty$ *minden $\beta > \alpha$-ra.*

Bizonyítás. Az állítás következik a

$$\frac{|p_{k+1} - p|}{|p_k - p|^\beta} = \frac{|p_{k+1} - p|}{|p_k - p|^\alpha} \frac{1}{|p_k - p|^{\beta-\alpha}}$$

összefüggésből. □

A tételből következik, hogy ha egy p_k sorozat (legalább) α rendben konvergál, és az aszimptotikus hibakonstans létezik és nem 0, akkor a konvergencia rendje pontosan α.

2.29. példa. Tekintsük újra a 2.22. példában vizsgált Newton-iterációt! A 2.8. táblázat utolsó három oszlopában feltüntettük a $|p_{k+1} - p|/|p_k - p|^\alpha$ kifejezések értékeit $\alpha = 1, 2$ és 3-ra, használva a $p = 0.5397851608092811$ értéket. Látható, hogy $\alpha = 1$-re a kifejezés 0-hoz tart, $\alpha = 2$-re korlátos marad, de nem tart 0-hoz, $\alpha = 3$-ra pedig a végtelenbe tart. (Természetesen egy sorozat első 4 tagjából még nem tudunk messzemenő következtetéseket levonni, de ha több tagját generáljuk a sorozatnak, ellenőrizhetjük az előbb említett eredményt.) Úgy tapasztaljuk tehát, hogy a sorozat konvergencia rendje 2. □

2.8. táblázat. Newton-módszer konvergencia rendje, $f(x) = e^x - 2\cos x$

			$\lvert p_k - p\rvert / \lvert p_{k-1} - p\rvert^\alpha$		
k	p_k	$f(p_k)$	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$
0	0.0000000000	-1.0000e+00			
1	1.0000000000	1.6377e+00	8.5259e-01	1.5795e+00	2.9262e+00
2	0.6279041258	2.5516e-01	1.9147e-01	4.1605e-01	9.0404e-01
3	0.5442066314	1.2164e-02	5.0176e-02	5.6941e-01	6.4619e+00
4	0.5397973257	3.3375e-05	2.7513e-03	6.2226e-01	1.4074e+02
5	0.5397851609	2.5388e-10	7.6071e-06	6.2533e-01	5.1404e+04

2.30. tétel. *Tegyük fel, hogy a p_k sorozat teljesíti a (2.16) egyenlőtlenséget valamely $c \ge 0$-ra és $\alpha > 1$-re. Ekkor a p_n sorozat lokálisan konvergál a p számhoz, valamint minden k-ra*

$$|p_k - p| \le c^{\frac{\alpha^k - 1}{\alpha - 1}} |p_0 - p|^{\alpha^k}. \tag{2.19}$$

Bizonyítás. Teljes indukcióval könnyen igazolható a (2.19) egyenlőtlenség. Ebből viszont következik, hogy

$$|p_k - p| \le c^{\frac{1}{1-\alpha}} \left(c^{\frac{1}{\alpha-1}} |p_0 - p| \right)^{\alpha^k},$$

így ha p_0 olyan, hogy $c^{\frac{1}{\alpha-1}}|p_0 - p| < 1$, akkor $p_k \to p$, azaz p_k lokálisan konvergál p-hez. □

2.31. példa. Legyenek $p_k \to p$ és $q_k \to q$ lineárisan ill. kvadratikusan konvergáló sorozatok, amelyek teljesítik a (2.17) ill. (2.16) egyenlőtlenségeket $c = 1/2$-re. Továbbá tegyük fel, hogy $|p_0 - p| < 1$ és $|q_0 - q| < 1$. Ekkor a (2.17) és (2.19) egyenlőtlenségekből kapjuk, hogy $|p_k - p| \le (1/2)^k$ ill. $|q_k - q| \le (1/2)^{2^k-1}$. A 2.9. táblázatban ezeket a hibakorlátokat soroltuk fel $k = 1, 2, \ldots, 5$-re. Látható, hogy a hiba mennyivel gyorsabban csökken (azaz a konvergencia mennyivel gyorsabb) a kvadratikus esetben. □

2.9. táblázat.

k	$(1/2)^k$	$(1/2)^{2^k-1}$
1	$5.0000 \cdot 10^{-1}$	$5.0000 \cdot 10^{-1}$
2	$2.5000 \cdot 10^{-1}$	$1.2500 \cdot 10^{-1}$
3	$1.2500 \cdot 10^{-1}$	$7.8125 \cdot 10^{-3}$
4	$6.2500 \cdot 10^{-2}$	$3.0518 \cdot 10^{-5}$
5	$3.1250 \cdot 10^{-2}$	$4.6566 \cdot 10^{-10}$
6	$1.5625 \cdot 10^{-2}$	$1.0842 \cdot 10^{-19}$

2.32. tétel. *Legyen $g \in C^m[a,b]$, $p \in (a,b)$ és $p = g(p)$. Tekintsük a $p_{k+1} = g(p_k)$ fixpont interációt.*

1. *Ha $|g'(p)| < 1$, akkor a fixpont iteráció lokálisan lineárisan konvergál p-hez.*
2. *Ha $g'(p) = g''(p) = \ldots = g^{(m-1)}(p) = 0$, akkor a fixpont iteráció lokálisan m-edrendben konvergál p-hez a $g^{(m)}(p)/m!$ aszimptotikus hibakonstanssal.*

Bizonyítás. Az 1. állítás a 2.15. tétel bizonyításából következik.

A 2. állítás bizonyításához vegyük a g függvény p-körüli $(m-1)$-edrendű Taylor-közelítését:

$$g(p_k) = g(p) + g'(p)(p_k - p) + \cdots + \frac{g^{(m-1)}(p)}{(m-1)!}(p_k - p)^{m-1} + \frac{g^{(m)}(\xi_k)}{m!}(p_k - p)^m,$$

ahol $\xi_k \in \langle p_k, p \rangle$. Ebből következik, használva, hogy az első $m-1$ derivált 0 a p pontban, $g(p) = p$ és $g(p_k) = p_{k+1}$, hogy

$$|p_{k+1} - p| = \frac{|g^{(m)}(\xi_k)|}{m!}|p_k - p|^m \le c|p_k - p|^m. \tag{2.20}$$

Az utolsó becslésnél használtuk, hogy $g \in C^m[a,b]$, azaz $g^{(m)}$ folytonos, így korlátos p egy környezetében. A (2.18) határérték létezése következik az előbbiekből, hiszen $\xi_k \to p$ ha $k \to \infty$, mivel $|\xi_k - p| \le |p_k - p|$, és ezért

$$\lim_{k\to\infty} \frac{p_{k+1} - p}{(p_k - p)^m} = \lim_{k\to\infty} \frac{g^{(m)}(\xi_k)}{m!} = \frac{g^{(m)}(p)}{m!}.$$

□

A tételből következik, hogy a fixpont iteráció konvergenciájának rendje mindig egész szám (feltéve hogy a g függvény elegendően sokszor differenciálható). A 2.36. tételben meg fogjuk mutatni, hogy ez általában nem igaz többlépéses iterációs módszerekre.

Szükségünk lesz a többszörös gyök fogalmára. A $p \in (a,b)$ számot az $f \in C[a,b]$ függvény *m-szeres gyökének* nevezzük, ha létezik olyan $q \in C[a,b]$ függvény, hogy $q(p) \neq 0$ és

$$f(x) = (x-p)^m q(x), \qquad x \in (a,b). \tag{2.21}$$

Könnyen igazolható a következő állítás:

2.33. tétel. *Legyen $f \in C^m[a,b]$, $p \in (a,b)$.*

1. *Legyen p m-szeres gyöke f-nek, és a (2.21) azonosságot teljesítő q függvény m-szer differenciálható. Ekkor*
$$f(p) = f'(p) = f''(p) = \ldots = f^{(m-1)}(p) = 0, \quad \text{és } f^{(m)}(p) \neq 0. \tag{2.22}$$
2. *Ha (2.22) teljesül, akkor p m-szeres gyöke f-nek.*
3. *Tegyük fel, hogy az f függvény akárhányszor differenciálható, f-et előállítja a p-körüli Taylor-sora, és f teljesíti a (2.22) relációkat. Ekkor p m-szeres gyöke f-nek, és a (2.21) azonosságot teljesítő q függvény is akárhányszor differenciálható, valamint q is Taylor-sorba fejthető p-körül.*

A következő tétel szerint ha f-nek p egyszeres gyöke, akkor a Newton-módszer kvadratikusan, ha pedig többszörös gyöke, akkor lineárisan konvergál.

2.34. tétel. *Legyen $f \in C^2[a,b]$.*

1. *Ha $f(p) = 0$ és $f'(p) \neq 0$, akkor a Newton-iteráció lokálisan kvadratikusan konvergál p-hez.*
2. *Ha $f(x) = (x-p)^m q(x)$, ahol $q \in C^2[a,b]$, $q(p) \neq 0$, $m > 1$, akkor a Newton-iteráció lokálisan lineárisan konvergál p-hez.*

Bizonyítás. Az 1. állítás következik a 2.32. tétel 2. állításából, hiszen a Newton-iteráció egy fixpont iteráció a (2.8) egyenlettel definiált g iterációs függvénnyel, és $g'(p) = 0$ a (2.9) reláció szerint.

Mivel a

$$g(x) := \begin{cases} x - \frac{f(x)}{f'(x)}, & x \neq p, \\ p & x = p \end{cases}$$

függvényre

$$g(x) = x - \frac{(x-p)q(x)}{mq(x) + (x-p)q'(x)},$$

ezért g folytonosan differenciálható p-ben, és $g'(p) = 1 - \frac{1}{m}$. Így a 2.32. tétel 2. pontja szerint a konvergencia rendje lineáris. □

2.35. példa. Keressük meg az $f(x) = x^3 + x^2 - 8x - 12$ polinom egy gyökét a Newton–Raphson módszerrel, a $p_0 = 0$ kiindulási értéket és a 10^{-5} tolerancia értéket használva! Könnyen látható, hogy $x = -2$ kétszeres gyöke, $x = 3$ pedig egyszeres gyöke a polinomnak. A 2.10. táblázatban található futásnál a $p_0 = 0$, a 2.11. táblázat generálásakor pedig a $p_0 = 2$ kezdeti értékből indultunk ki. Az első esetben a sorozat -2-höz konvergál, a második esetben pedig 3-hoz. A táblázatokból látható, hogy az első esetben csak lineáris, a másodikban pedig kvadratikus a konvergencia rendje. □

2.10. táblázat. Newton-módszer, $f(x) = x^3 + x^2 - 8x - 12$

k	p_k	$f(p_k)$	$\frac{\lvert p_k - p\rvert}{\lvert p_{k-1} - p\rvert^\alpha}$ $\alpha = 1$	$\alpha = 2$
0	0.0000000000	-1.2000e+01		
1	-1.5000000000	-1.1250e+00	2.5000e-01	1.2500e-01
2	-1.7647058824	-2.6379e-01	4.7059e-01	9.4118e-01
3	-1.8853313477	-6.4237e-02	4.8734e-01	2.0712e+00
4	-1.9433465411	-1.5866e-02	4.9406e-01	4.3086e+00
5	-1.9718365260	-3.9436e-03	4.9712e-01	8.7747e+00
6	-1.9859582600	-9.8308e-04	4.9858e-01	1.7703e+01
7	-1.9929890302	-2.4542e-04	4.9929e-01	3.5558e+01
8	-1.9964969780	-6.1313e-05	4.9965e-01	7.1267e+01
9	-1.9982491032	-1.5323e-05	4.9982e-01	1.4268e+02
10	-1.9991247050	-3.8300e-06	4.9991e-01	2.8552e+02
11	-1.9995623908	-9.5743e-07	4.9996e-01	5.7119e+02
12	-1.9997812050	-2.3935e-07	4.9998e-01	1.1425e+03
13	-1.9998906049	-5.9835e-08	4.9999e-01	2.2852e+03
14	-1.9999453030	-1.4959e-08	4.9999e-01	4.5705e+03
15	-1.9999726517	-3.7396e-09	5.0000e-01	9.1412e+03
16	-1.9999863259	-9.3491e-10	5.0000e-01	1.8283e+04
17	-1.9999931629	-2.3373e-10	5.0000e-01	3.6565e+04

2.11. táblázat. Newton-módszer, $f(x) = x^3 + x^2 - 8x - 12$

k	p_k	$f(p_k)$	$\frac{\lvert p_k - p\rvert}{\lvert p_{k-1} - p\rvert^\alpha}$ $\alpha = 1$	$\alpha = 2$
0	2.0000000000	-1.6000e+01		
1	4.0000000000	3.6000e+01	1.0000e+00	1.0000e+00
2	3.2500000000	6.8906e+00	2.5000e-01	2.5000e-01
3	3.0217391304	5.4821e-01	8.6957e-02	3.4783e-01
4	3.0001866020	4.6654e-03	8.5837e-03	3.9485e-01
5	3.0000000139	3.4816e-07	7.4632e-05	3.9996e-01
6	3.0000000000	1.9400e-15	5.5721e-09	4.0011e-01

2.36. tétel. *Ha f-nek p egyszeres gyöke, akkor a szelőmódszer $\alpha = (1+\sqrt{5})/2 \approx 1.618$ rendben lokálisan konvergál p-hez.*

Bizonyítás. Használjuk a 2.27. tétel bizonyításában bevezetett jelöléseket és az ott kapott eredményeket. A (2.13) egyenlőtlenség szerint

$$|p_{k+1} - p| \le M|p_k - p||p_{k-1} - p|.$$

Ebből kiindulva, és a $|p_k - p| \le \frac{1}{M}\varepsilon^{q_k}$ becslést használva kapjuk

$$\begin{aligned} |p - p_{k+1}| &\le |p_k - p|^{r_0} M |p_k - p|^{1-r_0} |p_{k-1} - p| \\ &\le |p_k - p|^{r_0} M \left(\frac{1}{M}\varepsilon^{q_k}\right)^{1-r_0} \frac{1}{M}\varepsilon^{q_{k-1}} \\ &= |p_k - p|^{r_0} M^{r_0-1} \varepsilon^{q_k + q_{k-1} - r_0 q_k} \\ &= |p_k - p|^{r_0} M^{r_0-1} \varepsilon^{q_{k+1} - r_0 q_k} \\ &= |p_k - p|^{r_0} M^{r_0-1} \varepsilon^{r_1^{k+1}}. \end{aligned}$$

Megjegyezzük, hogy az utolsó lépés a (2.15) egyenlőségből következik (kis számolással). Mivel $r_1^{k+1} \to 0$ ha $k \to \infty$, kapjuk, hogy létezik olyan c konstans, hogy $|p - p_{k+1}| \le c|p_k - p|^{r_0}$, azaz a konvergencia rendje $r_0 = \frac{1+\sqrt{5}}{2}$. □

Láttuk, hogy a Newton-módszer többszörös gyökökre alkalmazva csak lineárisan konvergál. Belátható, hogy ez a szelőmódszerre is érvényes. Most azzal foglalkozunk, hogy lehet ezekben az esetekben felgyorsítani a konvergenciát.

Legyen $f \in C^3[a,b]$, és tegyük fel, hogy $p \in (a,b)$ többszörös gyöke f-nek, pontosabban feltesszük, hogy $f(x) = (x-p)^m q(x)$ alakú, ahol $m > 1$ és $q \in C^3[a,b]$. Definiáljuk a

$$\mu(x) = \begin{cases} \frac{f(x)}{f'(x)}, & \text{ha } x \neq p, \\ 0, & x = p \end{cases}$$

függvényt. Könnyen ellenőrizhető, hogy

$$\mu(x) = \frac{(x-p)q(x)}{mq(x) + (x-p)q'(x)},$$

ezért $\mu \in C^2[a,b]$, továbbá $\mu'(p) = \frac{1}{m}$, így p csak egyszeres gyöke μ-nek. Ezért ha f helyett a μ függvényre alkalmazzuk a Newton-módszert, kvadratikus konvergenciát kapunk. Ennek a módszernek a definíciója tehát

$$p_{k+1} = p_k - \frac{\mu(p_k)}{\mu'(p_k)} = p_k - \frac{f(p_k)f'(p_k)}{(f'(p_k))^2 - f(p_k)f''(p_k)}. \tag{2.23}$$

Feladatok

1. Mutassa meg, hogy az intervallumfelezés módszere lineárisan konvergens!
2. Lássa be a (2.19) egyenlőtlenséget!
3. Legyen $a > 0$. Mutassa meg, hogy a
$$p_{k+1} = \frac{p_k(p_k^2 + 3a)}{3p_k^2 + a}$$
egy harmadrendű lokálisan konvergens iterációs módszer $\sqrt{a}$ kiszámolására!
4. Adja meg a $p_k = \frac{1}{k}$ sorozat konvergencia rendjét! Mi a konvergencia rendje a $p_k = \frac{1}{k^n}$ sorozatnak?
5. Mutassa meg, hogy a $p_k = 10^{-2^k}$ sorozat másodrendben konvergál 0-hoz! Adjon meg egy olyan sorozatot, amely α-ad rendben konvergens!
6. Mutassa meg, hogy a $\sin^2 x$ függvénynek $x = 0$ kétszeres gyöke!
7. Igazolja a 2.33. tételt!
8. Tekintsük a következő iterációs módszereket:
   (a) (Halley-módszer:) $p_{k+1} = p_k - \dfrac{1}{a_k}$, ahol $a_k = \dfrac{f'(p_k)}{f(p_k)} - \dfrac{1}{2}\dfrac{f''(p_k)}{f'(p_k)}$,
   (b) (Olver-módszer:) $p_{k+1} = p_k - \dfrac{f(p_k)}{f'(p_k)} - \dfrac{1}{2}\dfrac{f''(p_k)}{f'(p_k)}\left(\dfrac{f(p_k)}{f'(p_k)}\right)^2$,
   Határozza meg az egyes módszerek konvergencia rendjét! Alkalmazza ezeket a módszereket a 2.3. szakasz 1. feladatában felsorolt egyenletekre!
9. Keresse meg az $f(x) = (x^2-2)^3$ függvény egy gyökét a Newton-iterációval, a szelőmódszerrel, a (2.23) iterációval és a
$$p_{k+1} = p_k - m\frac{f(p_k)}{f'(p_k)}$$
iterációval, ahol m a gyök multiplicitása! Hasonlítsa össze a módszerek konvergenciájának sebességét! Mi ez utóbbi módszer konvergenciájának rendje?

10. Tegyük fel, hogy egy f függvénynek már meghatároztuk az x_1 közelítő gyökét. Ha ezután a $g(x) = f(x)/(x - x_1)$ függvényre alkalmazunk egy gyökkereső eljárást, akkor azzal f egy másik gyökét is (vagy x_1-et újra, ha x_1 többszörös gyök volt) megkaphatjuk. Ezzel az ún. *deflációs* eljárással határozza meg az

$$\text{(a)} \quad f(x) = x^3 - 3x^2 + 4, \qquad \text{(b)} \quad f(x) = x^4 - 5x^3 + 9x^2 - 7x + 2$$

polinomok összes gyökét az egyes gyökök multiplicitásával együtt, tetszőleges gyökkereső algoritmust használva!

2.8. Iterációs módszerek megállási feltételei

Az eddigi módszerek mindegyike az f függvény egy gyökének meghatározására egy p_k sorozatot generált, amely (adott feltételek teljesülése esetén) konvergált az f függvény egy p gyökéhez. A gyök, azaz a sorozat határértékének közelítésére a sorozat egy p_k tagját használjuk, ahol k „elég nagy". Azt, hogy „meddig kell elmenni" a sorozat generálásában, többféle stratégiát használva dönthetjük el. Itt a három leggyakrabban használttal foglalkozunk. Előre megadunk $\varepsilon_1 > 0$, $\varepsilon_2 > 0$ és $\varepsilon_3 > 0$ tolerancia értékeket. A sorozat k-adik tagját, p_k-t tekintjük p közelítésének, ha

$$1.\ |p_k - p_{k-1}| < \varepsilon_1, \qquad 2.\ \frac{|p_k - p_{k-1}|}{|p_k|} < \varepsilon_2, \quad \text{vagy} \quad 3.\ |f(p_k)| < \varepsilon_3.$$

Az 1. feltétel a közelítés hibájának, $|p_k - p|$-nek numerikus megfelelője. Azt mondja, hogy ha a sorozat új tagja az előzőtől egy adott tolerancia értéknél kevesebbel tér el, akkor úgy gondoljuk, hogy azért változik csak kicsit az új érték a régihez képest, mert mindkettő már közel van a határértékhez, és ezért megszakítjuk a sorozat generálását.

A 2. feltétellel a közelítés relatív hibáját, $|p_k - p|/|p|$-et közelítjük numerikusan. Mint az előző feltételnél, itt is azt vizsgáljuk, hogy mennyit változik a sorozat következő tagja az előzőhöz képest, de a különbség képzésénél figyelembe vesszük a tagok nagyságrendjét.

A 3. feltétel szerint ha a függvényérték kicsi, akkor feltesszük, hogy közel vagyunk a gyökhöz, és megállunk.

Ezenkívül minden iterációs algoritmusba érdemes beépíteni az iteráció lépésszámának követését, és egy adott lépésszámot túllépve megállítani a program futását. Ezzel megakadályozhatjuk a program végtelen ciklusba kerülését, és kiszűrjük a túl lassú konvergenciát.

Az első két feltétel minden iterációs módszerre alkalmazható, a harmadik természetesen az ebben a fejezetben vizsgált feladatra, az f függvény gyökének meghatározására vonatkozik. Más feladatoknál többnyire meg lehet adni hasonló feltételt arra vonatkozólag, hogy egy adott közelítő megoldás „mennyire" elégíti ki az adott problémát (lásd pl. a 4.4. szakaszt később).

Mindegyik feltételhez lehet példát megadni, ahol a feltétel teljesülése nem vonja maga után azt, hogy a gyöknek jó közelítését kapjuk. Ezért a gyakorlatban, hogy az egyes feltételekkel kapcsolatos lehetséges problémákat kiszűrjük, ezeknek a megállási kritériumoknak kombinációit szokták használni.

Feladatok

1. Tegyük fel, hogy egy iterációs módszer a $p_k = \sum_{i=1}^{k} \frac{1}{i}$ sorozatot generálja, és tegyük fel, hogy az ebben a szakaszban leírt 1. feltételt használjuk csak megállási feltételként. Mit tapasztalunk? Konvergens-e a sorozat? Mit tapasztalunk ha csak a 2. megállási feltételt használjuk?
2. Legyen $f(x) = x^8$, és tegyük fel, hogy egy módszer a $p_k = 1/k$ sorozatot generálja f gyökének közelítésére. Tegyük fel, hogy csak az 1. feltételt használjuk megállási feltételként az $\varepsilon_1 = 10^{-8}$ tolerancia értékkel. Mi lesz az algoritmussal megadott közelítő gyök értéke? Mi lesz a gyök, ha csak a 2. és mi, ha csak a 3. feltételt használjuk az $\varepsilon_2 = 10^{-8}$ ill. $\varepsilon_3 = 10^{-8}$ tolerancia értékekkel?

2.9. Többváltozós analízis előismeretek

Ebben a szakaszban összefoglaljuk azokat a többváltozós analízis ismereteket, jelöléseket, amelyekre szükségünk lesz a fejezet hátralevő részében a nemlineáris egyenletrendszerek tárgyalásakor.

2.37. tétel. *Legyen $E \subset \mathbb{R}^n$ korlátos zárt halmaz, $f\colon E \to \mathbb{R}$ folytonos függvény. Ekkor f felveszi maximumát és minimumát E-n, azaz létezik olyan $\mathbf{c}, \mathbf{d} \in E$, hogy*

$$f(\mathbf{c}) = \max_{\mathbf{x}\in E} f(\mathbf{x}) \qquad \text{és} \qquad f(\mathbf{d}) = \min_{\mathbf{x}\in E} f(\mathbf{x}).$$

Legyen $E \subset \mathbb{R}^n$ és tekintsük az $f : E \to \mathbb{R}$ n-változós függvényt. Az $f = f(\mathbf{x}) = f(x_1, \ldots, x_n)$ függvény x_i változója szerinti parciális deriváltját $\frac{\partial f}{\partial x_i}$ jelöli. Ha az f függvény összes m-edrendű parciális deriváltja létezik és folytonos, akkor a függvényt m-szer folytonosan parciálisan differenciálhatónak nevezzük. Ezt a tulajdonságot az $f \in C^m$ jelöléssel rövidítjük. Ha $f \in C^1$, akkor f' az f függvény *gradiensvektorát* jelöli, azaz

$$f'(\mathbf{x}) := \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right)^T.$$

Ha $f \in C^2$, akkor $f''(\mathbf{x})$ jelöli az ún. *Hesse-mátrixot*:

$$f''(\mathbf{x}) := \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x}) \end{pmatrix}$$

Szükségünk lesz a Taylor-tétel többváltozós függvényekre vonatkozó alakjára.

2.38. tétel (Taylor-formula). *Legyen $E \subset \mathbb{R}^n$ nyílt halmaz, $f\colon E \to \mathbb{R}$, $f \in C^{m+1}$, és legyen $\mathbf{a} \in E$. Ekkor minden $\mathbf{x} \in E$-hez létezik olyan $\xi = \xi(\mathbf{x}) \in E$, hogy $\xi = \mathbf{x} + t(\mathbf{a} - \mathbf{x})$ valamely $t \in (0,1)$-re (azaz ξ az $\mathbf{a}$ és $\mathbf{x}$ vektorokat összekötő szakasz valamely pontja), és*

$$\begin{aligned} & f(x_1, \ldots, x_n) \\ &= f(a_1, \ldots, a_n) + \sum_{i=1}^{n} \frac{\partial f(a_1, \ldots, a_n)}{\partial x_i}(x_i - a_i) \\ &\quad + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f(a_1, \ldots, a_n)}{\partial x_i \, \partial x_j}(x_i - a_i)(x_j - a_j) \\ &\quad + \cdots + \frac{1}{m!} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \frac{\partial^m f(a_1, \ldots, a_n)}{\partial x_{i_1} \cdots \partial x_{i_m}}(x_{i_1} - a_{i_1}) \cdots (x_{i_m} - a_{i_m}) \\ &\quad + \frac{1}{(m+1)!} \sum_{i_1=1}^{n} \cdots \sum_{i_{m+1}=1}^{n} \frac{\partial^{m+1} f(\xi_1, \ldots, \xi_n)}{\partial x_{i_1} \cdots \partial x_{i_{m+1}}}(x_{i_1} - a_{i_1}) \cdots (x_{i_{m+1}} - a_{i_{m+1}}). \end{aligned}$$

A többváltozós Taylor-formulát többnyire $m = 1$-re vagy $m = 2$-re fogjuk használni, azaz egy függvényt elsőrendű vagy másodrendű Taylor-polinommal fogunk közelíteni. Az előbbi formulából könnyen ellenőrizhető, hogy a gradiensvektor és a Hesse-mátrix jelölést alkalmazva az $f \in C^3$ függvény másodrendű Taylor-közelítése az

$$f(\mathbf{x}) \approx f(\mathbf{a}) + f'(\mathbf{a})^T(\mathbf{x}-\mathbf{a}) + \frac{1}{2}(\mathbf{x}-\mathbf{a})^T f''(\mathbf{a})(\mathbf{x}-\mathbf{a})$$

alakban írható fel. Ez indokolja az f' és f'' jelölést a gradiensvektorra és a Hesse-mátrixra. A másik indok persze az, hogy egyszer illetve kétszer folytonosan parciálisan differenciálható függvényekre f' és f'' az f ill. f' függvény többváltozós analízisből ismert ún. totális vagy Fréchet-deriváltja. Erre a fogalomra nem lesz szükségünk a továbbiakban, így f'-t és f''-t mi jelölésnek tekinthetjük a gradiensvektorra ill. a Hesse-mátrixra.

Legyen $I \subset \mathbb{R}$, $g : I \to \mathbb{R}^n$. g komponensfüggvényeit jelölje g_i, azaz legyen $g(t) = (g_1(t), \dots, g_n(t))^T$. Ekkor g-t differenciálhatónak nevezzük, ha minden komponensfüggvénye differenciálható, és a deriváltján a

$$g' \colon I \to \mathbb{R}^n, \qquad g'(t) := (g_1'(t), \dots, g_n'(t))^T$$

függvényt értjük. g-t folytonosan differenciálhatónak nevezzük, ha minden komponensfüggvénye folytonosan differenciálható.

Érvényes a következő tétel.

2.39. tétel (láncszabály). *Legyen $f \colon \mathbb{R}^n \to \mathbb{R}$, $f \in C^1$ és $g \colon \mathbb{R} \to \mathbb{R}^n$ folytonosan differenciálható. Ekkor az $f \circ g \colon \mathbb{R} \to \mathbb{R}$ összetett függvény is folytonosan differenciálható, és*

$$\frac{d}{dt} f(g(t)) = f'(g(t))^T g'(t).$$

A láncszabály következményeként beláthatjuk a Lagrange-tétel következő általánosítását többváltozós valós függvényekre.

2.40. tétel (Lagrange-féle középértéktétel). *Legyen $E \subset \mathbb{R}^n$ nyílt, konvex halmaz, $f \colon E \to \mathbb{R}$ folytonosan parciálisan differenciálható. Ekkor minden $\mathbf{x}, \mathbf{y} \in E$-hez létezik olyan $\xi \in (0,1)$, hogy*

$$f(\mathbf{x}) - f(\mathbf{y}) = f'(\mathbf{y} + \xi(\mathbf{x}-\mathbf{y}))^T(\mathbf{x}-\mathbf{y}).$$

Bizonyítás. Definiáljuk a $g(t) = f(\mathbf{y} + t(\mathbf{x}-\mathbf{y}))$ egyváltozós valós függvényt $[0,1]$-en. Az egyváltozós valós függvényekre vonatkozó Lagrange-féle középértéktétel és a láncszabály szerint

$$f(\mathbf{x}) - f(\mathbf{y}) = g(1) - g(0) = g'(\xi) = f'(\mathbf{x} + \xi(\mathbf{y}-\mathbf{x}))^T(\mathbf{x}-\mathbf{y}).$$

□

Legyen $E \subset \mathbb{R}^n$ és $\mathbf{f}\colon E \to \mathbb{R}^n$. Az $\mathbf{f}$ függvény komponensfüggvényeit jelölje f_i, azaz

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_n(\mathbf{x}))^T .$$

Az $\mathbf{f}$ függvényt m-szer folytonosan parciálisan differenciálhatónak nevezzük, ha minden komponensfüggvényének minden m-edrendű parciális deriváltja létezik és folytonos. $\mathbf{f} \in C^m$ jelöli röviden azt, hogy $\mathbf{f}$ m-szer folytonosan parciálisan differenciálható. Az $\mathbf{f} \in C^1$ függvény *Jacobi-mátrixának* vagy derivált mátrixának az

$$\mathbf{f}'(\mathbf{x}) := \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f_n}{\partial x_n}(\mathbf{x}) \end{pmatrix}$$

$n \times n$-es mátrixot hívjuk.

Legyen $\mathbf{a}$ rögzített. Ha az $\mathbf{f}$ függvény komponensfüggvényeit az $\mathbf{a}$-körüli elsőrendű Taylor-polinomjaival közelítjük, akkor kapjuk, hogy

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{pmatrix} \approx \begin{pmatrix} f_1(\mathbf{a}) + f_1'(\mathbf{a})^T(\mathbf{x}-\mathbf{a}) \\ \vdots \\ f_n(\mathbf{a}) + f_n'(\mathbf{a})^T(\mathbf{x}-\mathbf{a}) \end{pmatrix} = \mathbf{f}(\mathbf{a}) + \mathbf{f}'(\mathbf{a})(\mathbf{x}-\mathbf{a}).$$

Az $\mathbf{f}(\mathbf{a}) + \mathbf{f}'(\mathbf{a})(\mathbf{x}-\mathbf{a})$ kifejezést az $\mathbf{f}$ függvény $\mathbf{a}$-körüli *lineáris közelítésének* hívjuk.

2.10. Vektor- és mátrixnormák, vektor- és mátrixsorozatok konvergenciája

Az $\mathbf{x} \in \mathbb{R}^n$ vektor komponenseit $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$-tal jelöljük. Az $\|\cdot\|\colon \mathbb{R}^n \to \mathbb{R}$ függvényt *vektornormának* nevezzük, ha

1. $\|\mathbf{x}\| \geq 0$ minden $\mathbf{x} \in \mathbb{R}^n$-re, és $\|\mathbf{x}\| = 0$ akkor és csak akkor, ha $\mathbf{x} = \mathbf{0}$,
2. $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$ minden $c \in \mathbb{R}$ és $\mathbf{x} \in \mathbb{R}^n$-re,
3. (háromszög-egyenlőtlenség:) $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ minden $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$-re.

2.41. tétel. *Egy tetszőleges $\|\cdot\|$ vektornormára*

1. $\Big|\|\mathbf{x}\| - \|\mathbf{y}\|\Big| \leq \|\mathbf{x}-\mathbf{y}\|$,
2. *$\|\cdot\|$ folytonos függvény $\mathbb{R}^n$-en.*

Bizonyítás. A háromszög-egyenlőtlenség alapján $\|\mathbf{x}\| = \|\mathbf{x}-\mathbf{y}+\mathbf{y}\| \leq \|\mathbf{x}-\mathbf{y}\| + \|\mathbf{y}\|$, amiből $\|\mathbf{x}\| - \|\mathbf{y}\| \leq \|\mathbf{x}-\mathbf{y}\|$ következik. Ugyanígy $\|\mathbf{y}\| - \|\mathbf{x}\| \leq \|\mathbf{x}-\mathbf{y}\|$ is teljesül, így az 1. állítás igaz. A $\|\cdot\|$ norma függvény folytonossága következik az 1. pontban bizonyított egyenlőtlenségből. □

Legyen $p \geq 1$, és definiáljuk az ún. *p-normát*:

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} .$$

Belátható, hogy $\|\cdot\|_p$ teljesíti a vektornorma definícióját minden $p \geq 1$-re. A $p = 2$-höz tartozó $\|\cdot\|_2$ normát *euklideszi normának* is szokás nevezni. Egy gyakran használt speciális eset az 1-norma:

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

Egy másik gyakran használt vektornorma az ún. *végtelen norma*

$$\|\mathbf{x}\|_\infty := \max_{i=1,\ldots,n} |x_i|.$$

Az olvasóra bízzuk annak igazolását, hogy $\|\cdot\|_1$ és $\|\cdot\|_\infty$ teljesítik a norma tulajdonságait (1. feladat). Az euklideszi norma is nyilvánvalóan teljesíti a norma definíciójának 1. és 2. tulajdonságát. A háromszög-egyenlőtlenség igazolásához viszont szükség van a következő, önmagában is igen fontos egyenlőtlenségre.

2.42. tétel (Cauchy–Bunyakovszkij–Schwarz egyenlőtlenség). *Minden $x_1, \ldots, x_n$, $y_1, \ldots, y_n \in \mathbb{R}$-re teljesül a*

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2$$

egyenlőtlenség, ahol akkor és csak akkor áll fenn egyenlőség, ha létezik olyan $\lambda \in \mathbb{R}$, hogy $y_i = \lambda x_i$ minden $i = 1, 2, \ldots, n$-re.

Bizonyítás. Tekintsük a $p(t) := t^2 \sum_{i=1}^{n} x_i^2 - 2t \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2$ másodfokú polinomot. Ekkor $p(t) = \sum_{i=1}^{n} (tx_i - y_i)^2 \geq 0$ teljesül minden t-re, így p-nek nem lehet két valós gyöke, azaz p diszkriminánsa nem lehet pozitív:

$$4\left(\sum_{i=1}^{n} x_i y_i\right)^2 - 4\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2 \leq 0.$$

Ebből kapjuk a tétel állításában szereplő egyenlőtlenséget. p-nek akkor és csak akkor lehet pontosan egy valós gyöke, ha a diszkriminánsa egyenlő nullával, azaz a tétel állításában egyenlőség szerepel. Másrészt $p(t) = 0$ akkor és csak akkor teljesül valamely $t = \lambda$-ra, ha minden $i = 1, 2, \ldots, n$-re $y_i = \lambda x_i$. □

A Cauchy–Bunyakovszkij–Schwarz egyenlőtlenség mindkét oldalából gyököt vonva és vektoriális jelölést alkalmazva kapjuk:

2.43. következmény. *Tetszőleges $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$-re*

$$|\mathbf{x}^T \mathbf{y}| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

teljesül, ahol egyenlőség akkor és csak akkor van, ha $\mathbf{y} = \lambda \mathbf{x}$ valamely $\lambda \in \mathbb{R}$-re.

A Cauchy–Bunyakovszkij–Schwarz egyenlőtlenség alapján

$$\begin{aligned}
\|\mathbf{x}+\mathbf{y}\|_2^2 &= \sum_{i=1}^{n}(x_i+y_i)^2 \\
&= \sum_{i=1}^{n}x_i^2 + 2\sum_{i=1}^{n}x_iy_i + \sum_{i=1}^{n}y_i^2 \\
&\le \sum_{i=1}^{n}x_i^2 + 2\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2} + \sum_{i=1}^{n}y_i^2 \\
&= \left(\sqrt{\sum_{i=1}^{n}x_i^2} + \sqrt{\sum_{i=1}^{n}y_i^2}\right)^2 \\
&= (\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)^2,
\end{aligned}$$

ami igazolja, hogy az euklideszi norma teljesíti a háromszög-egyenlőlenséget.

A normák segítségével értelmezhetjük vektorok hosszát, távolságát, valamint vektorsorozatok határértékét. A $\|\mathbf{x}\|$-t az $\mathbf{x}$ *vektor hosszának*, azaz a $\mathbf{0}$-tól való távolságának nevezzük. Az $\mathbf{x}$ és $\mathbf{y}$ *vektorok távolságán* az $\|\mathbf{x}-\mathbf{y}\|$ számot értjük. Legyen $\mathbf{p}^{(k)}$ n-dimenziós vektoroknak egy sorozata, és $\|\cdot\|$ egy vektornorma $\mathbb{R}^n$-en. Azt mondjuk, hogy a $\mathbf{p}^{(k)}$ sorozat a $\mathbf{p}$ vektorhoz konvergál, ha

$$\lim_{k\to\infty}\|\mathbf{p}^{(k)}-\mathbf{p}\| = 0.$$

Belátható, hogy a konvergencia fogalma független a definícióban használt vektornorma választásától, azaz ha egy sorozat egy vektornormában konvergens, akkor egy tetszőleges másik vektornormában is az, és ugyanahhoz a vektorhoz konvergál. (Ezt a tulajdonságot hívják az analízisben úgy, hogy $\mathbb{R}^n$-en a vektornormák ekvivalensek.)

2.44. tétel. *Legyen $|\cdot|$ és $\|\cdot\|$ két vektornorma, és $\mathbf{p}^{(k)}$ egy vektorsorozat $\mathbb{R}^n$-en. Ekkor $\lim_{k\to\infty}|\mathbf{p}^{(k)}-\mathbf{p}| = 0$ akkor és csak akkor, ha $\lim_{k\to\infty}\|\mathbf{p}^{(k)}-\mathbf{p}\| = 0$.*

Bizonyítás. Elegendő megmutatni, hogy $\|\mathbf{p}^{(k)}-\mathbf{p}\|\to 0$ akkor és csak akkor, ha $\|\mathbf{p}^{(k)}-\mathbf{p}\|_1\to 0$, ahol $\|\cdot\|$ egy tetszőleges norma $\mathbb{R}^n$-en. Ez teljesül, ha belátjuk, hogy léteznek olyan m és M konstansok, hogy

$$m\|\mathbf{p}^{(k)}-\mathbf{p}\|_1 \le \|\mathbf{p}^{(k)}-\mathbf{p}\| \le M\|\mathbf{p}^{(k)}-\mathbf{p}\|_1. \tag{2.24}$$

Legyen $E := \{\|\mathbf{x}\in\mathbb{R}^n : \|\mathbf{x}\|_1 = 1\}$. E korlátos és zárt részhalmaza $\mathbb{R}^n$-nek, ezért a 2.37. és 2.41 tételek szerint a $\|\cdot\|$ folytonos függvény felveszi maximumát és minimumát E-n. Legyenek ezek M és m. Legyen $\mathbf{x} = (\mathbf{p}^{(k)}-\mathbf{p})/\|\mathbf{p}^{(k)}-\mathbf{p}\|_1$. Ekkor $\mathbf{x}\in E$, ezért $m\le\|\mathbf{x}\|\le M$, amit beszorozva $\|\mathbf{p}^{(k)}-\mathbf{p}\|_1$-val kapjuk a (2.24) egyenlőtlenséget. □

2.45. tétel. *Legyen a $\mathbf{p}^{(k)}$ és a $\mathbf{p}$ vektor i-edik komponense $p_i^{(k)}$ ill. p_i. Ekkor a $\mathbf{p}^{(k)}$ vektorsorozat akkor és csak akkor konvergál a $\mathbf{p}$ vektorhoz, ha $p_i^{(k)}\to p_i$ minden $i = 1,2,\dots,n$-re, ha $k\to\infty$.*

Bizonyítás. A 2.44. tétel szerint $\|\mathbf{p}^{(k)}-\mathbf{p}\|\to 0$ akkor és csak akkor, ha $\|\mathbf{p}^{(k)}-\mathbf{p}\|_1 = \sum_{i=1}^{n}|\mathbf{p}_i^{(k)}-p_i|\to 0$, ami pontosan akkor teljesül, ha $p_i^{(k)}\to p_i$ minden $i = 1,2,\dots,n$-re. □

Az $n \times n$-es valós mátrixok halmazát $\mathbb{R}^{n\times n}$-nel jelöljük. Legyen $\|\cdot\|$ egy vektornorma $\mathbb{R}^n$-en. Az

$$\|\mathbf{A}\| := \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|}$$

képlettel definiált $\|\cdot\|\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ függvényt az $\|\cdot\|$ vektornorma által generált *mátrixnormának* nevezzük. (A jelölésben nem teszünk különbséget a vektornorma és az általa generált mátrixnorma között.) Megmutatható, hogy a mátrixnorma definícióban szereplő sup (azaz legkisebb felső korlát) max-ra cserélhető, azaz létezik olyan $\mathbf{x}$ vektor, amelyre $\|\mathbf{A}\| = \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|}$. Könnyen beláthatók a mátrixnorma következő tulajdonságai:

2.46. tétel. *Minden $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$-re*

1. $\|\mathbf{A}\| \geq 0$, *és* $\|\mathbf{A}\| = 0$ *akkor és csak akkor, ha* $\mathbf{A} = \mathbf{0}$,
2. $\|c\mathbf{A}\| = |c|\|\mathbf{A}\|$ *minden* $c \in \mathbb{R}$*-re,*
3. *(háromszög-egyenlőtlenség:)* $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$,
4. $\|\mathbf{Ax}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$, *minden* $\mathbf{x} \in \mathbb{R}^n$*-re,*
5. $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$,
6. $\|\mathbf{A}\| = \sup\{\|\mathbf{Ay}\|\colon \|\mathbf{y}\| = 1\}$.

Bizonyítás. Az 1., 2. és 3. állítások bizonyítását az olvasóra hagyjuk. A 4. állítás következik az

$$\frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \leq \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{Ay}\|}{\|\mathbf{y}\|} = \|\mathbf{A}\|$$

egyenlőtlenségből. A 4. állítást felhasználva

$$\frac{\|\mathbf{ABx}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\| \frac{\|\mathbf{Bx}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\|\mathbf{B}\|,$$

ezért

$$\|\mathbf{AB}\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{ABx}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\|\mathbf{B}\|.$$

Végül a 6. állítás következik az

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

egyenlőségből. □

Megjegyezzük, hogy a mátrixnormát általánosabban is lehet definiálni a vektornorma definíciójához hasonlóan: egy olyan $\|\cdot\|\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ függvény, amely teljesíti a 2.46. tétel első 1.–3. és 5. tulajdonságait. Vannak olyan mátrixnormák, amelyek nem vektornormák által generált mátrixnormák. Nekünk a továbbiakban elegendő csak a vektornormák által generált mátrixnormákat használni, ezért fogalmaztuk így a definíciót.

A következő tétel szerint bármely két mátrixnorma ekvivalens.

2.47. tétel. *Jelöljön $|\cdot|$ és $\|\cdot\|$ két vektornormát ill. az általa generált mátrixnormát. Legyen $\mathbf{A}^{(k)}$ egy vektorsorozat $\mathbb{R}^{n\times n}$-en. Ekkor $\lim_{k\to\infty} |\mathbf{A}^{(k)} - \mathbf{A}| = 0$ akkor és csak akkor, ha $\lim_{k\to\infty} \|\mathbf{A}^{(k)} - \mathbf{A}\| = 0$.*

Bizonyítás. Most is, mint a 2.44. tétel bizonyításában, elegendő megmutatni, hogy léteznek olyan l, L nemnegatív konstansok, hogy

$$l|\mathbf{B}| \le \|\mathbf{B}\| \le L|\mathbf{B}|, \qquad \mathbf{B} \in \mathbb{R}^{n\times n}.$$

A 2.44. tétel bizonyításából következik, hogy létezik olyan $m, M > 0$, hogy

$$m|\mathbf{x}| \le \|\mathbf{x}\| \le M|\mathbf{x}|, \qquad x \in \mathbb{R}^n.$$

Ekkor

$$\frac{m}{M}|\mathbf{B}| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{m|\mathbf{Bx}|}{M|\mathbf{x}|} \le \|\mathbf{B}\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Bx}\|}{\|\mathbf{x}\|} \le \sup_{\mathbf{x}\neq\mathbf{0}} \frac{M|\mathbf{Bx}|}{m|\mathbf{x}|} = \frac{M}{m}|\mathbf{B}|,$$

amiből következik a tétel állítása. □

A gyakorlatban az 1-es és a végtelen vektornormák által generált mátrixnormákat használjuk leggyakrabban. Ezek kiszámolására vonatkozik a következő tétel:

2.48. tétel. *Legyen* $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n\times n}$. *Ekkor az* $\|\cdot\|_1$ *és* $\|\cdot\|_\infty$ *vektornormák által generált mátrixnorma*

$$\|\mathbf{A}\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|,$$

illetve

$$\|\mathbf{A}\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|.$$

Bizonyítás. Csak az első képletet indokoljuk, a második bizonyítását az olvasóra hagyjuk. Az $\|\cdot\|_1$ vektornorma definíciója és a háromszög-egyenlőtlenség alapján

$$\begin{aligned}
\|\mathbf{Ax}\|_1 &= \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \\
&\le \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}x_j| \\
&= \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| \\
&\le \left(\max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}| \right) \sum_{j=1}^{n} |x_j| \\
&= \left(\max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}| \right) \|\mathbf{x}\|_1,
\end{aligned}$$

ezért $\|\mathbf{A}\|_1 \le \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|$. Tegyük fel, hogy $\max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}| = \sum_{i=1}^{n} |a_{ik}|$. Az egyenlőséget abból kapjuk, hogy ha az $\mathbf{e}^{(k)} = (0,\dots,0,1,0,\dots,0)^T$ vektorra alkalmazzuk $\mathbf{A}$-t, ahol $e_i^{(k)} = 0$ ha $i \neq k$ és $e_k^{(k)} = 1$, akkor $\mathbf{Ae}^{(k)} = (a_{1k}, a_{2k}, \dots, a_{nk})^T$, így $\|\mathbf{Ae}^{(k)}\|_1 = \sum_{i=1}^{n} |a_{ik}|$. □

A valós számsorozatok tulajdonságainak egyszerű általánosításából kapjuk:

2.49. tétel.

1. *Ha a* $\mathbf{p}^{(k)}$ *vektorsorozat konvergens, akkor a határérték egyértelmű.*
2. *Ha* $\mathbf{p}^{(k)} \to \mathbf{p}$ *és* $\mathbf{q}^{(k)} \to \mathbf{q}$, $\alpha, \beta \in \mathbb{R}$, *akkor* $\alpha\mathbf{p}^{(k)} + \beta\mathbf{q}^{(k)}$ *konvergens, és* $\alpha\mathbf{p}^{(k)} + \beta\mathbf{q}^{(k)} \to \alpha\mathbf{p} + \beta\mathbf{q}$.
3. *Ha* $c_k \to c$ *valós számsorozat és* $\mathbf{p}^{(k)} \to \mathbf{p}$, *akkor* $c_k\mathbf{p}^{(k)} \to c\mathbf{p}$.
4. *Ha* $\mathbf{p}^{(k)} \to \mathbf{p}$, *akkor* $\mathbf{A}\mathbf{p}^{(k)} \to \mathbf{A}\mathbf{p}$ *minden* $\mathbf{A} \in \mathbb{R}^{n\times n}$*-re.*
5. *(Cauchy-féle konvergenciakritérium)* $\mathbf{p}^{(k)}$ *akkor és csak akkor konvergens, ha Cauchy-sorozat, azaz bármely* $\varepsilon > 0$*-hoz létezik olyan* $k_0 > 0$ *küszöbszám, hogy* $\|\mathbf{p}^{(k)} - \mathbf{p}^{(m)}\| < \varepsilon$ *minden* $k, m > k_0$*-ra.*

Mátrixokra értelemszerűen kiterjeszthető a hosszúság, távolság és a konvergencia fogalma, és érvényes a 2.44., a 2.45. és 2.49. tételek megfelelő kiterjesztése.

A vektor- és mátrixnorma alkalmazásával általánosítható a Lagrange-féle középértéktétel többváltozós vektor értékű függvényekre.

2.50. tétel (Lagrange-féle középértéktétel). *Jelöljön* $\|\cdot\|$ *egy tetszőleges vektornormát* $\mathbb{R}^n$*-en illetve az általa generált mátrixnormát. Legyen* $E \subset \mathbb{R}^n$ *nyílt konvex halmaz,* $\mathbf{f}\colon E \to \mathbb{R}^n$ *folytonosan parciálisan differenciálható,* $\mathbf{x}, \mathbf{y} \in E$. *Ekkor*

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \le \max_{t\in[0,1]} \|\mathbf{f}'(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))\| \cdot \|\mathbf{x} - \mathbf{y}\|.$$

Bizonyítás. Az állítást csak a $\|\cdot\| = \|\cdot\|_2$ speciális esetben bizonyítjuk. Nyilván feltehető, hogy $\mathbf{f}(\mathbf{x}) \neq \mathbf{f}(\mathbf{y})$. Legyen

$$\mathbf{h} := \frac{\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})}{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|_2}.$$

Ekkor $\|\mathbf{h}\|_2 = 1$. Legyen $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_n(\mathbf{x}))^T$, $\mathbf{h} = (h_1, \dots, h_n)^T$. Definiáljuk a

$$g(t) := \mathbf{h}^T\mathbf{f}(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) = \sum_{i=1}^{n} h_i f_i(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))$$

valós függvényt. Ekkor az egyváltozós függvényekre vonatkozó Lagrange-féle középértéktétel és a láncszabály szerint

$$\begin{aligned}
\mathbf{h}^T(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})) &= g(1) - g(0) \\
&= g'(\xi) \\
&= \sum_{i=1}^{n} h_i f_i'(\mathbf{y} + \xi(\mathbf{x} - \mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \\
&= \mathbf{h}^T\mathbf{f}'(\mathbf{y} + \xi(\mathbf{x} - \mathbf{y}))(\mathbf{x} - \mathbf{y})
\end{aligned}$$

valamely $\xi \in (0, 1)$-re. Így $\mathbf{h}$ definíciója, a Cauchy-Bunyakovszkij-Schwarz egyenlőtlenség vektoriális alakja, $\|\mathbf{h}\|_2 = 1$ és a mátrixnorma 2.46. tétel 5. tulajdonsága alapján

$$\begin{aligned}
\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|_2 &= \mathbf{h}^T(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})) \\
&= \mathbf{h}^T\mathbf{f}'(\mathbf{y} + \xi(\mathbf{x} - \mathbf{y}))(\mathbf{x} - \mathbf{y}) \\
&\le \|\mathbf{h}\|_2\|f'(\mathbf{y} + \xi(\mathbf{x} - \mathbf{y}))(\mathbf{x} - \mathbf{y})\|_2 \\
&\le \|f'(\mathbf{y} + \xi(\mathbf{x} - \mathbf{y}))\|_2\|\mathbf{x} - \mathbf{y}\|_2,
\end{aligned}$$

amiből következik a tétel állítása. □

Feladatok

1. Mutassa meg, hogy $\|\cdot\|_1$ és $\|\cdot\|_\infty$ teljesítik a vektornorma tulajdonságait!
2. Számítsa ki az $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$ és $\|\mathbf{x}\|_\infty$, ill. az $\|\mathbf{A}\|_1$ és $\|\mathbf{A}\|_\infty$ normákat, ha
$$\text{(a)}\quad \mathbf{x}=(3,-1,0,5)^T, \qquad \text{(b)}\quad \mathbf{x}=(-3,-2,-1,4,-1)^T,$$
illetve
$$\text{(c)}\quad \mathbf{A}=\begin{pmatrix} -1 & 3 & -2 \\ 2 & -4 & 0 \\ 0 & 3 & 2 \end{pmatrix}, \qquad \text{(d)}\quad \mathbf{A}=\begin{pmatrix} -1 & 2 & 4 \\ 2 & -3 & 5 \\ 7 & -2 & 3 \end{pmatrix}.$$
3. Rajzolja fel az
$$\text{(a)}\quad \{\mathbf{x}\in\mathbb{R}^2\colon \|\mathbf{x}\|_1=1\}, \qquad \text{(b)}\quad \{\mathbf{x}\in\mathbb{R}^2\colon \|\mathbf{x}\|_\infty=1\}$$
síkbeli halmazokat!
4. Lássa be a 2.46. tétel 1.–3. állításait!
5. Igazolja a 2.48. tétel 2. állítását!
6. Bizonyítsa be a 2.49. tételt!

2.11. Fixpont tétel n-dimenzióban

Az egyváltozós függvényekre definiált fixpont és a fixpont iteráció fogalmát és annak tulajdonságait könnyen általánosíthatjuk többváltozós függvényekre.

2.51. példa. Tekintsük a
$$\begin{aligned} 4x_1 - e^{x_1x_2} - 3 &= 0 \\ x_1 - x_2^2 - 3x_2 - 1 &= 0. \end{aligned} \tag{2.25}$$
egyenletrendszert. Ennek megoldása $x_1=1$ és $x_2=0$. Alakítsuk át a (2.25) rendszert a következő módon. Fejezzük ki az első egyenletből x_1-et, a másodikból pedig x_2-t:
$$\begin{aligned} x_1 &= \tfrac{1}{4}(e^{x_1x_2}+3) \\ x_2 &= \tfrac{1}{3}(x_1-x_2^2-1) \end{aligned} \tag{2.26}$$
A (2.26) egyenletrendszert röviden az $\mathbf{x}=\mathbf{g}(\mathbf{x})$ alakban írhatjuk fel a vektoriális jelölést alkalmazva, ahol $\mathbf{x}=(x_1,x_2)^T$ és
$$\mathbf{g}(\mathbf{x})=\mathbf{g}(x_1,x_2)=\begin{pmatrix} \frac{1}{4}(e^{x_1x_2}+3) \\ \frac{1}{3}(x_1-x_2^2-1) \end{pmatrix}. \tag{2.27}$$
Az egyváltozós fixpont iterációhoz hasonlóan (2.26) megoldására definiáljuk a következő iterációt $k=0,1,2,\ldots$-re:
$$\begin{aligned} p_1^{(k+1)} &= \tfrac{1}{4}(e^{p_1^{(k)}p_2^{(k)}}+3) \\ p_2^{(k+1)} &= \tfrac{1}{3}\left(p_1^{(k)}-(p_2^{(k)})^2-1\right). \end{aligned} \tag{2.28}$$
A $p_1^{(0)}=-2$ és $p_2^{(0)}=-2$ kezdőértékekből kiindulva kiszámoltuk a $p_1^{(k)}$ és $p_2^{(k)}$ sorozatok első néhány tagját a 2.12. táblázatban. Látható, hogy a sorozatok konvergálnak 1-hez ill. 0-hoz.

Definiálva a $\mathbf{p}^{(k)}=(p_1^{(k)},p_2^{(k)})^T$ vektorsorozatot, a (2.28) egyenletrendszert röviden a $\mathbf{p}^{(k+1)}=\mathbf{g}(\mathbf{p}^{(k)})$ alakban írhatjuk fel. □

Legyen $E\subset\mathbb{R}^n$, és tekintsünk egy $\mathbf{g}\colon E\to\mathbb{R}^n$ függvényt. Az egyváltozós esethez hasonlóan, a $\mathbf{p}\in E$ vektort a $\mathbf{g}$ függvény fixpontjának nevezzük, ha $\mathbf{p}=\mathbf{g}(\mathbf{p})$.

Egy $\mathbf{g}\colon E\to\mathbb{R}^n$ függvény *kontrakció* az E halmazon a $\|\cdot\|$ vektornormában, ha létezik egy $0\le c<1$ szám, hogy $\|\mathbf{g}(\mathbf{x})-\mathbf{g}(\mathbf{y})\|\le c\|\mathbf{x}-\mathbf{y}\|$ minden $\mathbf{x},\mathbf{y}\in E$-re.

2.12. táblázat. Fixpont iteráció

k	$p_1^{(k)}$	$p_2^{(k)}$
0	-2.000000000	-2.000000000
1	14.399537510	-2.333333333
2	0.750000000	2.651697690
3	2.576641266	-2.427166879
4	0.750480717	-1.438165931
5	0.834956989	-0.772613509
6	0.881152644	-0.253991549
7	0.949867689	-0.061119687
8	0.985899367	-0.017955976
9	0.995613247	-0.004807684
10	0.998806211	-0.001469956
11	0.999633219	-0.000398650
12	0.999900394	-0.000122313

2.52. tétel (fixpont tétel). *Legyen $E \subset \mathbb{R}^n$ zárt, $\mathbf{g}: E \to E$, és legyen $\mathbf{g}$ kontrakció az E halmazon valamely $\|\cdot\|$ normában. Ekkor $\mathbf{g}$-nek létezik egyértelmű $\mathbf{p} \in E$ fixpontja, és a $\mathbf{p}^{(k+1)} = \mathbf{g}(\mathbf{p}^{(k)})$ fixpont iteráció $\mathbf{p}$-hez konvergál minden $\mathbf{p}^{(0)} \in E$ kezdeti értékre. A konvergencia rendje legalább lineáris.*

Bizonyítás. Belátjuk, hogy a $\mathbf{p}^{(k)}$ sorozat Cauchy-sorozat. Legyen c a $\mathbf{g}$ függvény Lipschitz-konstansa, és legyen $k > m$. Az egyváltozós esethez hasonlóan a fixpont sorozat definíciója és a kontrakciós tulajdonságból kapjuk

$$\begin{aligned}
&\|\mathbf{p}^{(k)} - \mathbf{p}^{(m)}\| \\
&\quad\le \|\mathbf{p}^{(k)} - \mathbf{p}^{(k-1)}\| + \|\mathbf{p}^{(k-1)} - \mathbf{p}^{(k-2)}\| + \cdots + \|\mathbf{p}^{(m+1)} - \mathbf{p}^{(m)}\| \\
&\quad= \|\mathbf{g}(\mathbf{p}^{(k-1)}) - \mathbf{g}(\mathbf{p}^{(k-2)})\| + \|\mathbf{g}(\mathbf{p}^{(k-2)}) - \mathbf{g}(\mathbf{p}^{(k-3)})\| \\
&\qquad\quad + \cdots + \|\mathbf{g}(\mathbf{p}^{(m)}) - \mathbf{g}(\mathbf{p}^{(m-1)})\| \\
&\quad\le c(\|\mathbf{p}^{(k-1)} - \mathbf{p}^{(k-2)}\| + \|\mathbf{p}^{(k-2)} - \mathbf{p}^{(k-3)}\| + \cdots + \|\mathbf{p}^{(m)} - \mathbf{p}^{(m-1)}\|) \\
&\quad\le (c^{k-1} + c^{k-2} + \cdots + c^m)\|\mathbf{p}^{(1)} - \mathbf{p}^{(0)}\| \\
&\quad= c^m(c^{k-m-1} + c^{k-m-2} + \cdots + 1)\|\mathbf{p}^{(1)} - \mathbf{p}^{(0)}\| \\
&\quad\le c^m \sum_{i=0}^{\infty} c^i \|\mathbf{p}^{(1)} - \mathbf{p}^{(0)}\|.
\end{aligned}$$

Ebből adódik hogy $\|\mathbf{p}^{(k)} - \mathbf{p}^{(m)}\| \to 0$, ha $m \to \infty$, tehát $\mathbf{p}^{(k)}$ Cauchy-sorozat. A 2.49. tétel 5. pontja szerint $\mathbf{p}^{(k)}$ konvergál egy $\mathbf{p}$ vektorhoz. A $\mathbf{g}$ függvény folytonossága alapján ekkor $\mathbf{p}^{(k+1)} = \mathbf{g}(\mathbf{p}^{(k)}) \to \mathbf{g}(\mathbf{p})$, ezért $\mathbf{p} = \mathbf{g}(\mathbf{p})$, azaz $\mathbf{p}$ fixpontja $\mathbf{g}$-nek.

A konvergencia rendje legalább lineáris, hiszen

$$\|\mathbf{p}^{(k+1)} - \mathbf{p}\| = \|\mathbf{g}(\mathbf{p}^{(k)}) - \mathbf{g}(\mathbf{p})\| \le c\|\mathbf{p}^{(k)} - \mathbf{p}\|.$$

Tegyük fel, hogy $\mathbf{p}$ és $\bar{\mathbf{p}}$ fixpontjai $\mathbf{g}$-nek. A $\mathbf{g}$ függvény kontrakciós tulajdonsága alapján $\|\mathbf{p} - \bar{\mathbf{p}}\| = \|\mathbf{g}(\mathbf{p}) - \mathbf{g}(\bar{\mathbf{p}})\| \le c\|\mathbf{p} - \bar{\mathbf{p}}\|$, amiből $\mathbf{p} = \bar{\mathbf{p}}$ következik. □

2.53. tétel. *Legyen $E \subset \mathbb{R}^n$ nyílt halmaz, $\mathbf{g}: E \to \mathbb{R}^n$, $\mathbf{g} \in C^1$, és legyen $\mathbf{p}$ fixpontja $\mathbf{g}$-nek. Ha $\|\mathbf{g}'(\mathbf{p})\| < 1$ valamilyen $\|\cdot\|$ vektornorma által generált mátrixnormában, akkor a $\mathbf{p}^{(k+1)} = \mathbf{g}(\mathbf{p}^{(k)})$ fixpont iteráció lokálisan konvergál $\mathbf{p}$-hez.*

Bizonyítás. Mivel E nyílt halmaz, ezért létezik olyan $\bar{\delta} > 0$, hogy $\{\mathbf{x}: \|\mathbf{x} - \mathbf{p}\| < \bar{\delta}\} \subset E$. Válasszunk egy c számot, amelyre $\|\mathbf{g}'(\mathbf{p})\| < c < 1$. A $\mathbf{g}'$ függvény folytonos $\mathbf{p}$-ben, így létezik

olyan $0 < \delta \le \bar{\delta}$, hogy $\|\mathbf{g}'(\mathbf{x})\| \le c$ minden $\mathbf{x} \in V := \{\mathbf{x} : \|\mathbf{x} - \mathbf{p}\| \le \delta\}$-ra. A Lagrange-féle középértéktétel (2.50. tétel) alapján

$$\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| \le \max_{t\in(0,1)} \|\mathbf{g}'(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))\| \cdot \|\mathbf{x} - \mathbf{y}\| \le c\|\mathbf{x} - \mathbf{y}\|,$$

azaz $\mathbf{g}$ kontrakció.

Megmutatjuk, hogy a $\mathbf{g}$ függvény a V halmazt önmagába képezi. Legyen $\mathbf{x} \in V$. A $\mathbf{g}$ függvény kontrakciós tulajdonsága alapján $\|\mathbf{g}(\mathbf{x}) - \mathbf{p}\| = \|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{p})\| \le c\|\mathbf{x} - \mathbf{p}\| < \delta$, tehát $\mathbf{g}(\mathbf{x}) \in V$. Ha a $\mathbf{g}$ függvényt megszorítjuk a V halmazra, akkor erre a függvényre teljesülnek a 2.52. tétel feltételei, ezért a V halmazból indított fixpont iteráció konvergens, és $\mathbf{p}$-hez konvergál. □

2.54. példa. Számítsuk ki a 2.51. feladatban szereplő, a (2.27) képlettel definiált $\mathbf{g}$ függvény derivált mátrixát:

$$\mathbf{g}'(\mathbf{x}) = \begin{pmatrix} \frac{1}{4}x_2 e^{x_1x_2} & \frac{1}{4}x_1 e^{x_1x_2} \\ \frac{1}{3} & -\frac{2}{3}x_2 \end{pmatrix}.$$

Ennek a $\mathbf{g}$ függvény $(1,0)^T$ fixpontjában felvett értéke

$$\mathbf{g}'(1,0) = \begin{pmatrix} 0 & \frac{1}{4} \\ \frac{1}{3} & 0 \end{pmatrix}.$$

aminek 1-normája $\|\mathbf{g}'(1,0)\|_1 = \frac{1}{3} < 1$, ezért a 2.53. tétel szerint a fixpont sorozat lokálisan konvergens. □

2.55. tétel. *Legyen $E \subset \mathbb{R}^n$, $\mathbf{g} : E \to \mathbb{R}^n$, $\mathbf{g} \in C^2$, $\mathbf{g}(\mathbf{p}) = \mathbf{p}$, és $\mathbf{g}'(\mathbf{p}) = \mathbf{0}$. Ekkor létezik olyan $\delta > 0$ hogy a $\mathbf{p}^{(k+1)} = \mathbf{g}(\mathbf{p}^{(k)})$ fixpont iteráció konvergál $\mathbf{p}$-hez, ha $\|\mathbf{p}^{(0)} - \mathbf{p}\|_\infty < \delta$. Továbbá létezik olyan c konstans, hogy minden k-ra $\|\mathbf{p}^{(k+1)} - \mathbf{p}\|_\infty \le c\|\mathbf{p}^{(k)} - \mathbf{p}\|_\infty^2$ teljesül, azaz az iteráció másodrendben lokálisan konvergál $\mathbf{p}$-hez.*

Bizonyítás. A feltétel szerint $0 = \|\mathbf{g}'(\mathbf{p})\| < 1$, így a 2.53. tételből következik, hogy a fixpont iteráció lokálisan konvergens.

Most belátjuk, hogy a konvergencia kvadratikus. Vegyük a $\mathbf{g}$ függvény i-edik komponensfüggvényének a $\mathbf{p} = (p_1, \dots, p_n)^T$ pont körüli másodrendű Taylor-közelítését:

$$\begin{aligned} g_i(x_1, \dots, x_n) &= g_i(p_1, \dots, p_n) + \sum_{j=1}^{n} \frac{\partial g_i(p_1, \dots, p_n)}{\partial x_j}(x_j - p_j) \\ &\quad + \frac{1}{2}\sum_{j=1}^{n}\sum_{l=1}^{n} \frac{\partial^2 g_i(\xi_1, \dots, \xi_n)}{\partial x_j\, \partial x_l}(x_j - p_j)(x_l - p_l) \end{aligned}$$

Ezt az $(x_1, \dots, x_n)^T = (p_1^{(k)}, \cdots, p_n^{(k)})^T$ vektorra alkalmazva, és használva a $p_i = g_i(\mathbf{p})$ és $p_i^{(k+1)} = g_i(\mathbf{p}^{(k)})$ összefüggéseket, kapjuk

$$p_i^{(k+1)} - p_i = \frac{1}{2}\sum_{j=1}^{n}\sum_{l=1}^{n} \frac{\partial^2 g_i(\xi_1, \dots, \xi_n)}{\partial x_j\, \partial x_l}(p_j^{(k)} - p_j)(p_l^{(k)} - p_l).$$

Legyen M olyan, hogy $\left|\frac{\partial^2 g_i(x_1,\dots,x_n)}{\partial x_j\, \partial x_l}\right| \le M$ minden $i, j, l = 1, \dots, n$-re a $\mathbf{p}$ pont egy olyan környezetében, melyben minden $\mathbf{p}^{(k)}$ benne van. M definícióját használva

$$|p_i^{(k+1)} - p_i| \le \frac{1}{2}\sum_{j=1}^{n}\sum_{l=1}^{n} M|p_j^{(k)} - p_j||p_l^{(k)} - p_l| \le \frac{n^2}{2}M\|\mathbf{p}^{(k)} - \mathbf{p}\|_\infty^2.$$

Mivel ez a becslés minden $i = 1, \dots, n$-re teljesül, ezért

$$\|\mathbf{p}^{(k+1)} - \mathbf{p}\|_\infty \leq \frac{n^2}{2} M \|\mathbf{p}^{(k)} - \mathbf{p}\|_\infty^2,$$

azaz a konvergencia másodrendű. □

Feladatok

1. Alakítsa át a következő egyenleteket fixpont feladattá, majd keresse meg az egyenlet közelítő megoldását fixpont iterációval a $(0,0)^T$ kezdeti értékből kiindulva:

(a)
$$\begin{aligned} -2x^2 + 6x - y^2 &= 4 \\ x^2 + y^3 - 5y &= 3 \end{aligned}$$

(b)
$$\begin{aligned} 8x + \cos x - y^3 &= -7 \\ x^2 + 4y &= 8 \end{aligned}$$

(c)
$$\begin{aligned} x^2 + 7x + y^2 - 4y &= 3 \\ 2x + y^3 + 4y &= -5 \end{aligned}$$

(d)
$$\begin{aligned} \cos x - 5y &= 3 \\ x^2 - 6x + y^2 - 2y &= 4 \end{aligned}$$

2. Számítsa ki az előző feladatban használt fixpont függvény deriváltját és annak normáját a numerikusan kapott fixpontban!
3. Mutassa meg, hogy a 2.55. tétel feltételei mellett a $\mathbf{p}^{(k)}$ fixpont iteráció tetszőleges vektornormában lokálisan kvadratikusan konvergál!

2.12. Newton-módszer n-dimenzióban

Legyen $U \subset \mathbb{R}^n$ nyílt halmaz, $\mathbf{f}\colon U \to \mathbb{R}^n$, és tekintsük az

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}$$

egyenletrendszert. Rögzítsünk egy $\mathbf{p}^{(k)} \in U$ vektort. Az egyváltozós esethez hasonlóan közelítsük az $\mathbf{f}$ függvényt a lineáris részével, az $\mathbf{f}(\mathbf{p}^{(k)}) + \mathbf{f}'(\mathbf{p}^{(k)})(\mathbf{x} - \mathbf{p}^{(k)})$ függvénnyel. Ennek gyöke az $\bar{\mathbf{x}} = \mathbf{p}^{(k)} - (\mathbf{f}'(\mathbf{p}^{(k)}))^{-1}\mathbf{f}(\mathbf{p}^{(k)})$ vektor. Ezt a képletet használjuk a Newton-módszer definíciójára:

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} - \left(\mathbf{f}'(\mathbf{p}^{(k)})\right)^{-1} \mathbf{f}(\mathbf{p}^{(k)}). \tag{2.29}$$

2.56. tétel. *Legyen $\mathbf{f} \in C^2$, $\mathbf{f}(\mathbf{p}) = \mathbf{0}$ és $\mathbf{f}'(\mathbf{p})$ invertálható. Ekkor a (2.29) Newton-iteráció lokálisan kvadratikusan konvergál $\mathbf{p}$-hez.*

Bizonyítás. A Newton-módszer egy fixpont iteráció a

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} - (\mathbf{f}'(\mathbf{x}))^{-1}\mathbf{f}(\mathbf{x})$$

iterációs függvénnyel. Legyen $(\mathbf{f}'(\mathbf{x}))^{-1} = (b_{ij}(\mathbf{x}))_{n\times n}$. Ekkor

$$\sum_{j=1}^{n} b_{ij}(\mathbf{x}) \frac{\partial f_j(\mathbf{x})}{\partial x_l} = \delta_{il} := \begin{cases} 1, & i = l, \\ 0, & i \neq l. \end{cases} \tag{2.30}$$

Tekintsük $\mathbf{g}$ i-edik komponensét:

$$g_i(\mathbf{x}) = x_i - \sum_{j=1}^{n} b_{ij}(\mathbf{x}) f_j(\mathbf{x}).$$

Ezt deriválva x_l szerint

$$\frac{\partial g_i(\mathbf{x})}{\partial x_l} = \delta_{il} - \sum_{j=1}^{n} \left(\frac{\partial b_{ij}(\mathbf{x})}{\partial x_l} f_j(\mathbf{x}) + b_{ij}(\mathbf{x}) \frac{\partial f_j(\mathbf{x})}{\partial x_l} \right).$$

Az $\mathbf{x} = \mathbf{p}$ pontban az $f_j(\mathbf{p}) = 0$ és a (2.30) relációkat használva tehát

$$\frac{\partial g_i(\mathbf{p})}{\partial x_l} = \delta_{il} - \sum_{j=1}^{n} b_{ij}(\mathbf{p}) \frac{\partial f_j(\mathbf{p})}{\partial x_l} = 0.$$

Azt kaptuk, hogy $\mathbf{g}'(\mathbf{p}) = \mathbf{0}$, és így a 2.55. tétel szerint a fixpont sorozat lokálisan kvadratikusan konvergens. □

A (2.29) képlet alkalmazásakor mátrixot kell invertálni. Ehelyett a gyakorlatban a következőképpen járunk el: Vezessük be az $\mathbf{s}^{(k)} = \mathbf{p}^{(k+1)} - \mathbf{p}^{(k)}$ jelölést, és rendezzük át a (2.29) egyenletet az

$$\mathbf{f}'(\mathbf{p}^{(k)})\mathbf{s}^{(k)} = -\mathbf{f}(\mathbf{p}^{(k)})$$

alakba. Ezt megoldjuk $\mathbf{s}^{(k)}$-ra, majd legyen $\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} + \mathbf{s}^{(k)}$.

2.57. példa. Tekintsük a 2.51. példában vizsgált (2.25) egyenletrendszert! A Newton-módszert alkalmaztuk az egyenletre a $(-1.5, -1.5)^T$ kezdeti értéktől indulva. A kapott eredményt a 2.13. táblázatban foglaltuk össze. □

2.13. táblázat. Newton-módszer

k	$\mathbf{p}^{(k)}$	$\|\mathbf{p}^{(k)} - \mathbf{p}\|_\infty$
0	(-1.50000000000,-1.50000000000)^T	2.500000e+00
1	(-1.25000000000,-0.52120413480)^T	2.250000e+00
2	(0.53188386800,-0.10035922100)^T	4.681161e-01
3	(0.98873605300,-0.00042581408)^T	1.126395e-02
4	(0.99999868610,-0.00000037764)^T	1.313900e-06

Feladatok

1. Alkalmazza a Newton-módszert a 2.11. szakasz 1. feladatában szereplő egyenletek megoldására!

2.13. Kvázi-Newton módszerek, Broyden-módszer

A Newton-módszert gyors (lokális) konvergenciája miatt szeretjük alkalmazni. Hátránya viszont, hogy az $\mathbf{f}$ derivált mátrixát kell kiszámolni hozzá, aminek általában nagy a műveletigénye. Ezenkívül mátrixot kell invertálni vagy lineáris egyenletrendszert megoldani minden egyes iterációban, ami szintén műveletigényes. Ezen nehézségek kiküszöbölésére szolgálnak a *kvázi-Newton módszerek*, amelynek általános definíciója:

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} - \left(\mathbf{A}^{(k)}\right)^{-1} \mathbf{f}(\mathbf{p}^{(k)}). \tag{2.31}$$

A kvázi-Newton módszereknél tehát az $\mathbf{f}'(\mathbf{p}^{(k)})$ mátrixot közelítjük egy $\mathbf{A}^{(k)}$ mátrixszal. Attól függően, hogy milyen közelítést használunk, más és más módszereket tudunk definiálni. Az egyik gyakran használt módszer a deriváltat numerikusan közelíti: legyen $\mathbf{e}^{(j)} = (0, \dots, 0, 1, 0, \dots, 0)^T$ a j-edik egységvektor, $h > 0$ egy megadott kis lépésköz, és definiáljuk az $\mathbf{A}^{(k)}$ mátrix komponenseit az

$$a_{ij}^{(k)} = \frac{f_i(\mathbf{p}^{(k)} + h\mathbf{e}^{(j)}) - f_i(\mathbf{p}^{(k)})}{h}, \qquad i, j = 1, \dots, n \tag{2.32}$$

képlettel.

A továbbiakban az $\mathbf{A}^{(k)}$ mátrix megválasztásának egy másik, a gyakorlatban igen népszerű módszerét, a *Broyden-módszert* vizsgáljuk. Ez a módszer is, mint az előző, a szelőmódszer általánosításának tekinthető.

Skaláris egyenletekre a szelőmódszer az $f(x) = 0$ egyenletet az

$$f(p_k) + a_k(x - p_k) = 0$$

lineáris egyenlettel helyettesíti, ahol $a_k = (f(p_k) - f(p_{k-1}))/(p_k - p_{k-1})$. Ezt k helyett $k+1$-re felírva és átrendezve, kapjuk, hogy a_{k+1} megoldása az

$$a_{k+1}(p_{k+1} - p_k) = f(p_{k+1}) - f(p_k) \tag{2.33}$$

egyenletnek. Ez utóbbi alakot lehet könnyen általánosítani többváltozós függvényekre.

Válasszunk egy $\mathbf{p}^{(0)}$ kezdeti vektort és egy $\mathbf{A}^{(0)}$ kezdeti mátrixot. $\mathbf{A}^{(0)}$ választására többféle módszer használatos: használhatjuk az $\mathbf{A}^{(0)} = \mathbf{f}'(\mathbf{p}^{(0)})$ pontos értéket, vagy a (2.32) képlettel közelíthetjük a derivált mátrixot a $\mathbf{p}^{(0)}$ pontban, vagy veszünk egy tetszőleges invertálható $\mathbf{A}^{(0)}$ mátrixot.

Tegyük fel, hogy $\mathbf{p}^{(k)}$ és $\mathbf{A}^{(k)}$ már definiált. Ekkor a (2.31) képlettel értelmezzük $\mathbf{p}^{(k+1)}$-et. A (2.33) egyenlet analógiájára megköveteljük, hogy $\mathbf{A}^{(k+1)}$ teljesítse az

$$\mathbf{A}^{(k+1)}(\mathbf{p}^{(k+1)} - \mathbf{p}^{(k)}) = \mathbf{f}(\mathbf{p}^{(k+1)}) - \mathbf{f}(\mathbf{p}^{(k)}), \tag{2.34}$$

az ún. *szelő egyenletet*. Vezessük be a következő jelöléseket:

$$\mathbf{y}^{(k)} := \mathbf{f}(\mathbf{p}^{(k+1)}) - \mathbf{f}(\mathbf{p}^{(k)}) \qquad \text{és} \qquad \mathbf{s}^{(k)} := \mathbf{p}^{(k+1)} - \mathbf{p}^{(k)}.$$

Ezzel a jelöléssel a (2.31) iterációs formula az

$$\mathbf{A}^{(k)}\mathbf{s}^{(k)} = -\mathbf{f}(\mathbf{p}^{(k)}), \tag{2.35}$$

a (2.34) egyenlet pedig az

$$\mathbf{A}^{(k+1)}\mathbf{s}^{(k)} = \mathbf{y}^{(k)} \tag{2.36}$$

alakban írható fel. A (2.35) egyenlet megoldható $\mathbf{s}^{(k)}$-ra (feltéve hogy $\mathbf{A}^{(k)}$ invertálható), így a probléma redukálódott arra, hogy olyan $\mathbf{A}^{(k+1)}$ mátrixot keresünk, amely a (2.36) egyenletet teljesíti. Ez az egyenlet viszont nem határozza meg az $\mathbf{A}^{(k+1)}$ mátrixot egyértelmű módon, hiszen a vektor alakban írt egyenlet n db skaláris egyenlettel ekvivalens, $\mathbf{A}^{(k+1)}$-et viszont n^2 db komponense határozza meg. (2.36) csak annyit jelent, hogy az $\mathbf{A}^{(k+1)}$ mátrixszal meghatározott lineáris leképezés az $\mathbf{s}^{(k)}$ irányában meghatározott, de az erre merőleges $(n-1)$-dimenziós altéren nem meghatározott. Mivel a $k+1$-edik lépésben erről nincs új információnk, ezért úgy definiáljuk $\mathbf{A}^{(k+1)}$-et, hogy a mátrixhoz tartozó lineáris leképezésnek ugyanaz legyen a hatása ezen az altéren, mint az $\mathbf{A}^{(k)}$ leképezésnek. Azaz a (2.36) egyenleten kívül azt is megköveteljük, hogy

$$\mathbf{A}^{(k+1)}\mathbf{z} = \mathbf{A}^{(k)}\mathbf{z}, \qquad \text{minden } \mathbf{z} \perp \mathbf{s}^{(k)}\text{-ra}. \tag{2.37}$$

(2.36) és (2.37) együtt egyértelműen meghatározza az $\mathbf{A}^{(k+1)}$ mátrixot. Könnyen belátható (2. feladat), hogy az

$$\mathbf{A}^{(k+1)} = \mathbf{A}^{(k)} + \frac{(\mathbf{y}^{(k)} - \mathbf{A}^{(k)}\mathbf{s}^{(k)})(\mathbf{s}^{(k)})^T}{\|\mathbf{s}^{(k)}\|_2^2} \tag{2.38}$$

mátrix teljesíti a (2.36) és (2.37) egyenleteket.

A (2.31) rekurzív képletben igazából $(\mathbf{A}^{(k)})^{-1}$-re van szükségünk. Ennek kiszámítását teszi egyszerűbbé a következő tétel.

2.58. tétel (Sherman–Morrison–Woodbury). *Legyen $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, $\mathbf{u}, \mathbf{v} \neq \mathbf{0}$ és $\mathbf{A} \in \mathbb{R}^{n\times n}$ invertálható. Ekkor az $\mathbf{A} + \mathbf{u}\mathbf{v}^T$ mátrix akkor és csak akkor invertálható, ha $1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u} \neq 0$, és ekkor*

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}$$

teljesül.

Bizonyítás. Legyen $\gamma \in \mathbb{R}$, és tekintsük a következő szorzatot:

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^T)(\mathbf{A}^{-1} - \gamma\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}) = \mathbf{I} + \mathbf{u}\mathbf{v}^T\mathbf{A}^{-1} - \gamma\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1} - \gamma\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}.$$

Mivel $\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}$ skaláris szám, ezért az előző egyenlet átalakítható az

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^T)(\mathbf{A}^{-1} - \gamma\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}) = \mathbf{I} + (1 - \gamma - \gamma\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u})\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}$$

alakba, amiből következik az állítás. □

A 2.58. tételt használva a (2.38) összefüggésre, rövid számolással kapjuk:

$$\begin{aligned}
(\mathbf{A}^{(k+1)})^{-1} &= \left(\mathbf{A}^{(k)} + \frac{(\mathbf{y}^{(k)} - \mathbf{A}^{(k)}\mathbf{s}^{(k)})(\mathbf{s}^{(k)})^T}{\|\mathbf{s}^{(k)}\|_2^2}\right)^{-1} \\
&= (\mathbf{A}^{(k)})^{-1} - \frac{(\mathbf{A}^{(k)})^{-1}\left(\frac{\mathbf{y}^{(k)} - \mathbf{A}^{(k)}\mathbf{s}^{(k)}}{\|\mathbf{s}^{(k)}\|_2^2}\right)(\mathbf{s}^{(k)})^T(\mathbf{A}^{(k)})^{-1}}{1 + (\mathbf{s}^{(k)})^T(\mathbf{A}^{(k)})^{-1}\frac{\mathbf{y}^{(k)} - \mathbf{A}^{(k)}\mathbf{s}^{(k)}}{\|\mathbf{s}^{(k)}\|_2^2}} \\
&= (\mathbf{A}^{(k)})^{-1} - \frac{\left((\mathbf{A}^{(k)})^{-1}\mathbf{y}^{(k)} - \mathbf{s}^{(k)}\right)(\mathbf{s}^{(k)})^T(\mathbf{A}^{(k)})^{-1}}{(\mathbf{s}^{(k)})^T(\mathbf{A}^{(k)})^{-1}\mathbf{y}^{(k)}}.
\end{aligned} \tag{2.39}$$

Ismerve $(\mathbf{A}^{(k)})^{-1}$-t, csak mátrixszorzásokat alkalmazva kiszámítható $(\mathbf{A}^{(k+1)})^{-1}$, így ehhez csak n^2 nagyságrendű művelet kell, szemben azzal, hogy a mátrix invertáláshoz, mint azt majd a következő fejezetben megmutatjuk, n^3 nagyságrendű műveletre van szükség.

Megmutatható, hogy a Broyden-módszer lokálisan konvergál az $\mathbf{f}$ függvény egy $\mathbf{p}$ gyökéhez, és ha $\mathbf{A}^{(0)}$ elegendően közel van $\mathbf{f}'(\mathbf{p})$-hez, akkor a konvergencia rendje szuperlineáris, azaz

$$\lim_{k\to\infty} \frac{\|\mathbf{p}^{(k+1)} - \mathbf{p}\|}{\|\mathbf{p}^{(k)} - \mathbf{p}\|} = 0.$$

Ezek bizonyításával itt nem foglalkozunk. A Broyden-módszer egy lehetséges változatát a következő algoritmusban közöljük.

2.59. algoritmus. Broyden-módszer

INPUT: $\mathbf{f}$ - függvény,
$\mathbf{p}^{(0)}$ - kezdeti érték,
h - lépésköz $\mathbf{A}^{(0)}$ számításához,
$\|\cdot\|$ - vektornorma
TOL - tolerancia,
$MAXIT$ - maximális iterációszám,
OUTPUT: $\mathbf{p}$ - közelítő gyök.

($\mathbf{A} = (a_{ij}) = A^{(0)}$ *kiszámítása)*
for $i = 1, \ldots, n$ **do**
  for $j = 1, \ldots, n$ **do**
    $a_{ij} \leftarrow (f_i(\mathbf{p}^{(0)} + h\mathbf{e}^{(j)}) - f_i(\mathbf{p}^{(0)}))/h$
  end do
end do
$\mathbf{A} \leftarrow \mathbf{A}^{-1}$
$\mathbf{q} \leftarrow \mathbf{p}^{(0)}$
$k \leftarrow 1$ *(lépésszám)*
while $k < MAXIT$ **do**
  $\mathbf{s} \leftarrow -\mathbf{A}\mathbf{f}(\mathbf{q})$
  $\mathbf{p} \leftarrow \mathbf{q} + \mathbf{s}$
  if $\|\mathbf{s}\| < TOL$ **do**
    output$(\mathbf{p})$
    stop
  end do
  $\mathbf{y} \leftarrow \mathbf{f}(\mathbf{p}) - \mathbf{f}(\mathbf{q})$
  $\mathbf{A} \leftarrow \mathbf{A} - \dfrac{(\mathbf{A}\mathbf{y} - \mathbf{s})\mathbf{s}^T\mathbf{A}}{\mathbf{s}^T\mathbf{A}\mathbf{y}}$
  $\mathbf{q} \leftarrow \mathbf{p}$
  $k \leftarrow k + 1$
end do
output(Maximális iterációszám túllépve)

2.60. példa. Tekintsük újra a 2.51. és 2.57. példákban vizsgált (2.25) egyenletrendszert! A 2.59. algoritmus eredménye erre az egyenletre a 2.14. táblázatban látható. Az utolsó oszlop mutatja, hogy a módszer szuperlineárisan konvergált. □

Feladatok

1. Alkalmazza a Broyden-módszert a 2.11. szakasz 1. feladatában szereplő egyenletek megoldására!
2. Mutassa meg, hogy a (2.38) képlettel definiált $\mathbf{A}^{(k+1)}$ mátrix teljesíti a (2.36) és (2.37) egyenleteket!

2.14. táblázat. Broyden-módszer

k	$\mathbf{p}^{(k)}$	$\|\mathbf{p}^{(k)}-\mathbf{p}\|_\infty$	$\frac{\|\mathbf{p}^{(k)}-\mathbf{p}\|_\infty}{\|\mathbf{p}^{(k-1)}-\mathbf{p}\|_\infty}$
0	(-1.5000000000, -1.5000000000)^T	2.5000000000	
1	(-1.2490215360, -0.5215363883)^T	2.2490215360	0.8996086144
2	(-0.4968297655, -0.9366983828)^T	1.4968297660	0.6655471022
3	(-0.3045368940, -0.3621731989)^T	1.3045368940	0.8715332389
4	(0.5414891937, -0.0587408442)^T	0.4585108063	0.3514740046
5	(0.9527177435, -0.0515250779)^T	0.0515250779	0.1123748387
6	(1.0003263340, 0.0319681269)^T	0.0319681269	0.6204382061
7	(1.0000051000, -0.0040567750)^T	0.0040567750	0.1269006155
8	(1.0000069210, -0.0000347010)^T	0.0000347010	0.0085538489
9	(1.0000001100, 0.0000012682)^T	0.0000012682	0.0365458110
10	(1.0000000050, 0.0000000576)^T	0.0000000576	0.0453865979

3. fejezet
Lineáris egyenletrendszerek

Ebben a fejezetben lineáris egyenletrendszerek direkt módszerekkel történő numerikus megoldásait és vele kapcsolatos lineáris algebrai feladatokat vizsgálunk. Megismerjük a Gauss- és Gauss–Jordan-eliminációt és variánsait, valamint azok alkalmazását a mátrix inverzió feladatára.

3.1. Lineáris algebrai előismeretek

Ebben a szakaszban néhány, a későbbiekben használt lineáris algebrai jelölést, fogalmat, állítást elevenítünk fel. A továbbiakban, ha másképp nem mondjuk, $\mathbf{A} = (a_{ij})$ egy $n \times n$-es mátrixot, $\mathbf{x}$ pedig egy n-dimenziós oszlopvektort jelöl. Az $\mathbf{A}$ mátrix determinánsát $\det(\mathbf{A})$-val, az $n \times n$-es egységmátrixot $\mathbf{I}$-vel jelöljük. Az $\mathbf{A}$ mátrix ill. az $\mathbf{x}$ oszlopvektor transzponáltját $\mathbf{A}^T$ ill. $\mathbf{x}^T$ jelöli. Azt a diagonális mátrixot, amelynek főátlójában rendre $a_1, a_2, \ldots, a_n$ áll, $\operatorname{diag}(a_1, a_2, \ldots, a_n)$ jelöli.

A determinánsok néhány ismert tulajdonságát foglaltuk össze a következő tételben:

3.1. tétel. *Legyen $\mathbf{A}, \mathbf{B}$ $n \times n$-es mátrixok. Ekkor*

1. $\det(\mathbf{A}) = 0$, *ha* $\mathbf{A}$ *egy sora (vagy oszlopa) azonosan nulla;*
2. $\det(\mathbf{A}) = 0$, *ha* $\mathbf{A}$ *két sora (oszlopa) azonos;*
3. $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$;
4. $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})$;
5. $\det(\mathbf{A}^T) = \det(\mathbf{A})$;
6. *Ha* $\mathbf{B}$*-t úgy kapjuk az* $\mathbf{A}$ *mártixból, hogy annak valamely sorát (oszlopát) megszorozzuk egy c konstanssal, akkor* $\det(\mathbf{B}) = c \det(\mathbf{A})$.
7. *Ha* $\mathbf{B}$*-t úgy kapjuk az* $\mathbf{A}$ *mártixból, hogy annak két sorát (oszlopát) felcseréljük, akkor* $\det(\mathbf{B}) = -\det(\mathbf{A})$.
8. *Ha* $\mathbf{B}$*-t úgy kapjuk az* $\mathbf{A}$ *mártixból, hogy annak egyik sorához (oszlopához) egy másik sor (oszlop) c-szeresét ($c \in \mathbb{R}$ tetszőleges) hozzáadjuk, akkor* $\det(\mathbf{B}) = \det(\mathbf{A})$.
9. *Jelölje* $\mathbf{A}_{ij}$ *azt az $(n-1) \times (n-1)$-es mátrixot, amelyet az* $\mathbf{A}$ *mátrixból annak i-edik sora és j-edik oszlopa elhagyásával kapunk. Ekkor a determináns i-edik sora szerinti sorfejtése*

$$\det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{A}_{ij}),$$

a j-edik oszlop szerinti sorfejtése pedig

$$\det(\mathbf{A}) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{A}_{ij}).$$

Az $\mathbf{A}^{-1}$ $n \times n$-es mátrixot az $\mathbf{A}$ $n \times n$-es mátrix *inverzének* nevezzük, ha $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. Egy négyzetes mátrixot *invertálhatónak* nevezünk, ha létezik az inverze. Egy $\mathbf{A}$ négyzetes mátrixot *szinguláris*nak nevezünk, ha nem létezik az inverze. Az invertálható mátrixokat szokás *nemszinguláris* vagy *reguláris* mátrixoknak is hívni.

3.2. tétel. *Legyen* $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$. *A következő állítások ekvivalensek:*

1. $\det(\mathbf{A}) \neq 0$,
2. *az* $\mathbf{A}$ *mátrix invertálható,*
3. *az* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *egyenletnek létezik egyértelmű megoldása minden* $\mathbf{b}$ *vektorra.*

3.3. tétel. *Az* $\mathbf{A}\mathbf{x} = \mathbf{0}$ *egyenletnek akkor és csak akkor van nemtriviális (azaz nemnulla) megoldása, ha* $\mathbf{A}$ *szinguláris, azaz* $\det(\mathbf{A}) = 0$.

3.4. tétel. *Ha* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ *invertálható, akkor* $\mathbf{AB}$ *is invertálható, és* $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Az $\mathbf{A}$ négyzetes mátrixot *felülről (alulról) triangulárisnak* vagy más szóval *felső (alsó) háromszög mátrixnak* nevezzünk, ha $a_{ij} = 0$ minden $i > j$-re ($i < j$-re), azaz a mátrix főátlója alatti (feletti) minden elem 0.

3.5. tétel. *Egy* $\mathbf{A}$ *trianguláris mátrix deteminánsa* $det(\mathbf{A}) = a_{11}a_{22}\cdots a_{nn}$.

3.6. tétel. *Felülről (alulról) trianguláris mátrixok szorzata felülről (alulról) trianguláris mátrix. Felülről (alulról) trianguláris invertálható mátrix inverze felülről (alulról) trianguláris mátrix.*

Egy olyan $\mathbf{P}$ négyzetes mátrixot, amelyet az egységmátrixból sorok (vagy oszlopok) felcserélésével (permutációjával) kapunk, *permutációs mátrixnak* mátrixnak nevezünk. A következő tétel szerint mátrixok sorainak (oszlopainak) felcserélése egy megfelelő permutációs mátrixszal való szorzással ekvivalens.

3.7. tétel. *Legyen* $k_1, \ldots, k_n$ *az* $1, \ldots, n$ *számok egy permutációja (átrendezése), és legyen* $\mathbf{P} \in \mathbb{R}^{n\times n}$ *az a permutációs mátrix, amelyet az egységmátrixból úgy kapunk, hogy annak első sorát a* k_1*-edik sorba, ..., az* n*-edik sorát pedig a* k_n*-edik sorba helyezzük el. Legyen* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *tetszőleges. Ekkor a* $\mathbf{PA}$ *mátrix (*$\mathbf{AP}$ *mátrix) megkapható az* $\mathbf{A}$ *mátrixból úgy, hogy annak első sorát (oszlopát) a* k_1*-edik sorba (oszlopba), ..., az* n*-edik sorát (oszlopát) pedig a* k_n*-edik sorba (oszlopba) helyezzük el.*

Az $\mathbf{A}$ négyzetes mátrixot *soronként diagonálisan dominánsnak* vagy röviden *diagonálisan dominánsnak* nevezzük, ha

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|$$

teljesül minden $i = 1, \ldots, n$-re. Ehhez hasonlóan az $\mathbf{A}$ mátrixot *oszloponként diagonálisan dominánsnak* nevezzük, ha $\mathbf{A}^T$ diagonálisan domináns, azaz

$$|a_{jj}| > \sum_{\substack{i=1\\ i\neq j}}^{n} |a_{ij}|$$

teljesül minden $j = 1, \ldots, n$-re.

3.8. tétel. *Ha* $\mathbf{A}$ *diagonálisan domináns, akkor* $\mathbf{A}$ *invertálható.*

Bizonyítás. Tegyük fel, hogy $\mathbf{A}$ nem invertálható. Ekkor az $\mathbf{Ax} = \mathbf{0}$ egyenletnek létezik $\mathbf{x} \neq \mathbf{0}$ nemtriviális megoldása. Legyen k olyan, hogy $|x_k| = \max\{|x_i| : i = 1, \ldots, n\}$. Ekkor $x_k \neq 0$. Mivel $\sum_{j=1}^{n} a_{ij}x_j = 0$ minden $i = 1, \ldots, n$-re, kapjuk, hogy $a_{kk}x_k = -\sum_{j=1, j\neq k}^{n} a_{kj}x_j$. Ekkor a háromszög-egyenlőtlenség alapján $|a_{kk}x_k| \leq \sum_{j=1, j\neq k}^{n} |a_{kj}x_j|$, és így

$$|a_{kk}| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}| \frac{|x_j|}{|x_k|} \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}|,$$

ami ellentmondás. □

Egy $\mathbf{A}$ mátrixot *pozitív definitnek* (*negatív definitnek*) nevezünk, ha $\mathbf{A}$ szimmetrikus és $\mathbf{x}^T\mathbf{Ax} > 0$ (ill. $\mathbf{x}^T\mathbf{Ax} < 0$) minden $\mathbf{x} \neq \mathbf{0}$-ra. $\mathbf{A}$-t *pozitív szemidefinitnek* (*negatív szemidefinitnek*) nevezzük, ha $\mathbf{A}$ szimmetrikus és $\mathbf{x}^T\mathbf{Ax} \geq 0$ (ill. $\mathbf{x}^T\mathbf{Ax} \leq 0$) minden $\mathbf{x}$-re.

3.9. tétel. *Ha* $\mathbf{A}$ *pozitív definit, akkor*

1. $\mathbf{A}$ *invertálható,*
2. $a_{ii} > 0$ *minden* $i = 1, \ldots, n$*-re.*

3.10. tétel. *Az* $\mathbf{A}$ *négyzetes szimmetrikus mátrix akkor és csak akkor pozitív definit, ha az összes bal felső főminorjai pozitívak, azaz*

$$\det \begin{pmatrix} a_{11} & \cdots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{ii} \end{pmatrix} > 0, \qquad i = 1, 2, \ldots, n.$$

Az $\mathbf{A}$ négyzetes mátrixot ortogonálisnak nevezzük, ha $\mathbf{AA}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}$, azaz $\mathbf{A}$ invertálható és $\mathbf{A}^{-1} = \mathbf{A}^T$.

3.11. tétel. *Ortogonális mátrixok szorzata ortogonális.*

A $\lambda \in \mathbb{C}$ komplex számot az $\mathbf{A}$ mátrix *sajátértékének* nevezzük, ha az

$$\mathbf{Ax} = \lambda\mathbf{x}$$

egyenletnek létezik nemtriviális ($\mathbf{x} \neq \mathbf{0}$) megoldása. Az egyenlet egy $\mathbf{x} \neq 0$ megoldását az $\mathbf{A}$ mátrix λ sajátértékéhez tartozó *sajátvektorának* nevezzük.

3.12. tétel. *Az* $\mathbf{A}$ $n \times n$*-es mátrixnak* n *db sajátértéke van, amelyek a*

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

n*-edfokú algebrai egyenlet, az ún.* karakterisztikus egyenlet *gyökei.*

3.13. tétel. *Legyen* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *az* $\mathbf{A}$ *mátrix sajátértékei. Ekkor*

1. *$det(\mathbf{A}) = \lambda_1\lambda_2\cdots\lambda_n$;*
2. *$\mathbf{A}$ akkor és csak akkor invertálható, ha $\lambda_i \neq 0$ minden $i = 1, 2, \ldots, n$-re;*
3. *ha $\mathbf{A}$ invertálható, akkor $\mathbf{A}^{-1}$ sajátértékei az $1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_n$ számok;*
4. *az $\mathbf{A}^k$ mátrix sajátértékei a $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$ számok.*

3.14. tétel. *Egy trianguláris $\mathbf{A}$ mátrix sajátértékei a főátlóban álló a_{11}, a_{22}, …, a_{nn} számok.*

Legyen $\mathbf{A}$ és $\mathbf{B}$ két azonos dimenziójú négyzetes mátrix. Azt mondjuk, hogy $\mathbf{A}$ és $\mathbf{B}$ *hasonló*, ha létezik olyan $\mathbf{P}$ invertálható mátrix, hogy $\mathbf{A} = \mathbf{P}^{-1}\mathbf{B}\mathbf{P}$. Megjegyezzük, hogy ekkor nyilván $\mathbf{B} = \mathbf{P}\mathbf{A}\mathbf{P}^{-1}$, azaz a hasonlóság szimmetrikus reláció. A $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ mátrixhoz tartozó lineáris transzformációt *hasonlósági transzformációnak* nevezzük.

3.15. tétel. *Hasonló mátrixok sajátértékei megegyeznek.*

Bizonyítás. Legyen $\mathbf{A} = \mathbf{P}^{-1}\mathbf{B}\mathbf{P}$. Ekkor a determinánsok tulajdonságait felhasználva $\mathbf{A}$ karakterisztikus polinomjára

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \det(\mathbf{P}^{-1}\mathbf{B}\mathbf{P} - \lambda\mathbf{I}) = \det(\mathbf{P}^{-1})\det(\mathbf{B} - \lambda\mathbf{I})\det(\mathbf{P}) = \det(\mathbf{B} - \lambda\mathbf{I})$$

teljesül, amiből következik a tétel. □

A $\rho(\mathbf{A}) := \max\{|\lambda|\colon \lambda \text{ sajátértéke } \mathbf{A}\text{-nak}\}$ számot az $\mathbf{A}$ mátrix *spektrálsugarának* nevezzük.

3.16. tétel. *Legyen k pozitív egész, és $\|\cdot\|$ egy tetszőleges mátrixnorma. Ekkor*

1. *$\rho(\mathbf{A}^k) = (\rho(\mathbf{A}))^k$,*
2. *$\rho(\mathbf{A}) \leq \|\mathbf{A}\|$.*

3.17. tétel. *Minden $\mathbf{A}$ mátrixhoz és $\varepsilon > 0$ számhoz létezik olyan $\|\cdot\|$ mátrixnorma, amelyre $\|\mathbf{A}\| \leq \rho(\mathbf{A}) + \varepsilon$.*

3.18. tétel. *Egy tetszőleges négyzetes $\mathbf{A}$ mátrixra $\|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^T\mathbf{A})}$. Ha $\mathbf{A}$ szimmetrikus, akkor $\|\mathbf{A}\|_2 = \rho(\mathbf{A})$.*

Legyenek $a_1, \ldots, a_n$ komplex számok. A

$$\det\begin{pmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^{n-1} \end{pmatrix} \tag{3.1}$$

determinánst *Vandermonde-féle determinánsnak* nevezzük.

3.19. tétel. *A (3.1) Vandermonde-féle determináns akkor és csak akkor nem nulla, ha az a_i számok páronként különbözők.*

Feladatok

1. Határozza meg az α és β paraméterek lehetséges értékeit, hogy az
$$\mathbf{A} = \begin{pmatrix} \alpha & 1 & 0 \\ \beta & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$
mátrix
   (a) szinguláris,
   (b) diagonálisan domináns,
   (c) szimmetrikus,
   (d) pozitív definit legyen.
2. Igazolja, hogy ha $\mathbf{A}$ és $\mathbf{B}$ pozitív definit $n \times n$-es mátrixok, akkor
   (a) $\mathbf{A}^T$,
   (b) $\mathbf{A} + \mathbf{B}$,
   (c) $\mathbf{A}^2$
   is pozitív definit.
3. Bizonyítsa be a 3.6. tételt!
4. Bizonyítsa be a 3.7. tételt!
5. Bizonyítsa be a 3.9. tételt!
6. Bizonyítsa be a 3.11. tételt!
7. Bizonyítsa be a 3.12. tételt!
8. Bizonyítsa be a 3.14. tételt!
9. Bizonyítsa be a 3.19. tételt! (Útmutatás: A 3.1. determináns képletében helyettesítsük a_1-et x-szel. Mutassa meg, hogy a kapott determináns egy $n-1$-edfokú polinom x-ben! Soroljon fel $n-1$ db különböző gyökét a kapott polinomnak!)
10. Mutassa meg, hogy a 3.1. Vandermonde-determináns értéke
$$\prod_{i>j}(a_i - a_j).$$
(Útmutatás: Tekintse az előző feladat megoldását!)

3.2. Trianguláris egyenletrendszerek

3.20. példa. Oldjuk meg a következő egyenletrendszert:

$$\begin{array}{rcrcrcrcr} 2x_1 & - & x_2 & + & 3x_3 & + & x_4 & = & 3 \\ & & 3x_2 & - & x_3 & + & 2x_4 & = & 13 \\ & & & & 2x_3 & - & x_4 & = & -2 \\ & & & & & & 3x_4 & = & 12 \end{array}$$

A negyedik egyenletet x_4-re megoldhatjuk: $x_4 = 4$. Ezt visszahelyettesítve a harmadik egyenletbe kapjuk $x_3 = (-2 + x_4)/2 = 1$, majd a második egyenletből $x_2 = (13 + x_3 - 2x_4)/3 = 2$. Végül az első egyenletből $x_1 = (3 + x_2 - 3x_3 - x_4)/2 = -1$. □

Az előző példát általánosítva, egy n-dimenziós felülről trianguláris egyenletrendszer, $\mathbf{Ax} = \mathbf{b}$, azaz

$$\begin{array}{rcrcccrcl} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ & & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ & & & & \ddots & & \vdots & & \vdots \\ & & & & & & a_{nn}x_n & = & b_n \end{array} \tag{3.2}$$

megoldásának módszerét, az ún. *visszahelyettesítés módszerét* a következő algoritmussal adhatjuk meg:

3.21. algoritmus. Trianguláris egyenletrendszer megoldása visszahelyettesítéssel

INPUT: $a_{ij}, \quad (i = 1, \ldots, n, \quad j = 1, \ldots, n), \quad b_i, \quad (i = 1, \ldots, n)$
OUTPUT: $x_1, \ldots, x_n$

$x_n \leftarrow b_n / a_{nn}$
for $i = n-1, \ldots, 1$ **do**

$$x_i \leftarrow (b_i - \sum_{j=i+1}^{n} a_{ij} x_j) / a_{ii}$$

end do
output$(x_1, x_2, \ldots, x_n)$

A visszahelyettesítés módszere akkor és csak akkor hajtható végre, ha $a_{ii} \neq 0$ minden $i = 1, \ldots, n$-re. Mivel $\det(\mathbf{A}) = a_{11} a_{22} \cdots a_{nn}$, így ez akkor és csak akkor teljesül, ha a (3.2) egyenletnek létezik egyértelmű megoldása, azaz $\det(\mathbf{A}) \neq 0$.

A módszer műveletigénye:

	osztás/szorzás	összeadás/kivonás
1. lépés:	1	0
2. lépés:	2	1
$\vdots$	$\vdots$	$\vdots$
n. lépés:	n	$n-1$

Azaz a módszer végrehajtásához összesen $1 + 2 + \cdots + n = n(n+1)/2$ osztás ill. szorzás, valamint $1 + 2 + \cdots + n - 1 = (n-1)n/2$ összeadás ill. kivonás szükséges. Ezt szokás úgy is írni, hogy $n^2/2 + \mathcal{O}(n)$ nagyságrendű osztás/szorzás, és hasonlóan $n^2/2 + \mathcal{O}(n)$ nagyságrendű összeadás/kivonás kell a módszerhez. Itt és a továbbiakban $\mathcal{O}(n^k)$ egy legfeljebb k-adrendű polinomot jelöl.

Feladatok

1. Oldja meg a következő trianguláris egyenletrendszereket:

(a)

$$\begin{array}{rcrcrcrcr} 3x_1 & + & x_2 & - & x_3 & + & 2x_4 & = & -4 \\ & & 4x_2 & - & 2x_3 & + & x_4 & = & 5 \\ & & & & 6x_3 & - & 2x_4 & = & -7 \\ & & & & & & 2x_4 & = & 4 \end{array}$$

(b)

$$\begin{array}{rcrcrcrcr} 1.2x_1 & + & 2.1x_2 & - & 3.2x_3 & + & 2.0x_4 & + & 1.4x_5 & = & 81.5 \\ & & 2.5x_2 & - & 1.1x_3 & + & 6.1x_4 & - & 3.0x_5 & = & 159.7 \\ & & & & 2.6x_3 & - & 1.1x_4 & & & = & 12.8 \\ & & & & & & 2.2x_4 & + & 4.1x_5 & = & 46.9 \\ & & & & & & & & 1.3x_5 & = & 6.5 \end{array}$$

3.3. Gauss-elimináció, főelemkiválasztási stratégiák

3.22. példa. Tekintsük az

$$\begin{array}{rcrcrcrcr} x_1 & - & 2x_2 & - & 2x_3 & - & 2x_4 & = & -11 \\ 2x_1 & - & x_2 & + & 2x_3 & + & 4x_4 & = & -8 \\ -x_1 & + & 2x_2 & + & 3x_3 & - & 4x_4 & = & 27 \\ 2x_1 & + & x_2 & + & 4x_3 & - & 2x_4 & = & 28 \end{array} \tag{3.3}$$

egyenletrendszert. Az első egyenlet segítségével a második, harmadik és negyedik egyenletből az x_1 változó kiejthető a következő módon: az első egyenlet 2-szeresét, -1-szeresét, ill. 2-szeresét kivonjuk a második, harmadik, ill. a negyedik egyenletből:

$$\begin{array}{rcrcrcrcr} x_1 & - & 2x_2 & - & 2x_3 & - & 2x_4 & = & -11 \\ & & 3x_2 & + & 6x_3 & + & 8x_4 & = & 14 \\ & & & & x_3 & - & 6x_4 & = & 16 \\ & - & 3x_2 & & & - & 6x_4 & = & 6 \end{array} \tag{3.4}$$

Ekkor az eredetivel ekvivalens egyenletrendszert kapunk. Ezt mátrixok segítségével a következőképpen írhatjuk le röviden: A (3.3) egyenletrendszer együtthatóit egy 4×4-es mátrixban leírjuk, majd azt kibővítjük egy ötödik oszloppal, ahol az egyenletrendszer jobb oldalát írjuk le. Ekkor kapjuk az

$$\begin{pmatrix} 1 & -2 & -2 & -2 & -11 \\ 2 & -1 & 2 & 4 & -8 \\ -1 & 2 & 3 & -4 & 27 \\ 2 & 1 & 4 & -2 & 28 \end{pmatrix} \tag{3.5}$$

ún. *kibővített mátrixot.* A (3.4) egyenletrendszert leíró kibővített mátrixot tehát úgy kapjuk, hogy a (3.5) mátrix első sorát megszorozzuk 2, -1 és 2-vel, és a kapott sorokat kivonjuk rendre a második, harmadik és a negyedik sorból:

$$\begin{pmatrix} 1 & -2 & -2 & -2 & -11 \\ 0 & 3 & 6 & 8 & 14 \\ 0 & 0 & 1 & -6 & 16 \\ 0 & -3 & 0 & -6 & 6. \end{pmatrix}. \tag{3.6}$$

Az x_2 változó hiányzik a harmadik sorból, és a második egyenlet segítségével kiküszöböljük a a negyedik sorból x_2-t, azaz a (3.6) mátrixban a második oszlopban a főátló alatti elemeket „kinullázzuk” a második sor segítségével: a második sor -1-szeresét kivonjuk a negyedik sorból:

$$\begin{pmatrix} 1 & -2 & -2 & -2 & -11 \\ 0 & 3 & 6 & 8 & 14 \\ 0 & 0 & 1 & -6 & 16 \\ 0 & 0 & 6 & 2 & 20 \end{pmatrix}. \tag{3.7}$$

Végül beszorozzuk a harmadik sort 6-tal, és kivonjuk a negyedikből:

$$\begin{pmatrix} 1 & -2 & -2 & -2 & -11 \\ 0 & 3 & 6 & 8 & 14 \\ 0 & 0 & 1 & -6 & 16 \\ 0 & 0 & 0 & 38 & -76 \end{pmatrix}. \tag{3.8}$$

Ez az

$$\begin{array}{rcrcrcrcr} x_1 & - & 2x_2 & - & 2x_3 & - & 2x_4 & = & -11 \\ & & 3x_2 & + & 6x_3 & + & 8x_4 & = & 14 \\ & & & & x_3 & - & 6x_4 & = & 16 \\ & & & & & & 38x_4 & = & -76 \end{array}$$

triangularis egyenletrendszerrel ekvivalens. Ezt megoldva a visszahelyettesítés módszerével kapjuk, hogy a megoldás $x_1 = -3$, $x_2 = 2$, $x_3 = 4$ és $x_4 = -2$. A kibővített mátrixokkal a számolást röviden a következő alakban szoktuk leírni:

$$\begin{pmatrix} 1 & -2 & -2 & -2 & -11 \\ 2 & -1 & 2 & 4 & -8 \\ -1 & 2 & 3 & -4 & 27 \\ -2 & 1 & 4 & -2 & 28 \end{pmatrix} \sim \begin{pmatrix} 1 & -2 & -2 & -2 & -11 \\ 0 & 3 & 6 & 8 & 14 \\ 0 & 0 & 1 & -6 & 16 \\ 0 & -3 & 0 & -6 & 6. \end{pmatrix} \sim$$

$$\begin{pmatrix} 1 & -2 & -2 & -2 & -11 \\ 0 & 3 & 6 & 8 & 14 \\ 0 & 0 & 1 & -6 & 16 \\ 0 & 0 & 6 & 2 & 20 \end{pmatrix} \sim \begin{pmatrix} 1 & -2 & -2 & -2 & -11 \\ 0 & 3 & 6 & 8 & 14 \\ 0 & 0 & 1 & -6 & 16 \\ 0 & 0 & 0 & 38 & -76 \end{pmatrix}.$$

□

Az előző példa módszerét alkalmazva az

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \dots & + & a_{nn}x_n & = & b_n \end{array} \tag{3.9}$$

általános n-dimenziós lineáris egyenletrendszerre kapjuk a *Gauss-elimináció* módszerét: Az együtthatókat és az egyenlet bal oldalát az ún. *kibővített mátrixban* tároljuk:

$$\tilde{\mathbf{A}}^{(0)} = (\mathbf{A}, \mathbf{b}) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & a_{1,n+1} \\ a_{21} & a_{22} & \dots & a_{2n} & a_{2,n+1} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} & a_{n,n+1} \end{pmatrix},$$

ahol $a_{i,n+1} := b_i$, $(i = 1, \dots, n)$. Az $\tilde{\mathbf{A}}^{(0)}$ mátrixból képezzük az egymással ekvivalens egyenleteket leíró $\tilde{\mathbf{A}}^{(1)}$, $\tilde{\mathbf{A}}^{(2)}$, …, $\tilde{\mathbf{A}}^{(n-1)}$ mátrixokat a következő módon:

$$\tilde{\mathbf{A}}^{(1)} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & a_{1,n+1} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} & a_{2,n+1}^{(1)} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} & a_{n,n+1}^{(1)} \end{pmatrix},$$

ahol $a_{ij}^{(1)} = a_{ij} - l_{i1}a_{1j}, \quad l_{i1} = \frac{a_{i1}}{a_{11}}, \quad i = 2, \dots, n, \quad j = 2, \dots, n+1$, (feltéve, hogy $a_{11} \neq 0$). Ha már $\tilde{\mathbf{A}}^{(1)}, \dots, \tilde{\mathbf{A}}^{(k-1)}$ definiált, ahol $k \leq n-1$, akkor legyen

$$\tilde{\mathbf{A}}^{(k)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,k} & a_{1,k+1} & \cdots & a_{1,n} & a_{1,n+1} \\ 0 & a_{22}^{(1)} & \cdots & a_{2,k}^{(1)} & a_{2,k+1}^{(1)} & \cdots & a_{2,n}^{(1)} & a_{2,n+1}^{(1)} \\ & & \ddots & & & & & \\ 0 & 0 & \cdots & a_{k,k}^{(k-1)} & a_{k,k+1}^{(k-1)} & \cdots & a_{k,n}^{(k-1)} & a_{k,n+1}^{(k-1)} \\ 0 & 0 & \cdots & 0 & a_{k+1,k+1}^{(k)} & \cdots & a_{k+1,n}^{(k)} & a_{k+1,n+1}^{(k)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n,k+1}^{(k)} & \cdots & a_{n,n}^{(k)} & a_{n,n+1}^{(k)} \end{pmatrix},$$

ahol $a_{ij}^{(k)} = a_{ij}^{(k-1)} - l_{ik}a_{kj}^{(k-1)}, \;\; l_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$, $i = k+1, \dots, n, \;\; j = k+1, \dots, n+1$. Ezeket az ún. *eliminációs lépéseket* $k = 1, \dots, n-1$-re hajtjuk végre. Ezután az $\tilde{\mathbf{A}}^{(n-1)}$ mátrixhoz tartozó trianguláris egyenletrendszert a visszahelyettesítés módszerével megoldjuk.

A Gauss-elimináció végrehajtása után az együtthatómátrix főátlójában szereplő a_{11}, $a_{22}^{(1)}$, …, $a_{nn}^{(n-1)}$ számokat *főelemeknek* nevezzük. Nyilvánvalóan, a Gauss-elimináció akkor és csak akkor hajtható végre, ha az összes főelem nem nulla.

Ha a Gauss-elimináció lépéseit csak az együtthatómátrixon végezzük, akkor az iterációs lépésekben kapott mátrixokat $\mathbf{A}^{(0)} := \mathbf{A}$, $\mathbf{A}^{(1)}$, …, $\mathbf{A}^{(n-1)}$-gyel jelöljük.

3.23. algoritmus. Gauss-elimináció

INPUT: a_{ij}, $(i = 1, \ldots, n, \quad j = 1, \ldots, n+1)$ - kibővített együtthatómátrix
OUTPUT: x_1, ..., x_n

(elimináció:)
for $k = 1, \ldots, n-1$ **do**
  for $i = k+1, \ldots, n$ **do**
    $l_{ik} \leftarrow a_{ik}/a_{kk}$
    for $j = k+1, \ldots, n+1$ **do**
      $a_{ij} \leftarrow a_{ij} - l_{ik}a_{kj}$
    end do
  end do
end do
(visszahelyettesítés:)
$x_n \leftarrow a_{n,n+1}/a_{nn}$
for $i = n-1, \ldots, 1$ **do**
  $x_i \leftarrow (a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii}$
end do
output$(x_1, x_2, \ldots, x_n)$

A fenti algoritmust úgy fogalmaztuk meg, hogy minden egyes eliminációs lépésben az új együtthatómátrix elemeivel felülírjuk az előző lépés együtthatómátrixát. Megjegyezzük, hogy a 3.23. algoritmus a „kinullázott" elemeket se nem számítja, se nem tárolja. Azaz az algoritmus végén a főátló alatti elemek tartalma nem használható, ott előző lépésekből megmaradt tartalom van csak. Ha szükséges, ezeket az elemeket nullázzuk ki direkt módon.

A Gauss-elimináció műveletigénye:

	osztás/szorzás	összeadás/kivonás
1. lépés:	$(n-1)(n+1)$	$(n-1)n$
2. lépés:	$(n-2)n$	$(n-2)(n-1)$
$\vdots$	$\vdots$	$\vdots$
$n-1$-edik lépés:	$1 \cdot 3$	$1 \cdot 2$
összesen:	$\sum_{i=1}^{n-1} i(i+2)$	$\sum_{i=1}^{n-1} i(i+1)$

Az $1^2 + 2^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1)$ azonosságot alkalmazva könnyen kiszámítható, hogy összesen $n^3/3 + n^2/2 - 5n/6$ szorzás ill. osztás, valamint $(n^3 - n)/3$ összeadás ill. kivonás szükséges az együtthatómátrix trianguláris alakra hozásához. A visszahelyettesítéssel együtt pedig összesen $n^3/3 + n^2/2 - 5n/6 + n^2/2 + n/2 = n^3/3 + n^2 - n/3 = n^3/3 + \mathcal{O}(n^2)$ osztás ill. szorzás, valamint $(n^3 - n)/3 + n^2/2 - n/2 = n^3/3 + n^2/2 - 5n/6 = n^3/3 + \mathcal{O}(n^2)$ összeadás ill. kivonás szükséges a Gauss-elimináció végrehajtásához. Röviden azt mondjuk, hogy $n^3/3$ nagyságrendű műveletigénye van a módszernek.

3.24. példa. Oldjuk meg az

$$\begin{array}{rcrcrcrcr} 2x_1 & - & x_2 & & & - & 3x_4 & = & 8 \\ 2x_1 & - & x_2 & + & x_3 & + & 5x_4 & = & 2 \\ -3x_1 & + & x_2 & + & x_3 & - & 2x_4 & = & -5 \\ 2x_1 & + & 4x_2 & & & - & x_4 & = & 21 \end{array}$$

egyenletrendszert Gauss-eliminációval! Egy Gauss-eliminációs lépést elvégezve kapjuk

$$\begin{pmatrix} 2 & -1 & 0 & -3 & 8 \\ 2 & -1 & 1 & 5 & 2 \\ -3 & 1 & 1 & -2 & -5 \\ 2 & 4 & 0 & -1 & 21 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 0 & -3 & 8 \\ 0 & 0 & 1 & 8 & -6 \\ 0 & -1/2 & 1 & -13/2 & 7 \\ 0 & 5 & 0 & 2 & 13 \end{pmatrix}.$$

A második sor második oszlopában levő elem 0, ezért nem tudjuk tovább folyatatni a 3.23. algoritmust. Könnyen látható, hogy az egyenletrendszernek viszont létezik egyértelmű megoldása: $x_1 = 4$, $x_2 = 3$, $x_3 = 2$ és $x_4 = -1$. Ha felcseréljük az utolsó lépésben kapott kibővített mátrix második és harmadik sorát, akkor ezzel természetesen a hozzá tartozó egyenletrendszer nem változik, viszont folytathatók az eliminációs lépések:

$$\begin{pmatrix} 2 & -1 & 0 & -3 & 8 \\ 0 & 0 & 1 & 8 & -6 \\ 0 & -1/2 & 1 & -13/2 & 7 \\ 0 & 5 & 0 & 2 & 13 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 0 & -3 & 8 \\ 0 & -1/2 & 1 & -13/2 & 7 \\ 0 & 0 & 1 & 8 & -6 \\ 0 & 5 & 0 & 2 & 13 \end{pmatrix} \sim$$
$$\begin{pmatrix} 2 & -1 & 0 & -3 & 8 \\ 0 & -1/2 & 1 & -13/2 & 7 \\ 0 & 0 & 1 & 8 & -6 \\ 0 & 0 & 10 & -63 & 83 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 0 & -3 & 8 \\ 0 & -1/2 & 1 & -13/2 & 7 \\ 0 & 0 & 1 & 8 & -6 \\ 0 & 0 & 0 & -143 & 143 \end{pmatrix},$$

amelyből következik az egyenletrendszer megoldása. □

3.25. példa. Oldjuk meg a

$$\begin{array}{rcrcr} 0.0002x_1 & - & 30.5x_2 & = & -60.99 \\ 5.060x_1 & - & 1.05x_2 & = & 250.9 \end{array}$$

egyenletrendszert a Gauss-eliminációval 4-jegyű aritmetikát használva a számolásokhoz! A 3.23. algoritmust követve, először kiszámoljuk az $l_{21} = 5.060/0.0002 = 25300$ szorzótényezőt (4 értékes jegyre kerekítve), ezzel beszorozzuk az első egyenletet, és a kapott sort kivonjuk a másodikból:

$$\begin{pmatrix} 0.0002 & -30.5 & -60.99 \\ 5.06 & -1.05 & 250.9 \end{pmatrix} \sim \begin{pmatrix} 0.0002 & -30.5 & -60.99 \\ 0 & 771700 & 1543000 \end{pmatrix}.$$

(Megjegyezzük, hogy a 3.23. algoritmussal a 2. sorban levő 0-t nem numerikusan számoljuk.) Ezt megoldva kapjuk az $\tilde{x}_1 = -100.0$ és $\tilde{x}_2 = 1.999$ numerikus megoldást. Könnyen ellenőrizhetjük, hogy az egyenletrendszer pontos megoldása $x_1 = 50$ és $x_2 = 2$. A számolt megoldásokban tehát 300% ill. 0.05%-os relatív hiba van! Különösen hatalmas a hiba az első változó értékében.

Végezzük most el ugyanezt a számolást az egyenletrendszeren úgy, hogy először felcseréljük a két egyenletet. Kapjuk:

$$\begin{pmatrix} 5.06 & -1.05 & 250.9 \\ 0.0002 & -30.5 & -60.99 \end{pmatrix} \sim \begin{pmatrix} 5.06 & -1.05 & 250.9 \\ 0 & -30.5 & -61.0 \end{pmatrix}.$$

amiből következik, hogy $x_1 = 50.00$ és $x_2 = 2.000$, amelyek pontosan megegyeznek a tényleges megoldás értékekkel!

Mi a különbség a két számolásban? Az első esetben l_{21} kiszámolásakor egy kis számmal (0.0002) kellett osztani, ami a kerekítési hiba jelentős növekedéséhez vezetett. A második esetben viszont 5.06-gyel osztottunk l_{21} kiszámításakor, és a végső eredményben nem kaptunk kerekítési hibát. □

Részleges főelemkiválasztás

Az előző két példa mutatja, hogy néha kell, és sok esetben célszerű módosítani a 3.23. algoritmust. Erre az egyik legegyszerűbb stratégia a következő, *részleges főelemkiválasztásnak* (vagy egyszerűen csak *főelemkiválasztásnak*) nevezett módszer: a Gauss-elimináció k-adik lépése előtt keressük meg a k-adik oszlopban a főátlóban és az alatta álló elemek közül a legnagyobb abszolút értékűt, azaz legyen

$$|a_{lk}| = \max\{|a_{ik}|\colon\ i = k, \ldots, n\}.$$

(A maximális elem az l-edik sorban van.) Cseréljük fel a k-adik és l-edik sort, és folytassuk az eliminációt. Ezzel a 3.24. és 3.25. példákban vizsgált problémákat ki tudjuk küszöbölni: ha $a_{kk}^{(k-1)} = 0$, akkor a sorcsere után nemnulla elem kerül erre a pozícióra (feltéve ha van nemnulla elem $a_{kk}^{(k-1)}$ alatt), valamint folytatva a Gauss-eliminációt a sorcserékkel elérhető lehető legnagyobb abszolút értékű számmal fogunk osztani, ami a kerekítési hibákat csökkenti.

3.26. tétel. *A következő állítások ekvivalensek:*

1. *az* $\mathbf{Ax} = \mathbf{b}$ *egyenlet egyértelműen megoldható Gauss-eliminációval részleges főelemkiválasztást használva,*
2. $\det(\mathbf{A}) \neq 0$*,*
3. *az* $\mathbf{A}$ *mátrix invertálható,*
4. *az* $\mathbf{Ax} = \mathbf{b}$ *egyenletnek létezik megoldása minden* $\mathbf{b}$ *vektorra.*

Bizonyítás. Lineáris algebrából ismert, hogy a 2., 3. és 4. állítások ekvivalensek (lásd a 3.2. tételt). Így most azt látjuk be, hogy 1. és 2. ekvivalens.

Tegyük fel először, hogy 1. teljesül. Legyen $\mathbf{A}^{(0)} = \mathbf{A}$, és jelöljük $\mathbf{A}^{(k)}$-val a Gauss-elimináció k-adik lépésekor kapott együtthatómátrixot. A determinánsok tulajdonságából következik, hogy $\det(\mathbf{A}^{(k)}) = \det(\mathbf{A}^{(k-1)})$, ha nem történt sorcsere a k-adik lépesben, ill. $\det(\mathbf{A}^{(k)}) = -\det(\mathbf{A}^{(k-1)})$, ha volt sorcsere. Mivel a feltétel szerint a Gauss-elimináció elvégezhető, ezért az $\mathbf{A}^{(n-1)}$ mátrixhoz tartozó trianguláris egyenletrendszer megoldható, azaz $\det(\mathbf{A}^{(n-1)}) \neq 0$. Ebből viszont következik, hogy $\det(\mathbf{A}) = \pm\det(\mathbf{A}^{(n-1)}) \neq 0$.

Belátjuk, hogy ha a részleges főelemkiválasztással végzett Gauss-elimináció k-adik lépése nem hajtható végre, akkor $\det(\mathbf{A}) = 0$. A k-adik lépés akkor és csak akkor nem hajtható végre, ha $a_{ik}^{(k-1)} = 0$ minden $i = k, \ldots, n$-re, azaz:

$$\mathbf{A}^{(k-1)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,k-1} & a_{1k} & a_{k,k+1} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2,k-1}^{(1)} & a_{2k}^{(1)} & a_{2,k+1}^{(1)} & \cdots & a_{2n}^{(1)} \\ & & \ddots & & & & & \\ 0 & 0 & \cdots & a_{k-1,k-1}^{(k-2)} & a_{k-1,k}^{(k-2)} & a_{k-1,k+1}^{(k-2)} & \cdots & a_{k-1,n}^{(k-2)} \\ 0 & 0 & \cdots & 0 & 0 & a_{k,k+1}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & a_{n,k+1}^{(k-1)} & \cdots & a_{nn}^{(k-1)} \end{pmatrix}.$$

Ezért

$$\det(\mathbf{A}^{(k-1)}) = a_{11}a_{22}^{(1)}\cdots a_{k-1,k-1}^{(k-2)} \det \begin{pmatrix} 0 & a_{k,k+1}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n,k+1}^{(k-1)} & \cdots & a_{nn}^{(k-1)} \end{pmatrix} = 0,$$

és így $\det(\mathbf{A}) = \pm\det(\mathbf{A}^{(k-1)}) = 0$. □

3.27. példa. Tekintsük újra a 3.22. példa egyenletrendszerét, és oldjuk meg a feladatot Gauss-elimináció̧val részleges főelemkiválasztást használva! A kibővített mátrixok sorozata a következő:

$$\begin{pmatrix} 2 & -1 & 0 & -3 & 8 \\ 2 & -1 & 1 & 5 & 2 \\ -3 & 1 & 1 & -2 & -5 \\ 2 & 4 & 0 & -1 & 21 \end{pmatrix} \sim \begin{pmatrix} -3 & 1 & 1 & -2 & -5 \\ 2 & -1 & 1 & 5 & 2 \\ 2 & -1 & 0 & -3 & 8 \\ 2 & 4 & 0 & -1 & 21 \end{pmatrix} \sim$$

$$\begin{pmatrix} -3 & 1 & 1 & -2 & -5 \\ 0 & -1/3 & 5/3 & 11/3 & -4/3 \\ 0 & -1/3 & 2/3 & -13/3 & 14/3 \\ 0 & 14/3 & 2/3 & -7/3 & 53/3 \end{pmatrix} \sim \begin{pmatrix} -3 & 1 & 1 & -2 & -5 \\ 0 & 14/3 & 2/3 & -7/3 & 53/3 \\ 0 & -1/3 & 2/3 & -13/3 & 14/3 \\ 0 & -1/3 & 5/3 & 11/3 & -4/3 \end{pmatrix} \sim$$

$$\begin{pmatrix} -3 & 1 & 1 & -2 & -5 \\ 0 & 14/3 & 2/3 & -7/3 & 53/3 \\ 0 & 0 & 5/7 & -9/2 & 83/14 \\ 0 & 0 & 12/7 & 7/2 & -1/14 \end{pmatrix} \sim \begin{pmatrix} -3 & 1 & 1 & -2 & -5 \\ 0 & 14/3 & 2/3 & -7/3 & 53/3 \\ 0 & 0 & 12/7 & 7/2 & -1/14 \\ 0 & 0 & 5/7 & -9/2 & 83/14 \end{pmatrix} \sim$$

$$\begin{pmatrix} -3 & 1 & 1 & -2 & -5 \\ 0 & 14/3 & 2/3 & -7/3 & 53/3 \\ 0 & 0 & 12/7 & 7/2 & -1/14 \\ 0 & 0 & 0 & -143/24 & 143/24 \end{pmatrix}$$

Látható, hogy az első és a harmadik eliminációs lépés előtt volt sorcsere. A trianguláris egyenletet megoldva kapjuk: $x_1 = 4$, $x_2 = 3$, $x_3 = 2$ és $x_4 = -1$. □

Tegyük fel, hogy egy $\mathbf{A}$ együtthatómátrixon részleges főelemkiválasztással elvégzett Gauss-elimináció közben szükséges sorcseréket összegyűjtjük. Végezzük el ezeket a sorcseréket egyszerre előre, az első eliminációs lépés előtt. Ezután a kapott mátrixon sorcsere nélkül végrehajtható lesz a Gauss-elimináció (és az eredménye ugyanaz, mint az $\mathbf{A}$ mátrixon részleges főelemkiválasztással elvégzett Gauss-eliminációé). A 3.7. tétel szerint a sorcserék hatása egy megfelelő permutációs $\mathbf{P}$ mátrixszal (balról) történő szorzással ekvivalens. A 3.26. tételből tehát rögtön következik az alábbi eredmény:

3.28. tétel. *Ha* $\det(\mathbf{A}) \neq 0$*, akkor létezik olyan* $\mathbf{P}$ *permutációs mátrix, hogy a* $\mathbf{PAx} = \mathbf{Pb}$ *egyenletrendszer egyértelműen megoldható Gauss-eliminációval (sorcserék nélkül) minden* $\mathbf{b}$ *vektorra.*

Teljes főelemkiválasztás

A kerekítési hibák további kiküszöbölésére használhatjuk a részleges főelemkiválasztás következő módosítását, az ún. *teljes főelemkiválasztás* módszerét: a Gauss-elimináció k-adik lépése előtt keressük meg az első olyan l és m sor- és oszlopindexet, amelyre

$$|a_{lm}| = \max\{|a_{ij}|\colon\ i = k, \dots, n,\ j = k, \dots, n\}.$$

(A maximális elem az l-edik sorban és m-edik oszlopban van.) Cseréljük fel a k-adik és l-edik sort és a k-adik és m-edik oszlopot. Jegyezzük meg, hogy az oszlopcserével melyik oszlop melyik ismeretlen együtthatóit tartalmazza, és folytassuk az eliminációt.

Ennek a módszernek a hátránya a részleges főelemkiválasztáshoz képest az, hogy sokkal több összehasonlításra van szükség, ami lassítja a módszert.

3.29. példa. Tekintsük újra a 3.22. és 3.27 példa egyenletrendszerét, és oldjuk meg a feladatot most

Gauss-elmiminációval teljes főelemkiválasztást használva:

$$\left(\begin{array}{rrrrr} 1 & -2 & -2 & -2 & -11 \\ 2 & -1 & 2 & 4 & -8 \\ -1 & 2 & 3 & -4 & 27 \\ -2 & 1 & 4 & -2 & 28 \\ \mathtt{x}_1 & \mathtt{x}_2 & \mathtt{x}_3 & \mathtt{x}_4 & \end{array}\right) \sim \left(\begin{array}{rrrrr} 2 & -1 & 2 & 4 & -8 \\ 1 & -2 & -2 & -2 & -11 \\ -1 & 2 & 3 & -4 & 27 \\ -2 & 1 & 4 & -2 & 28 \\ \mathtt{x}_1 & \mathtt{x}_2 & \mathtt{x}_3 & \mathtt{x}_4 & \end{array}\right) \sim$$

$$\left(\begin{array}{rrrrr} 4 & -1 & 2 & 2 & -8 \\ -2 & -2 & -2 & 1 & -11 \\ -4 & 2 & 3 & -1 & 27 \\ -2 & 1 & 4 & -2 & 28 \\ \mathtt{x}_4 & \mathtt{x}_2 & \mathtt{x}_3 & \mathtt{x}_1 & \end{array}\right) \sim \left(\begin{array}{rrrrr} 4 & -1 & 2 & 2 & -8 \\ 0 & -5/2 & -1 & 2 & -15 \\ 0 & 1 & 5 & 1 & 19 \\ 0 & 1/2 & 5 & -1 & 24 \\ \mathtt{x}_4 & \mathtt{x}_2 & \mathtt{x}_3 & \mathtt{x}_1 & \end{array}\right) \sim$$

$$\left(\begin{array}{rrrrr} 4 & -1 & 2 & 2 & -8 \\ 0 & 1 & 5 & 1 & 19 \\ 0 & -5/2 & -1 & 2 & -15 \\ 0 & 1/2 & 5 & -1 & 24 \\ \mathtt{x}_4 & \mathtt{x}_2 & \mathtt{x}_3 & \mathtt{x}_1 & \end{array}\right) \sim \left(\begin{array}{rrrrr} 4 & 2 & -1 & 2 & -8 \\ 0 & 5 & 1 & 1 & 19 \\ 0 & -1 & -5/2 & 2 & -15 \\ 0 & 5 & 1/2 & -1 & 24 \\ \mathtt{x}_4 & \mathtt{x}_3 & \mathtt{x}_2 & \mathtt{x}_1 & \end{array}\right) \sim$$

$$\left(\begin{array}{rrrrr} 4 & 2 & -1 & 2 & -8 \\ 0 & 5 & 1 & 1 & 19 \\ 0 & 0 & -23/10 & 11/5 & -56/5 \\ 0 & 0 & -1/2 & -2 & 5 \\ \mathtt{x}_4 & \mathtt{x}_3 & \mathtt{x}_2 & \mathtt{x}_1 & \end{array}\right) \sim \left(\begin{array}{rrrrr} 4 & 2 & -1 & 2 & -8 \\ 0 & 5 & 1 & 1 & 19 \\ 0 & 0 & -23/10 & 11/5 & -56/5 \\ 0 & 0 & 0 & -57/23 & 171/23 \\ \mathtt{x}_4 & \mathtt{x}_3 & \mathtt{x}_2 & \mathtt{x}_1 & \end{array}\right)$$

Azért, hogy az oszlopcseréket követni tudjuk, kibővítettük a mátrixot egy plusz sorral, ahol azt jelöljük, hogy az adott oszlop melyik változó együtthatóit tartalmazza. Az első eliminációs lépés előtt felcseréltük az első és második sort és az első és negyedik oszlopot, mivel 4 volt a maximális elem az együtthatók abszolút értékei közül. (Lehetett volna az első és a harmadik sor és az első és negyedik oszlop felcserélésével -4-et behozni a főelem pozíciójába; vagy pedig az első és negyedik sor és az első és harmadik oszlop cseréjével is 4-et behozni az első főelem pozíciójába.) A második eliminációs lépés előtt felcseréltük a második és harmadik sort és a második és harmadik oszlopot. A harmadik eliminációs lépés előtt pedig nem volt sor vagy oszlop csere. A megoldást most is a trianguláris egyenletrendszert megoldva kapjuk, de például a 4. egyenletből most az x_1 értékét kapjuk meg. A végeredmény: $x_1 = -3$, $x_2 = 2$, $x_3 = 4$ és $x_4 = -2$.

Természetesen a részleges ill. a teljes főelemkiválasztás módszerének előnye csak akkor jelentkezik, ha numerikusan számoljuk végig az egyenletrendszert. □

Sorkiegyenlítés

Numerikus tapasztalat az, hogy ha az együtthatómátrix elemei között jelentős nagyságrendi eltérés van, akkor a kerekítési hiba megnőhet a számolás során (lásd a 3.25. példát). Ezért szokás az egyes egyenleteket beszorozni valamely nemnulla számokkal úgy, hogy a kapott egyenletrendszer együtthatói közel azonos nagyságrendűek legyenek. Ezt a beszorzást nevezzük *sorkiegyenlítésnek*. Hasonlóan, ha az egyenletrendszer megoldásai eltérő nagyságrendűek, akkor azokat is célszerű kiegyenlíteni, azaz az együtthatómátrix oszlopait beszorozni valamely nemnulla számokkal. Erre jelenleg nem ismert jó stratégia (az $\mathbf{A}$ mátrix és a $\mathbf{b}$ vektor ismeretében), ezért itt csak a sorkiegyenlítéssel foglalkozunk.

Keresünk tehát olyan $d_1, \ldots, d_n \neq 0$ számokat, hogy a $\mathbf{B} := \mathbf{DA}$ mátrix elemei közel azonos nagyságrendűek legyenek, ahol $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n)$. Ekkor az $\mathbf{Ax} = \mathbf{b}$ egyenletrendszer helyett a $\mathbf{DAx} = \mathbf{Db}$ egyenletrendszert oldjuk meg numerikusan. Az egyik egyszerű startégia szerint úgy választjuk $\mathbf{D}$-t, hogy $\max\{|b_{ij}| : 1 \leq j \leq n\} \approx 1$ legyen minden $i = 1, \ldots, n$-re. Ezt elérhetjük a $d_i := 1/s_i$, $s_i := \max\{|a_{ij}| : 1 \leq j \leq n\}$ választással. Ezzel az a probléma, hogy az osztások további kerekítési hibát vezetnek be a számolásba. Ezt kiküszöbölendő csinálhatjuk a következőt: legyen β a számábrázolás alapja a számítógépen, és legyen r_i a legkisebb egész, hogy $\beta^{r_i} \geq s_i$, és definiáljuk $b_{ij} := a_{ij}/\beta^{r_i}$ $(i, j = 1, \ldots, n)$. Ekkor az osztásnál nem lesz kerekítési hiba, és $1/\beta < \max_{1 \leq j \leq n} |b_{ij}| \leq 1$ teljesül minden $i = 1, \ldots, n$-re.

Könnyen igazolható a következő állítás:

3.30. tétel. *Tegyük fel, hogy egy* $\mathbf{A}$ *együtthatómátrixon sorkiegyenlítést végeztünk olyan* $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n)$ *szorzótényezőkkel (pl.* β *hatványokkal), amelyek nem eredményeztek kerekítési hibát. Ekkor ha a* $\mathbf{DA}$ *mátrixon végzett (részleges vagy teljes) főelemkiválasztás ugyanazokat a sorcseréket (és oszlopcseréket) eredményezi, mint az* $\mathbf{A}$ *mátrixon, akkor az* $\mathbf{Ax} = \mathbf{b}$ *és* $\mathbf{DAx} = \mathbf{Db}$ *egyenletek numerikus megoldásai pontosan ugyanazok lesznek.*

Ebből következik, hogy a kiegyenlítésnek csak a főelemkiválasztásra van hatása. A Gauss-eliminációnak a következő módosításában a súlyozás helyett csak ún. *implicit sorkiegyenlítést* végzünk, a főelemek kiválasztásához használjuk csak a súlyokat. Ez a módszer a gyakorlatban az egyik leggyakrabban használt algoritmus lineáris egyenletrendszerek megoldására.

3.31. algoritmus. Gauss-elimináció részleges főelemkiválasztással és implicit sorkiegyenlítéssel

INPUT: $a_{ij}, \quad (i = 1, \ldots, n, \quad j = 1, \ldots, n+1)$ - kibővített együtthatómátrix
OUTPUT: $x_1, \ldots, x_n$

(súlyok kiszámítása:)
for $i = 1, \ldots, n$ **do**
    $s_i \leftarrow \max\limits_{1 \le j \le n} |a_{ij}|$
end do
(eliminációs:)
for $k = 1, \ldots, n-1$ **do**
    legyen l a legkisebb olyan index, amelyre $\dfrac{|a_{lk}|}{s_l} = \max\limits_{k \le i \le n} \dfrac{|a_{ik}|}{s_i}$
    cseréljük fel az $\mathbf{A}$ mátrix k-adik és l-edik sorát
    for $i = k+1, \ldots, n$ **do**
        $l_{ik} \leftarrow a_{ik}/a_{kk}$
        for $j = k+1, \ldots, n+1$ **do**
            $a_{ij} \leftarrow a_{ij} - l_{ik} a_{kj}$
        end do
    end do
end do
(visszahelyettesítés:)
$x_n \leftarrow a_{n,n+1}/a_{nn}$
for $i = n-1, \ldots, 1$ **do**
    $x_i \leftarrow (a_{i,n+1} - \sum\limits_{j=i+1}^{n} a_{ij} x_j)/a_{ii}$
end do
output$(x_1, x_2, \ldots, x_n)$

Megjegyezzük, hogy az eddigi módszereknél gyakran kellett egy $\mathbf{A} = (a_{ij})$ mátrix két sorát felcserélni. Ez sok művelettel jár, ezért az algoritmusok programozásakor csinálhatjuk a következőt: Az $\mathbf{A}$ mátrixot tároljuk egy $a[i, j]$ tömbben. Definiálunk egy $m[i]$ vektort, amelynek kezdeti értéke $m[i] = i$, $(i = 1, \ldots, n)$. A k-adik és l-edik sor cseréjekor csak az $m[\cdot]$ vektor k-adik és l-edik elemeit cseréljük fel. Amikor az algoritmusban az $\mathbf{A}$ mátrix egy a_{ij} elemére kell hivatkozni, akkor használjuk az $a[m[i], j]$ elemet.

3.32. tétel. *Ha az* $\mathbf{A}$ *mátrix diagonálisan domináns, akkor a Gauss-elimináció főelemkiválasztás nélkül végrehajtható az* $\mathbf{Ax} = \mathbf{b}$ *egyenletrendszeren, és a módszer stabil a kerekítési hibákra nézve.*

Bizonyítás. Megjegyezzük, hogy ha az $\mathbf{A}$ mátrix diagonálisan domináns, akkor a 3.8. tétel szerint az $\mathbf{Ax} = \mathbf{b}$ egyenletrendszernek létezik egyértelmű megoldása.

Megmutatjuk, hogy a Gauss-eliminációval kapott $\mathbf{A}^{(1)}$, $\mathbf{A}^{(2)}$, ..., $\mathbf{A}^{(n-1)}$ mátrixok mindegyike képezhető és diagonálisan domináns. Mivel $\mathbf{A}^{(0)} = \mathbf{A}$ diagonálisan domináns, ezért $|a_{11}| > \sum_{j=2}^{n} |a_{1j}|$, így $a_{11} \neq 0$. Ebből következik, hogy az $\mathbf{A}^{(1)}$ mátrix képezhető. Megmutatjuk, hogy $\mathbf{A}^{(1)}$ diagonálisan domináns. Mivel $\mathbf{A}^{(1)}$ első sora megegyezik $\mathbf{A}$ első sorával, ezért az első sor diagonálisan domináns. Legyen $1 < i \leq n$. Használva, hogy $a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j}$, $(j = 2, \ldots, n)$, valamint $a_{i1}^{(1)} = 0$, kapjuk

$$\sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij}^{(1)}| = \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j}| \leq \sum_{\substack{j=2 \\ j\neq i}}^{n} (|a_{ij}| + \frac{|a_{i1}|}{|a_{11}|} |a_{1j}|) = \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij}| + \frac{|a_{i1}|}{|a_{11}|} \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{1j}|.$$

Mivel az $\mathbf{A}$ mátrix i-edik és az első sora is diagonálisan domináns, ezért

$$\begin{aligned} \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij}^{(1)}| &< |a_{ii}| - |a_{i1}| + \frac{|a_{i1}|}{|a_{11}|} (|a_{11}| - |a_{1i}|) \\ &= |a_{ii}| - \frac{|a_{i1}|}{|a_{11}|} |a_{1i}| \\ &\leq \left| a_{ii} - \frac{a_{i1}}{a_{11}} a_{1i} \right| \\ &= |a_{ii}^{(1)}|. \end{aligned}$$

Ezzel beláttuk, hogy $\mathbf{A}^{(1)}$ minden sora diagonálisan domináns, azaz a mátrix diagonálisan domináns.

Ehhez hasonlóan belátható, hogy $\mathbf{A}^{(2)}$,...,$\mathbf{A}^{(n-1)}$ mindegyike definiált és diagonálisan domináns.

A módszer stabilitását itt nem bizonyítjuk be. □

Belátható a következő tétel:

3.33. tétel. *Legyen* $\mathbf{A}$ *szimmetrikus* $n \times n$*-es mátrix. Ekkor* $\mathbf{A}$ *akkor és csak akkor pozitív definit, ha a Gauss-elimináció főelemkiválasztás nélkül végrehajtható az* $\mathbf{Ax} = \mathbf{b}$ *egyenletrendszeren, és a főelemek pozitívak. Továbbá ebben az esetben a módszer stabil a kerekítési hibákra nézve.*

Feladatok

1. Oldja meg a következő egyenletrendszereket Gauss-eliminációval:
   (i) főelemkiválasztás nélkül,
   (ii) részleges főelemkiválasztással,
   (iii) teljes főelemkiválasztással,
   (iv) részleges főelemkiválasztással és implicit sorkiegyenlítéssel:

(a)

$$\begin{array}{rcrcrcr} 2x_1 & + & 2x_2 & - & 2x_3 & = & -4 \\ -x_1 & + & 3x_2 & & & = & -11 \\ 4x_1 & + & 2x_2 & - & 3x_3 & = & -1 \end{array}$$

(b)

$$\begin{array}{rcrcrcrcr} -x_1 & - & 3x_2 & & & + & 2x_4 & = & 10 \\ -2x_1 & + & 3x_2 & & & + & x_4 & = & 8 \\ 4x_1 & + & x_2 & - & x_3 & - & 3x_4 & = & -21 \\ 2x_1 & + & x_2 & - & x_3 & + & 3x_4 & = & 7 \end{array}$$

2. Használjon 4-jegyű aritmetikát a számolásokhoz, és az előző feladat kérdését alkalmazza a következő egyenletekre:

(a)

$$\begin{array}{rcrcrcr} 1.03x_1 & - & 1.1x_2 & + & 8x_3 & = & -9.06 \\ -4.1x_1 & + & 10.1x_2 & - & 6x_3 & = & 106.2 \\ 2.11x_1 & - & 4.2x_2 & + & 12x_3 & = & -40.22 \end{array}$$

(pontos megoldás: (-2, 10, 0.5)),

(b)

$$\begin{array}{rcrcrcr} x_1 & + & \frac{1}{2}x_2 & + & \frac{1}{3}x_3 & = & 20 \\ \frac{1}{2}x_1 & + & \frac{1}{3}x_2 & + & \frac{1}{4}x_3 & = & 14 \\ \frac{1}{3}x_1 & + & \frac{1}{4}x_2 & + & \frac{1}{5}x_3 & = & 11 \end{array}$$

(pontos megoldás: (6, -12, 60))

3. Lássa be a 3.30. tételt!
4. Lássa be a 3.33. tételt (a stabilitásra vonatkozó állítás nélkül)!

3.4. Gauss–Jordan-elimináció

A Gauss-elimináció egyik módosítása a *Gauss–Jordan-elimináció* (vagy egyszerűen *Jordan-elimináció*), ahol a Gauss-elimináció lépéseivel egységmátrixra alakítjuk át az együtthatómátrixot, azaz az $(\mathbf{A}, \mathbf{b})$ kibővített mátrixot az $(\mathbf{I}, \mathbf{b}^{(n-1)})$ alakra hozzuk. Ekkor az egyenletrendszer megoldása $\mathbf{x} = \mathbf{b}^{(n-1)}$ lesz.

3.34. algoritmus. Gauss–Jordan-elimináció

INPUT: a_{ij}, $(i = 1, \ldots, n, \quad j = 1, \ldots, n+1)$ - kibővített együtthatómátrix
OUTPUT: $x_1, \ldots, x_n$

(együtthatómátrix diagonális alakra hozása:)
for $k = 1, \ldots, n$ **do**
  for $i = 1, \ldots, n$ **do**
    if $i \neq k$ **do**
      $l_{ik} \leftarrow a_{ik}/a_{kk}$
      for $j = k+1, \ldots, n+1$ **do**
        $a_{ij} \leftarrow a_{ij} - l_{ik}a_{kj}$
      end do
    end do
  end do
end do
for $i = 1, \ldots, n$ **do**
  $x_i \leftarrow a_{i,n+1}/a_{ii}$

end do
output$(x_1, x_2, \ldots, x_n)$

Ellenőrizhető, hogy a Gauss–Jordan-elimináció műveletigénye: $n^3/2 + \mathcal{O}(n^2)$ osztás/szorzás.

3.35. példa. Alkalmazzuk a Gauss–Jordan-eliminációt a 3.22. feladatban vizsgált egyenletrendszer megoldására:

$$\begin{pmatrix} 1 & -2 & -2 & -2 & -11 \\ 2 & -1 & 2 & 4 & -8 \\ -1 & 2 & -3 & -4 & 3 \\ -2 & 1 & 4 & -2 & 28 \end{pmatrix} \sim \begin{pmatrix} 1 & -2 & -2 & -2 & -11 \\ 0 & 3 & 6 & 8 & 14 \\ 0 & 0 & -5 & -6 & -8 \\ 0 & -3 & 0 & -6 & 6 \end{pmatrix} \sim$$
$$\begin{pmatrix} 1 & 0 & 2 & 10/3 & -5/3 \\ 0 & 3 & 6 & 8 & 14 \\ 0 & 0 & -5 & -6 & -8 \\ 0 & 0 & 6 & 2 & 20 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 14/15 & -73/15 \\ 0 & 3 & 0 & 4/5 & 22/5 \\ 0 & 0 & -5 & -6 & -8 \\ 0 & 0 & 0 & -26/5 & 52/5 \end{pmatrix} \sim$$
$$\begin{pmatrix} 1 & 0 & 0 & 0 & -3 \\ 0 & 3 & 0 & 0 & 6 \\ 0 & 0 & -5 & 0 & -20 \\ 0 & 0 & 0 & -26/5 & 52/5 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 0 & -3 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & -2 \end{pmatrix}$$

A megoldás leolvasható a mátrix utolsó oszlopáról: $x_1 = -3$, $x_2 = 2$, $x_3 = 4$ és $x_4 = -2$. □

A Gauss-eliminációnál megfogalmazott részleges ill. teljes főelemkiválasztás alkalmazható a Gauss–Jordan-elimináció esetében is.

3.36. példa. Alkalmazzuk a Gauss–Jordan-eliminációt részleges főelemkiválasztással a 3.22. feladatban vizsgált egyenletrendszer megoldására:

$$\begin{pmatrix} 1 & -2 & -2 & -2 & -11 \\ 2 & -1 & 2 & 4 & -8 \\ -1 & 2 & 3 & -4 & 27 \\ -2 & 1 & 4 & -2 & 28 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 2 & 4 & -8 \\ 1 & -2 & -2 & -2 & -11 \\ -1 & 2 & 3 & -4 & 27 \\ -2 & 1 & 4 & -2 & 28 \end{pmatrix} \sim$$
$$\begin{pmatrix} 2 & -1 & 2 & 4 & -8 \\ 0 & -3/2 & -3 & -4 & -7 \\ 0 & 3/2 & 4 & -2 & 23 \\ 0 & 0 & 6 & 2 & 20 \end{pmatrix} \sim \begin{pmatrix} 2 & 0 & 4 & 20/3 & -10/3 \\ 0 & -3/2 & -3 & -4 & -7 \\ 0 & 0 & 1 & -6 & 16 \\ 0 & 0 & 6 & 2 & 20 \end{pmatrix} \sim$$
$$\begin{pmatrix} 2 & 0 & 4 & 20/3 & -10/3 \\ 0 & -3/2 & -3 & -4 & -7 \\ 0 & 0 & 6 & 2 & 20 \\ 0 & 0 & 1 & -6 & 16 \end{pmatrix} \sim \begin{pmatrix} 2 & 0 & 0 & 16/3 & -50/3 \\ 0 & -3/2 & 0 & -3 & 3 \\ 0 & 0 & 6 & 2 & 20 \\ 0 & 0 & 0 & -19/3 & 38/3 \end{pmatrix} \sim$$
$$\begin{pmatrix} 2 & 0 & 0 & 0 & -6 \\ 0 & -3/2 & 0 & 0 & -3 \\ 0 & 0 & 6 & 0 & 24 \\ 0 & 0 & 0 & -19/3 & 38/3 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 0 & -3 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & -2 \end{pmatrix}$$

A megoldás tehát $x_1 = -3$, $x_2 = 2$, $x_3 = 4$ és $x_4 = -2$. □

Feladatok

1. Oldja meg a 3.3. szakasz 1. és 2. feladatában szereplő egyenletrendszereket Gauss–Jordan-eliminációval!
2. Lássa be, hogy a Gauss–Jordan-elimináció műveletigénye $n^3/2 + n^2 - n/2$ osztás ill. szorzás!

3.5. Tridiagonális egyenletrendszerek

Egy négyzetes (a_{ij}) mátrixot *tridiagonálisnak* nevezünk, ha $a_{ij} = 0$ minden $|i-j| > 1$-re, azaz nemnulla elemek csak a mátrix főátlójában, ill. közvetlen alatta vagy felette lehetnek. Tridiagonális lineáris egyenletrendszerek (azaz ahol az együtthatómátrix tridiagonális) gyakran előfordulnak alkalmazásokban, így ezek fontos speciális esetei a lineáris egyenletrendszereknek. A következő jelölést használjuk:

$$\begin{pmatrix} d_1 & c_1 & 0 & 0 & \cdots & 0 \\ a_1 & d_2 & c_2 & 0 & \cdots & 0 \\ 0 & a_2 & d_3 & c_3 & \cdots & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & 0 & \cdots & a_{n-2} & d_{n-1} & c_{n-1} \\ 0 & 0 & \cdots & 0 & a_{n-1} & d_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-1} \\ b_n \end{pmatrix}. \tag{3.10}$$

Egy tridiagonális mátrix elemeit célszerű a jelölés szerinti három vektorban tárolni: (a_i), (d_i) és (c_i), így összesen $3n-2$ tárolóhely kell az együtthatóknak.

Könnyen látható, hogy a Gauss-eliminációt alkalmazva a (3.10) egyenletrendszerre az a_i számok kinullázódnak az elimináció végére, a c_i számok viszont nem fognak megváltozni. A d_i és b_i értékek megváltoznak az elimináció során. A következő algoritmust úgy fogalmaztuk meg, hogy az elimináció során felülírja a (d_i) és (b_i) vektorokat.

3.37. algoritmus. Gauss-elimináció tridiagonális egyenletrendszerre

INPUT: a_i, c_i $(i = 1, \ldots, n-1)$, d_i, b_i $(i = 1, \ldots, n)$
OUTPUT: x_1,...,x_n

(elimináció:)
for $i = 2, \ldots, n$ **do**
  $temp \leftarrow a_{i-1}/d_{i-1}$
  $d_i \leftarrow d_i - temp \cdot c_{i-1}$
  $b_i \leftarrow b_i - temp \cdot b_{i-1}$
end do
(visszahelyettesítés:)
$x_n \leftarrow b_n/d_n$
for $i = n-1, \ldots, 1$ **do**
  $x_i \leftarrow (b_i - c_i x_{i+1})/d_i$
end do
output$(x_1, x_2, \ldots, x_n)$

A módszer műveletigénye meglepően kicsi: $5n-4$ szorzás/osztás. Ha ezt összehasonlítjuk a 3.23. algoritmus $n^3/3$ nagyságrendjével, akkor látjuk, hogy tridiagonális rendszerek megoldására feltétlenül ezt a speciális algoritmust kell használni.

A 3.32. tételből következik, hogy ha az $\mathbf{A}$ tridiagonális mátrix diagonálisan domináns, akkor a 3.37. algoritmus végrehajtható, azaz nincs szükség sorcserékre a Gauss-elimináció közben.

Feladatok

1. Oldja meg a következő tridiagonális egyenletrendszert:

$$\begin{array}{rcrcrcrcrcrcr} x_1 & - & 0.5x_2 & & & & & & & & = & 1.5 \\ 0.5x_1 & + & 4x_2 & - & 0.5x_3 & & & & & & = & -4.0 \\ & & 0.5x_2 & + & 2x_3 & - & 0.5x_4 & & & & = & 2.0 \\ & & & & 0.5x_3 & + & 4x_4 & - & 0.5x_5 & & = & -4.0 \\ & & & & & & 0.5x_4 & + & 2x_5 & - & 0.5x_6 & = & 2.0 \\ & & & & & & & & 0.5x_5 & + & x_6 & = & -0.5 \end{array}$$

2. Lássa be, hogy a 3.37. algoritmus műveletigénye $5n - 4$ osztás/szorzás!
3. Fogalmazzon meg a 3.37. algoritmushoz hasonló algoritmust olyan szalagmátrixokra, ahol nemnulla elemek csak a főátlóban és az az alatti és feletti 2-2 átlóban lehetnek.

3.6. Szimultán egyenletrendszerek

Gyakran előfordul, hogy ún. *szimultán egyenletrendszereket*, azaz olyan $\mathbf{Ax} = \mathbf{b}^{(i)}$ alakú egyenletrendszereket kell megoldanunk $i = 1, \ldots, m$-re, ahol az együtthatómátrix azonos, de az egyenletek jobb oldala különböző. Ezt röviden az $\mathbf{AX} = \mathbf{B}$ egyenlettel írhatjuk le, ahol az $n \times m$-es $\mathbf{B} = (\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \ldots, \mathbf{b}^{(m)})$ mátrix i-edik oszlopa $\mathbf{b}^{(i)}$, és az $n \times m$-es $\mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)})$ mátrix i-edik oszlopa $\mathbf{x}^{(i)}$, az $\mathbf{Ax}^{(i)} = \mathbf{b}^{(i)}$ egyenlet megoldása. Mivel a Gauss ill. a Gauss–Jordan-elimináció végrehajthatósága ill. főelemkiválasztásnál a cserék eldöntése csak az együtthatómátrixon múlik, alkalmazhatjuk ezeket a módszereket az $n \times (n + m)$-es $(\mathbf{A}, \mathbf{B})$ kibővített mátrixon. Pl. ha Gauss–Jordan-eliminációt végzünk, akkor az $(\mathbf{A}, \mathbf{B})$ kibővített mátrixot az $(\mathbf{I}, \mathbf{X})$ alakra hozzuk, és ekkor $\mathbf{X}$ lesz a szimultán egyenletrendszer megoldása.

Feladatok

1. Igazolja, hogy az $(\mathbf{A}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(m)})$ kibővített mátrixon végzett Gauss-elimináció műveletigénye $n^3/3 + mn^2 - n/3$ osztás/szorzás!
2. Igazolja, hogy az $(\mathbf{A}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(m)})$ kibővített mátrixon végzett Gauss–Jordan-elimináció műveletigénye $n^3/2 + mn^2 - n/2$ osztás/szorzás!
3. Fogalmazza át a 3.37. algoritmust szimultán tridiagonális együtthatójú egyenletrendszerek megoldására!
4. Lássa be, hogy az $\mathbf{Ax}^{(i)} = \mathbf{b}^{(i)}$, $i = 1, 2, \ldots, m$ egyenletrendszer ekvivalens az $\mathbf{AX} = \mathbf{B}$ mátrix egyenlettel, ahol $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)})$ és $\mathbf{B} = (\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(m)})$!

3.7. Mátrix invertálás és determináns számítás

Az $\mathbf{A}$ nemszinguláris négyzetes mátrix inverze teljesíti az $\mathbf{AA}^{-1} = \mathbf{I}$ mátrix egyenletet, ezért $\mathbf{A}^{-1}$ megoldása az $\mathbf{AX} = \mathbf{I}$ mátrix egyenletnek (azaz szimultán egyenletrendszernek). Ennek megoldására használhatjuk a Gauss–Jordan-eliminációt. Ellenőrizhető, hogy ennek műveletigénye $\frac{3}{2}n^3 + \mathcal{O}(n^2)$ osztás ill. szorzás.

3.38. példa. Invertáljuk az

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

mátrixot! A Gauss–Jordan-módszert használva:

$$\left(\begin{array}{rrrrrr} 1 & 0 & 2 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 1 & 0 \\ -2 & 0 & -1 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{rrrrrr} 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 1 & 0 \\ 0 & 0 & 3 & 2 & 0 & 1 \end{array}\right) \sim$$
$$\left(\begin{array}{rrrrrr} 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 1 & 0 \\ 0 & 0 & 3 & 2 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{rrrrrr} 1 & 0 & 0 & -1/3 & 0 & -2/3 \\ 0 & 1 & 0 & -1/3 & 1 & -2/3 \\ 0 & 0 & 3 & 2 & 0 & 1 \end{array}\right) \sim$$
$$\left(\begin{array}{rrrrrr} 1 & 0 & 0 & -1/3 & 0 & -2/3 \\ 0 & 1 & 0 & -1/3 & 1 & -2/3 \\ 0 & 0 & 1 & 2/3 & 0 & 1/3 \end{array}\right)$$

Tehát

$$\mathbf{A}^{-1} = \frac{1}{3}\left(\begin{array}{rrr} -1 & 0 & -2 \\ -1 & 3 & -2 \\ 2 & 0 & 1 \end{array}\right).$$

□

Természetesen a mátrix invertálás Gauss–Jordan-eliminációs módszerénél is használhatjuk a Gauss-eliminációnál megfogalmazott részleges főelemkiválasztás módszerét is a numerikus hiba csökkentése, illetve a nullával való osztás elkerülése érdekében.

A 3.26. tétel szerint az $\mathbf{A}$ mátrixon a Gauss-elimináció részleges főelemkiválasztással pontosan akkor hajtható végre, ha $\det(\mathbf{A}) \neq 0$. A tétel bizonyításából következik, hogy $\det(\mathbf{A}) = (-1)^s \det(\mathbf{A}^{(n-1)})$, ahol s a módszer közben végrehajtott sorcserék száma. Azaz a determináns egyenlő a főelemek megfelelő előjellel vett szorzatával: $\det(\mathbf{A}) = (-1)^s a_{11} a_{22}^{(1)} \cdots a_{nn}^{(n-1)}$.

3.39. példa. Tekintsük a 3.22. példa együtthatómátrixát, azaz legyen

$$\mathbf{A} = \left(\begin{array}{rrrr} 1 & -2 & -2 & -2 \\ 2 & -1 & 2 & 4 \\ -1 & 2 & 3 & -4 \\ -2 & 1 & 4 & -2 \end{array}\right).$$

Számítsuk ki a mátrix determinánsát! A 3.22. példában végigszámoltuk, hogy az $\mathbf{A}$ mátrixon végrehajtva a Gauss-eliminációt a végeredmény

$$\mathbf{A}^{(3)} = \left(\begin{array}{rrrr} 1 & -2 & -2 & -2 \\ 0 & 3 & 6 & 8 \\ 0 & 0 & 1 & -6 \\ 0 & 0 & 0 & 38 \end{array}\right).$$

Tehát $\det(\mathbf{A}) = \det(\mathbf{A}^{(3)}) = 1 \cdot 3 \cdot 1 \cdot 38 = 114$. □

Feladatok

1. Invertálja a következő mátrixokat:

$$(a) \quad \left(\begin{array}{rrr} -1 & 1 & 2 \\ -2 & 1 & 0 \\ 0 & 1 & -1 \end{array}\right) \quad (b) \quad \left(\begin{array}{rrr} -3 & 1 & 2 \\ 0 & 3 & 1 \\ -2 & -1 & 1 \end{array}\right) \quad (c) \quad \left(\begin{array}{rrrr} 1 & -1 & 0 & 2 \\ 2 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 1 & 2 & 2 & -1 \end{array}\right)$$

2. Igazolja, hogy az általános Gauss–Jordan-eliminációt használva $3n^3/2 - n/2$ osztás ill. szorzás kell a mátrix invertáláshoz!
3. Fogalmazza meg a Gauss–Jordan-eljárás algoritmusát a mátrix invertálás feladatára alkalmazva, figyelembe véve, hogy az $\mathbf{AX} = \mathbf{I}$ mátrix egyenletben $\mathbf{I}$ speciális alakú, azaz azt, hogy a nullával való szorzásokat nem kell végrehajtani! Lássa be, hogy az így kapott speciális Gauss–Jordan-elimináción alapuló mátrix invertálás műveletigénye n^3 osztás/szorzás!

4. Tesztelje az előző feladatban megfogalmazott algoritmust a

$$\begin{pmatrix} -2 & 1 & & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & 1 & -2 & 1 \\ & & & & & 1 & -2 \end{pmatrix}$$

   10×10-es mátrixon (ahol a hiányzó elemek nullák)! Lássa be, hogy a pontos inverz $\mathbf{A}^{-1} = (c_{ij})$, ahol $c_{ij} = c_{ji}$, és $c_{ij} = -i(11-j)/11$, $i \leq j$.
5. Számítsa ki az 1. feladatban megadott mátrixok determinánsát Gauss-eliminációt használva!

4. fejezet

Lineáris egyenletrendszerek megoldása iterációs módszerekkel

Ebben a fejezetben először a lineáris fixpont iteráció általános elméletét vizsgáljuk, majd ennek segítségével a lineáris egyenletrendszerek megoldásának néhány iterációs módszerét tárgyaljuk (Jacobi-, Gauss–Seidel-, relaxációs módszerek). A fejezet végén bevezetjük a mátrix kondíciószámának fogalmát, és a lineáris egyenletrendszerek perturbációjával foglalkozunk.

4.1. Lineáris fixpont iteráció

Ebben a szakaszban az

$$\mathbf{x}^{(k+1)} = \mathbf{T}\mathbf{x}^{(k)} + \mathbf{c}, \qquad k = 0, 1, 2, \ldots \tag{4.1}$$

lineáris n-dimenziós fixpont iterációval foglalkozunk. Először tekintsük a $\mathbf{c} = \mathbf{0}$ speciális esetet. Ekkor könnyen látható, hogy $\mathbf{x}^{(k)} = \mathbf{T}^k\mathbf{x}^{(0)}$ minden $k = 1, 2, \ldots$-re.

4.1. tétel. *A következő állítások ekvivalensek:*

1. $\lim_{k\to\infty} \mathbf{T}^k = \mathbf{0}$ *(zéró mátrix), azaz* $\lim_{k\to\infty} \|\mathbf{T}^k\| = 0$ *minden* $\|\cdot\|$ *mátrixnormára;*
2. $\lim_{k\to\infty} \mathbf{T}^k\mathbf{x} = \mathbf{0}$ *(zéró vektor) minden* $\mathbf{x} \in \mathbb{R}^n$*-re, azaz* $\lim_{k\to\infty} \|\mathbf{T}^k\mathbf{x}\| = 0$ *minden* $\mathbf{x} \in \mathbb{R}^n$*-re és minden* $\|\cdot\|$ *vektornormára;*
3. $\rho(\mathbf{T}) < 1$.

Bizonyítás. Az 1. állításból következik 2, mivel a mátrixnorma tulajdonsága alapján

$$\|\mathbf{T}^k\mathbf{x}\| \le \|\mathbf{T}^k\|\|\mathbf{x}\|$$

minden $x \in \mathbb{R}^n$-re és minden $\|\cdot\|$ normára.

Tegyük fel most, hogy 2. teljesül. Legyen λ egy tetszőleges sajátértéke $\mathbf{T}$-nek, és legyen $\mathbf{v}$ egy λ-hoz tartozó sajátérték. Ekkor $\mathbf{T}^k\mathbf{v} = \lambda^k\mathbf{v}$, ezért a $\mathbf{T}^k\mathbf{v} \to \mathbf{0}$ (ha $k \to \infty$) feltételből következik, hogy $|\lambda| < 1$, hiszen $\mathbf{v} \neq \mathbf{0}$. Mivel λ tetszőleges sajátérték volt, ezért $\rho(\mathbf{T}) < 1$ is teljesül.

Most tegyük fel, hogy 3. teljesül. A 3.17. tétel szerint létezik olyan $\|\cdot\|$ mátrixnorma és olyan $\varepsilon > 0$ szám, hogy $\|\mathbf{T}\| \le \rho(\mathbf{T}) + \varepsilon < 1$. Ekkor

$$\|\mathbf{T}^k\| \le \|\mathbf{T}\|^k \le (\rho(\mathbf{T}) + \varepsilon)^k \to 0,$$

ha $k \to \infty$. De ekkor a 2.47. tétel szerint $\|\mathbf{T}^k\| \to 0$ minden $\|\cdot\|$ mátrixnormában, azaz 1. teljesül. □

A következő tétel szerint $\|\mathbf{T}\| < 1$ elegendő feltétele annak, hogy $\|\mathbf{T}^k\| \to 0$ teljesüljön.

4.2. tétel. *Ha* $\|\mathbf{T}\| < 1$ *valamely* $\|\cdot\|$ *mátrixnormában, akkor* $\|\mathbf{T}^k\| \to 0$, *ha* $k \to \infty$.

Bizonyítás. Az állítás következik a $\|\mathbf{T}^k\| \le \|\mathbf{T}\|^k$ egyenlőtlenségből. □

Szükségünk lesz az $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots$ ún. *geometriai sor* vagy *Neumann-sor* konvergenciájának vizsgálatára.

4.3. tétel. *Ha* $\rho(\mathbf{A}) < 1$, *akkor az* $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots$ *végtelen mátrix sor konvergens, az* $\mathbf{I} - \mathbf{A}$ *mátrix invertálható, és*

$$(\mathbf{I} - \mathbf{A})^{-1} = \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots .$$

Fordítva, ha az $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots$ *geometriai sor konvergens, akkor* $\rho(\mathbf{A}) < 1$.

Bizonyítás. Legyen $\rho(\mathbf{A}) < 1$. Tegyük fel, hogy $\mathbf{I} - \mathbf{A}$ nem invertálható. Ekkor a 3.3. tétel szerint létezik olyan $\mathbf{x} \neq \mathbf{0}$ vektor, hogy $(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{0}$. Ezt átrendezve kapjuk, hogy $\mathbf{A}\mathbf{x} = \mathbf{x}$, azaz 1 sajátértéke $\mathbf{A}$-nak, ami ellentmond a feltevésnek, hogy $\rho(\mathbf{A}) < 1$. Tehát az $\mathbf{I} - \mathbf{A}$ mátrix invertálható.

Könnyű belátni az

$$(\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots + \mathbf{A}^m) = \mathbf{I} - \mathbf{A}^{m+1} \tag{4.2}$$

azonosságot. Ebből

$$\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots + \mathbf{A}^m = (\mathbf{I} - \mathbf{A})^{-1}(\mathbf{I} - \mathbf{A}^{m+1}),$$

és így, használva, hogy a 4.1. tétel szerint $\mathbf{A}^{m+1} \to 0$, kapjuk, hogy

$$\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots + \mathbf{A}^m \to (\mathbf{I} - \mathbf{A})^{-1},$$

ha $m \to \infty$.

Tegyük fel most, hogy az $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots$ geometriai sor konvergens. Ekkor könnyen igazolható, hogy $\mathbf{A}^m \to \mathbf{0}$, ezért a 4.1. tétel szerint $\rho(\mathbf{A}) < 1$. □

4.4. következmény. *Ha* $\|\mathbf{A}\| < 1$ *valamely* $\|\cdot\|$ *mátrixnormában, akkor* $\mathbf{I} - \mathbf{A}$ *invertálható, az* $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots$ *geometriai sor konvergens,* $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots = (\mathbf{I} - \mathbf{A})^{-1}$, *valamint*

$$\|(\mathbf{I} - \mathbf{A})^{-1}\| \le \frac{1}{1 - \|\mathbf{A}\|}.$$

Bizonyítás. Csak az utolsó állítást kell belátnunk, a többi rögtön következik a 4.3. és 3.16. tételekből. A mátrixnorma folytonosságát, a háromszög-egyenlőtlenséget és a norma tulajdonságait használva:

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{A})^{-1}\| &= \|\lim_{m\to\infty}(\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots + \mathbf{A}^m)\| \\
&= \lim_{m\to\infty}\|\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots + \mathbf{A}^m\| \\
&\le \lim_{m\to\infty}(\|\mathbf{I}\| + \|\mathbf{A}\| + \|\mathbf{A}^2\| + \|\mathbf{A}^3\| + \cdots + \|\mathbf{A}^m\|) \\
&\le \lim_{m\to\infty}(1 + \|\mathbf{A}\| + \|\mathbf{A}\|^2 + \|\mathbf{A}\|^3 + \cdots + \|\mathbf{A}\|^m) \\
&= \frac{1}{1 - \|\mathbf{A}\|}.
\end{aligned}$$

□

Az előző eredménynek van egy fontos következménye: ha $\mathbf{A}$ nemszinguláris mátrix, akkor az $\mathbf{A}$-hoz „közeli” mátrixok is nemszingulárisak.

4.5. tétel. *Legyenek* $\mathbf{A}$ *és* $\mathbf{B}$ $n \times n$*-es mátrixok. Legyen* $\mathbf{A}$ *nemszinguláris, és*

$$\|\mathbf{A}-\mathbf{B}\| < \frac{1}{\|\mathbf{A}^{-1}\|}.$$

Ekkor $\mathbf{B}$ *is nemszinguláris, továbbá*

$$\|\mathbf{B}^{-1}\| \le \frac{\|\mathbf{A}^{-1}\|}{1-\|\mathbf{A}^{-1}\|\|\mathbf{A}-\mathbf{B}\|} \tag{4.3}$$

és

$$\|\mathbf{A}^{-1}-\mathbf{B}^{-1}\| \le \frac{\|\mathbf{A}^{-1}\|^2\|\mathbf{A}-\mathbf{B}\|}{1-\|\mathbf{A}^{-1}\|\|\mathbf{A}-\mathbf{B}\|}. \tag{4.4}$$

Bizonyítás. Induljunk ki a $\mathbf{B} = \mathbf{A}-(\mathbf{A}-\mathbf{B}) = \mathbf{A}(\mathbf{I}-\mathbf{A}^{-1}(\mathbf{A}-\mathbf{B}))$ azonosságból. A feltétel szerint $\|\mathbf{A}^{-1}(\mathbf{A}-\mathbf{B})\| \le \|\mathbf{A}^{-1}\|\|\mathbf{A}-\mathbf{B}\| < 1$, ezért a 4.4. következmény szerint $\mathbf{I}-\mathbf{A}^{-1}(\mathbf{A}-\mathbf{B})$ invertálható. De ekkor $\mathbf{B}^{-1} = (\mathbf{I}-\mathbf{A}^{-1}(\mathbf{A}-\mathbf{B}))^{-1}\mathbf{A}^{-1}$ létezik. Ebből és az $\mathbf{A}^{-1}-\mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\mathbf{B}^{-1}$ azonosságból, valamint a 4.4. következményből kapjuk a (4.3) és (4.4) becsléseket. □

Térjünk most vissza a (4.1) fixpont feladathoz. Vizsgáljuk most az általános esetet. Könnyen látható, hogy a fixpont sorozat k-adik tagjának általános képlete

$$\mathbf{x}^{(k)} = \mathbf{T}^k\mathbf{x}^{(0)} + (\mathbf{T}^{k-1}+\mathbf{T}^{k-2}+\cdots+\mathbf{T}+\mathbf{I})\mathbf{c}, \qquad k = 1,2,\ldots.$$

A 4.1. és 4.3. tételekből kapjuk:

4.6. tétel. *Legyen* $\mathbf{c} \neq \mathbf{0}$*. Ekkor az* $\mathbf{x} = \mathbf{T}\mathbf{x}+\mathbf{c}$ *egyenletnek létezik egyértelmű megoldása, és a (4.1) iterációs sorozat akkor és csak akkor konvergál az egyenlet megoldásához minden* $\mathbf{x}^{(0)}$ *kezdeti értékre, ha* $\rho(\mathbf{T}) < 1$*.*

Bizonyítás. Legyen $\rho(\mathbf{T}) < 1$. Ekkor a 4.3. tétel szerint $\mathbf{I}-\mathbf{T}$ invertálható, ezért az $\mathbf{x} = \mathbf{T}\mathbf{x}+\mathbf{c}$ egyenletnek létezik egyértelmű megoldása: $\mathbf{x} = (\mathbf{I}-\mathbf{T})^{-1}\mathbf{c}$. A 4.1. és 4.3. tételekből következik, hogy $\mathbf{T}^k\mathbf{x}^{(0)} \to 0$ minden $\mathbf{x}^{(0)} \in \mathbb{R}^n$-re, és $(\mathbf{T}^{k-1}+\mathbf{T}^{k-2}+\cdots+\mathbf{T}+\mathbf{I})\mathbf{c} \to (\mathbf{I}-\mathbf{T})^{-1}\mathbf{c}$, ha $k \to \infty$.

Fordítva, legyen $\mathbf{x}$ az $\mathbf{x} = \mathbf{T}\mathbf{x}+\mathbf{c}$ egyenlet megoldása, és tegyük fel, hogy $\mathbf{x}^{(k)} \to \mathbf{x}$, ha $k \to \infty$. Ekkor az $\mathbf{x} = \mathbf{T}\mathbf{x}+\mathbf{c}$ és $\mathbf{x}^{(k+1)} = \mathbf{T}\mathbf{x}^{(k)}+\mathbf{c}$ egyenleteket egymásból kivonva $\mathbf{x}-\mathbf{x}^{(k+1)} = \mathbf{T}(\mathbf{x}-\mathbf{x}^{(k)})$, és így

$$\mathbf{x}-\mathbf{x}^{(k+1)} = \mathbf{T}(\mathbf{x}-\mathbf{x}^{(k)}) = \cdots = \mathbf{T}^{k+1}(\mathbf{x}-\mathbf{x}^{(0)}). \tag{4.5}$$

Legyen $\mathbf{z}$ egy tetszőleges vektor, és $\mathbf{x}^{(0)} = \mathbf{x}-\mathbf{z}$. Ekkor

$$\lim_{k\to\infty}\mathbf{T}^{k+1}\mathbf{z} = \lim_{k\to\infty}\mathbf{T}^{k+1}(\mathbf{x}-\mathbf{x}^{(0)}) = \lim_{k\to\infty}(\mathbf{x}-\mathbf{x}^{(k+1)}) = \mathbf{x}-\mathbf{x} = \mathbf{0}.$$

Alkalmazva a 4.1. tételt, kapjuk, hogy $\rho(\mathbf{T}) < 1$. □

4.7. következmény. *Ha* $\|\mathbf{T}\| < 1$ *valamely* $\|\cdot\|$ *mátrixnormában, akkor a (4.1) iterációs sorozat konvergens minden* $\mathbf{x}^{(0)}$ *kezdeti értékre, és*

$$\|\mathbf{x}-\mathbf{x}^{(k)}\| \le \|\mathbf{T}\|^k\|\mathbf{x}-\mathbf{x}^{(0)}\|. \tag{4.6}$$

A (4.6) egyenlőségből következik, hogy $\mathbf{x}^{(k)}$ annál gyorsabban konvergál, minél kisebb $\|\mathbf{T}\|$. A 3.17. tétel alapján ebből következik, hogy $\mathbf{x}^{(k)}$ annál gyorsabban konvergál (egy bizonyos normában), minél kisebb $\rho(\mathbf{T})$.

A továbbiakban vizsgáljuk a kerekítési hibák hatását a lineáris fixpont sorozat tagjaira. Tegyük fel, hogy a (4.1) sorozat helyett a

$$\begin{aligned} \mathbf{y}^{(k+1)} &= \mathbf{T}\mathbf{y}^{(k)} + \mathbf{c} + \mathbf{w}^{(k+1)}, \qquad k = 0, 1, \ldots, & (4.7)\\ \mathbf{y}^{(0)} &= \mathbf{x}^{(0)} + \mathbf{w}^{(0)} & (4.8) \end{aligned}$$

sorozatot generáljuk, ahol $\mathbf{w}^{(k+1)}$ reprezentálja a k-adik lépésben elkövetett kerekítési hibát, $\mathbf{w}^{(0)}$ pedig a kezdeti érték tárolásakor fellépő kerekítési hiba. Feltesszük, hogy a

$$\|\mathbf{w}^{(k)}\| \le \varepsilon, \qquad k = 0, 1, \ldots$$

becslés teljesül valamilyen vektornormában. Képezzük a (4.7) és (4.1) egyenletek különbségét:

$$\mathbf{y}^{(k+1)} - \mathbf{x}^{(k+1)} = \mathbf{T}(\mathbf{y}^{(k)} - \mathbf{x}^{(k)}) + \mathbf{w}^{(k+1)}.$$

Ekkor

$$\begin{aligned} \|\mathbf{y}^{(k+1)} - \mathbf{x}^{(k+1)}\| &\le \|\mathbf{T}(\mathbf{y}^{(k)} - \mathbf{x}^{(k)})\| + \|\mathbf{w}^{(k+1)}\| \\ &\le \|\mathbf{T}\|\|\mathbf{y}^{(k)} - \mathbf{x}^{(k)}\| + \varepsilon \\ &\vdots \\ &\le \|\mathbf{T}\|^{k+1}\|\mathbf{y}^{(0)} - \mathbf{x}^{(0)}\| + (\|\mathbf{T}\|^k + \cdots \|\mathbf{T}\| + 1)\varepsilon \\ &\le (\|\mathbf{T}\|^{k+1} + \|\mathbf{T}\|^k + \cdots \|\mathbf{T}\| + 1)\varepsilon. \end{aligned}$$

Ha $\|\mathbf{T}\| < 1$, akkor a legutolsó kifejezés tovább becsülhető a végtelen mértani sor összegével:

$$\|\mathbf{y}^{(k+1)} - \mathbf{x}^{(k+1)}\| \le \frac{1}{1 - \|\mathbf{T}\|}\varepsilon.$$

Ebből látható, hogy a számolás stabil a kerekítési hibára nézve, és a számolás közben fellépő kerekítési hiba annál kisebb, minél közelebb van $\|\mathbf{T}\|$ nullához.

Feladatok

1. Számítsa ki az $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \cdots$ geometriai sor értékét, ha

$$\text{(a)} \quad \mathbf{A} = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \text{(b)} \quad \mathbf{A} = \begin{pmatrix} 1/2 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/5 \end{pmatrix}.$$

2. Igazolja a (4.2) azonosságot!
3. Dolgozza ki a (4.3) és (4.4) egyenlőtlenségek bizonyításának részleteit!
4. Adja meg az összes α paraméterértéket, amelyre az

$$\begin{pmatrix} 1 & 2 \\ \alpha & 0 \end{pmatrix}^k$$

mátrixsorozat a $\mathbf{0}$ mátrixhoz konvergál!

4.2. Jacobi-iteráció

4.8. példa. Oldjuk meg a

$$\begin{array}{rcrcrcr} 5x_1 & + & 3x_2 & - & x_3 & = & -4 \\ 2x_1 & - & 10x_2 & + & x_3 & = & 25 \\ -3x_1 & + & 4x_2 & - & 12x_3 & = & -47. \end{array} \tag{4.9}$$

egyenletrendszert! Fejezzük ki az első egyenletből x_1-et, a másodikból x_2-t, a harmadikból pedig x_3-at:

$$\begin{aligned} x_1 &= (-4 - 3x_2 + x_3)/5 \\ x_2 &= (-25 + 2x_1 + x_3)/10 \\ x_3 &= (47 - 3x_1 + 4x_2)/12. \end{aligned} \tag{4.10}$$

A (4.10) egyenletrendszer egy lineáris háromdimenziós fixpont egyenlet, ezért definiáljuk a következő iterációs módszert $k = 0, 1, 2, \ldots$-re:

$$\begin{aligned} x_1^{(k+1)} &= (-4 - 3x_2^{(k)} + x_3^{(k)})/5 \\ x_2^{(k+1)} &= (-25 + 2x_1^{(k)} + x_3^{(k)})/10 \\ x_3^{(k+1)} &= (47 - 3x_1^{(k)} + 4x_2^{(k)})/12 \end{aligned} \tag{4.11}$$

A 4.1. táblázat az $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$ kezdeti értékekből számolt numerikus értékeket tartalmazza. Megfigyelhetjük, hogy erre a kezdeti értékre az iterációs sorozat konvergens, és a határértéke $x_1 = 1$, $x_2 = -2$, $x_3 = 3$, ami a (4.9) egyenletrendszer megoldása. A (4.11) iteráció röviden az

$$\mathbf{x}^{(k+1)} = \mathbf{T}\mathbf{x}^{(k)} + \mathbf{c} \tag{4.12}$$

alakban írható fel, ahol

$$\mathbf{T} = \begin{pmatrix} 0 & -3/5 & 1/5 \\ 2/10 & 0 & 1/10 \\ -3/12 & 4/12 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} -4/5 \\ -25/10 \\ 47/12 \end{pmatrix}.$$

A 4.7. következmény szerint a (4.12) iteráció konvergens, ha a $\mathbf{T}$ mátrix valamely mátrixnormája kisebb mint 1. Mivel $\|\mathbf{T}\|_\infty = \max\{4/5, 3/10, 7/12\} = 4/5 < 1$, ezért a (4.11) iteráció valóban konvergens. □

4.1. táblázat. Jacobi-iteráció

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000000	0.000000	0.000000
1	-0.800000	-2.500000	3.916667
2	1.483333	-2.268333	3.283333
3	1.217667	-1.875000	2.789722
4	0.882944	-1.977494	2.987250
⋮	⋮	⋮	⋮
14	0.999999	-1.999992	2.999990
15	0.999993	-2.000001	3.000003
16	1.000001	-2.000001	3.000001
17	1.000001	-2.000000	2.999999
18	1.000000	-2.000000	3.000000

Tekintsük az általános

$$\begin{array}{rcrcccrcl} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \ldots & + & a_{nn}x_n & = & b_n \end{array} \tag{4.13}$$

egyenletet. Ha $a_{ii} \neq 0$ minden $i = 1, \ldots, n$-re, akkor a (4.13) egyenletet átírhatjuk az

$$x_i = -\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{a_{ij}}{a_{ii}} x_j + \frac{b_i}{a_{ii}}, \qquad i = 1, \ldots, n \tag{4.14}$$

alakba, és definiálhatjuk az ún. *Jacobi-iterációt* $k = 0, 1, 2, \ldots$-re:

$$x_i^{(k+1)} = -\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{b_i}{a_{ii}}, \qquad i = 1, \ldots, n. \tag{4.15}$$

Ha $a_{ii} = 0$ valamely i-re, akkor megpróbáljuk sorcserékkel elérni, hogy $a_{ii} \neq 0$ legyen $i = 1, \ldots, n$-re. Vezessük be a következő jelölést: $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$, ahol

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_{21} & 0 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & 0 \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n} \\ 0 & 0 & 0 & \cdots & a_{3n} \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

és $\mathbf{D} = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$. $\mathbf{L}$ és $\mathbf{U}$ alulról ill. felülről trianguláris mátrixok (amelyeknek a főátlója is zéró). Ezzel a jelöléssel az $\mathbf{Ax} = \mathbf{b}$ egyenletrendszert a $\mathbf{Dx} = -(\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{b}$ alakra írjuk, majd beszorozzuk az egyenletet balról $\mathbf{D}^{-1}$-zel. Ennélfogva a Jacobi-iteráció a (4.12) képlettel definiálható, ahol $\mathbf{T} = \mathbf{T}_J \equiv -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$, és $\mathbf{c} = \mathbf{D}^{-1}\mathbf{b}$.

A 4.6. tétel és a 4.7. következményből rögtön kapjuk a Jacobi-iteráció konvergenciájára vonatkozó szükséges és elegendő, ill. elegendő feltételeket:

4.9. tétel. *A Jacobi-iteráció akkor és csak akkor konvergens, ha* $\rho(\mathbf{T}_J) < 1$.

4.10. következmény. *Ha* $\|\mathbf{T}_J\| < 1$ *valamely* $\|\cdot\|$ *mátrixnormában, akkor a Jacobi-iteráció konvergens bármely* $\mathbf{x}^{(0)}$ *kezdeti értékre.*

A gyakorlatban sokszor egyszerűen alkalmazható a következő tétel.

4.11. tétel. *Ha* $\mathbf{A}$ *diagonálisan domináns, akkor a Jacobi-iteráció konvergens bármely* $\mathbf{x}^{(0)}$ *kezdeti értékre.*

Bizonyítás. Mivel

$$\mathbf{T}_J = \begin{pmatrix} 0 & -a_{12}/a_{11} & -a_{13}/a_{11} & \cdots & -a_{1n}/a_{11} \\ -a_{21}/a_{22} & 0 & -a_{23}/a_{22} & \cdots & -a_{2n}/a_{22} \\ -a_{31}/a_{33} & -a_{32}/a_{33} & 0 & \cdots & -a_{3n}/a_{33} \\ \vdots & & & \ddots & \vdots \\ -a_{n1}/a_{nn} & -a_{n2}/a_{nn} & -a_{n3}/a_{nn} & \cdots & 0 \end{pmatrix},$$

ezért az $\mathbf{A}$ mátrix diagonális dominanciáját használva

$$\|\mathbf{T}_J\|_\infty = \max_{i=1,\ldots,n} \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right\} < 1,$$

amiből, a 4.10. következmény szerint kapjuk az állítást. □

Feladatok

1. A Jacobi-iterációt használva oldja meg a következő egyenletrendszereket:

(a)
$$\begin{array}{rcrcrcr} 6.2x_1 & + & 1.1x_2 & - & 3.4x_3 & = & 5.1 \\ -0.6x_1 & + & 2.9x_2 & + & 0.3x_3 & = & -7.2 \\ 1.1x_1 & - & 0.6x_2 & + & 4.4x_3 & = & 3.1 \end{array}$$

(b)
$$\begin{array}{rcrcrcrcr} -8x_1 & + & 3x_2 & - & 2x_3 & & & = & 6 \\ 2x_1 & + & 6x_2 & + & x_3 & - & 2x_4 & = & 3 \\ 3x_1 & - & 3x_2 & + & 10x_3 & + & x_4 & = & 5 \\ & & x_2 & - & 3x_3 & + & 7x_4 & = & -17 \end{array}$$

2. Mutassa meg, hogy a Jacobi-iteráció konvergens tetszőleges kezdeti értékre, ha $\mathbf{A}$ oszloponként diagonálisan domináns!

4.3. Gauss–Seidel-iteráció

4.12. példa. Tekintsük újra a (4.9) egyenletet és annak (4.10) alakját! Definiáljuk az

$$\begin{aligned} x_1^{(k+1)} &= (-4 - 3x_2^{(k)} + x_3^{(k)})/5 \\ x_2^{(k+1)} &= (-25 + 2x_1^{(k+1)} + x_3^{(k)})/10 \\ x_3^{(k+1)} &= (47 - 3x_1^{(k+1)} + 4x_2^{(k+1)})/12. \end{aligned} \tag{4.16}$$

iterációt! Az a különbség a (4.11) és (4.16) definíciók között, hogy ennél a módszernél amikor egy x_i változónak már kiszámoltuk az új értékét a k+1-edik iterációban, akkor ezt az új értéket már felhasználjuk a következő változó számításához: x_1 $k+1$-edik értékét az első egyenlettel számoljuk, az x_2 új értékének számításához már az x_1 új értékét (ami várhatóan jobb közelítése a megoldásnak mint $x_1^{(k)}$) használjuk a második egyenletben $x_3^{(k)}$-val együtt, mivel annak még nem számoltunk új értéket. A 4.2. táblázatban található a módszernek az $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$ kezdeti értékekhez tartozó numerikus eredménye. Láthatjuk, hogy ez az iterációs módszer gyorsabban konvergál ezen a feladaton mint a Jacobi-iteráció. □

4.2. táblázat. Gauss–Seidel-iteráció

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000000	0.000000	0.000000
1	-0.800000	-2.660000	3.230000
2	1.442000	-1.888600	2.926633
3	0.918487	-2.023639	3.012499
4	1.016683	-1.995413	2.997358
5	0.996720	-2.000920	3.000513
6	1.000655	-1.999818	2.999897
7	0.999870	-2.000036	3.000020
8	1.000026	-1.999993	2.999996
9	0.999995	-2.000001	3.000001
10	1.000001	-2.000000	3.000000
11	1.000000	-2.000000	3.000000

A (4.13) általános lineáris egyenletrendszer megoldására definiáljuk a *Gauss–Seidel-iterációt* $k = 0, 1, 2, \ldots$-re (ha $a_{ii} \neq 0$ minden $i = 1, \ldots, n$-re):

$$x_i^{(k+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(k+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{b_i}{a_{ii}}, \qquad i = 1, \ldots, n. \tag{4.17}$$

A (4.17) egyenletet átrendezhetjük a következő alakba:

$$\sum_{j=1}^{i} a_{ij}x_j^{(k+1)} = -\sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i, \qquad i = 1, \ldots, n,$$

azaz mátrix jelöléssel

$$(\mathbf{D}+\mathbf{L})\mathbf{x}^{(k+1)} = -\mathbf{U}\mathbf{x}^{(k)} + \mathbf{b},$$

ahol $\mathbf{L}$, $\mathbf{D}$, $\mathbf{U}$ ugyanaz, mint az előző szakaszban. Innen látható, hogy a Gauss–Seidel-iteráció is felírható a (4.12) alakban a $\mathbf{T} = \mathbf{T}_G \equiv -(\mathbf{D}+\mathbf{L})^{-1}\mathbf{U}$ és $\mathbf{c} = (\mathbf{D}+\mathbf{L})^{-1}\mathbf{b}$ választással.

Alkalmazva a 4.6. tételt és annak 4.7. következményét rögtön kapjuk:

4.13. tétel. *A Gauss–Seidel-iteráció akkor és csak akkor konvergens, ha* $\rho(\mathbf{T}_G) < 1$.

4.14. következmény. *Ha* $\|\mathbf{T}_G\| < 1$ *valamely* $\|\cdot\|$ *mátrixnormában, akkor a Gauss–Seidel-iteráció konvergens bármely* $\mathbf{x}^{(0)}$ *kezdeti értékre.*

Megmutatható, hogy a Jacobi-iterációhoz hasonlóan diagonálisan domináns mátrixokra a Gauss–Seidel-módszer is konvergens.

4.15. tétel. *Ha* $\mathbf{A}$ *diagonálisan domináns, akkor a Gauss–Seidel-iteráció konvergens bármely* $\mathbf{x}^{(0)}$ *kezdeti értékre.*

Bizonyítás. Jelölje $\mathbf{x} = (x_1, \ldots, x_n)^T$ a (4.13) egyenlet pontos megoldását. Ekkor a (4.13) egyenletrendszer i-edik egyenletéből x_i-t kifejezve és a kapott egyenletel kivonva a (4.17) egyenletből, kapjuk, hogy

$$x_i^{(k+1)} - x_i = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}}(x_j^{(k+1)} - x_j) - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}}(x_j^{(k)} - x_j).$$

Ebből következik, hogy

$$|x_i^{(k+1)} - x_i| \le \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| |x_j^{(k+1)} - x_j| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| |x_j^{(k)} - x_j|. \tag{4.18}$$

Legyen

$$\alpha_i \equiv \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| \qquad \text{és} \qquad \beta_i \equiv \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|.$$

Ezzel a jelöléssel a (4.18) egyenlőtlenségből kapjuk, hogy

$$|x_i^{(k+1)} - x_i| \le \alpha_i \|\mathbf{x}^{(k+1)} - \mathbf{x}\|_\infty + \beta_i \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty$$

teljesül minden $i = 1, \ldots, n$-re. Legyen l egy olyan index, amelyre $|x_l^{(k+1)} - x_l| = \|\mathbf{x}^{(k+1)} - \mathbf{x}\|_\infty$. Ekkor

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}\|_\infty \le \alpha_l \|\mathbf{x}^{(k+1)} - \mathbf{x}\|_\infty + \beta_l \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty.$$

$\mathbf{A}$ diaginálisan domináns, ezért $\alpha_l < 1$, és így

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}\|_\infty \le \frac{\beta_l}{1-\alpha_l} \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty.$$

Kapjuk tehát, hogy

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty \leq \left(\max_{l=1,\dots,n} \frac{\beta_l}{1-\alpha_l} \right)^k \|\mathbf{x}^{(0)} - \mathbf{x}\|_\infty.$$

Ebből következik, hogy a Gauss-Seidel módszer konvergens, hiszen a diagonális dominancia alapján könnyen ellenőrizhető, hogy

$$\frac{\beta_l}{1-\alpha_l} \leq \alpha_l + \beta_l < 1$$

teljesül minden $l = 1, \dots, n$-re, és ebből

$$\max_{l=1,\dots,n} \frac{\beta_l}{1-\alpha_l} \leq \max_{l=1,\dots,n} \{\alpha_l + \beta_l\} = \|\mathbf{T}_J\|_\infty < 1 \tag{4.19}$$

is következik. □

A (4.19) egyenlőtlenségből következik az is, hogy diagonálisan domináns mátrixok esetében a Gauss–Seidel-módszerre jobb hibabecslést tudunk adni, mint a Jacobi-iterációra, tehát várhatóan legalább olyan gyorsan konvergál, mint a Jacobi-iteráció. Az általános esetben az, hogy a Jacobi- vagy a Gauss–Seidel-iteráció konvergál-e gyorsabban, attól függ, hogy $\rho(\mathbf{T}_J)$ vagy $\rho(\mathbf{T}_G)$ kisebb-e. Ennek eldöntésére, az $\mathbf{A}$ mátrix együtthatói ismeretében, nem ismert egyszerű feltétel. Egy speciális esetre vonatkozik a következő tétel, amelyet bizonyítás nélkül közlünk.

4.16. tétel (Stein–Rosenberg). *Tegyük fel, hogy $a_{ij} \leq 0$ ha $i \neq j$ és $a_{ii} > 0$ minden $i = 1, \dots, n$-re. Ekkor a következő állítások közül pontosan egy teljesül:*

1. $0 \leq \rho(\mathbf{T}_G) < \rho(\mathbf{T}_J) < 1$,
2. $1 < \rho(\mathbf{T}_J) < \rho(\mathbf{T}_G)$,
3. $\rho(\mathbf{T}_J) = \rho(\mathbf{T}_G) = 0$,
4. $\rho(\mathbf{T}_J) = \rho(\mathbf{T}_G) = 1$.

A tételből következik, hogy a feltételeknek eleget tevő együtthatómátrixú egyenletrendszerek esetében a Jacobi-iteráció pontosan akkor konvergens, amikor a Gauss–Seidel-iteráció, és a Gauss–Seidel-iteráció mindig gyorsabban konvergál. Általában viszont nem igaz, hogy ha a Gauss–Seidel-iteráció konvergens, akkor a Jacobi is az, vagy fordítva.

Feladatok

1. A Gauss–Seidel-iterációt használva oldja meg az előző szakasz 1. feladatában megadott egyenletrendszereket!
2. Mutassa meg, hogy a Jacobi- és a Gauss–Seidel-iteráció is véges sok lépésben megadja az egyenlet pontos gyökét, feltéve, hogy $\mathbf{A}$ felülről trianguláris és $a_{ii} \neq 0$ $i = 1, \dots, n$-re!

4.4. Hibabecslés, iteratív finomítás

Az előző szakaszokban megismert iterációs módszerek megállási feltételei hasonlóak egy általános iterációs sorozat megállási feltételeihez. A 2.8. szakaszban tárgyalt feltételek mintájára három általános megállási feltétel valamelyikét, ill. ezek kombinációit használhatjuk:

$$1.\quad \|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|<\varepsilon,\quad 2.\quad \frac{\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|}{\|\mathbf{x}^{(k+1)}\|}<\varepsilon,\quad \text{és}\quad 3.\quad \|\mathbf{b}-\mathbf{A}\mathbf{x}^{(k)}\|<\varepsilon.$$

Ez utóbbi feltétellel foglalkozunk ebben a szakaszban.

Az $\mathbf{r}\equiv\mathbf{b}-\mathbf{A}\tilde{\mathbf{x}}$ vektort az $\tilde{\mathbf{x}}$ közelítő megoldáshoz tartozó *reziduális vektornak* nevezzük. A 3. feltétel azon a hipotézisen alapszik, hogy ha $\mathbf{r}$ normája kicsi, akkor $\tilde{\mathbf{x}}$ jó közelítése a pontos megoldásnak. Azt, hogy ez a hipotézis nem minden esetben igaz, az alábbi példa mutatja.

4.17. példa. A

$$\begin{pmatrix} 4 & 1\\ 4.03 & 1\end{pmatrix}\begin{pmatrix} x_1\\ x_2\end{pmatrix}=\begin{pmatrix} 5\\ 5.03\end{pmatrix} \tag{4.20}$$

egyenletrendszer pontos megoldása $\mathbf{x}=(1,\ 1)^T$. Tekintsük $\tilde{\mathbf{x}}=(2,\ -3)^T$-t egy „közelítő" megoldásnak. A hozzá tartozó reziduális vektor: $\mathbf{r}=\mathbf{b}-\mathbf{A}\tilde{\mathbf{x}}=(0,\ 0.03)^T$. Ennek végtelen normája $\|\mathbf{r}\|_\infty=0.03$, ami kicsi, annak ellenére, hogy $\tilde{\mathbf{x}}$ nyilván nem tekinthető a pontos megoldás jó közelítésének. □

A következő eredmény azt vizsgálja, hogy $\|\mathbf{r}\|$ kicsinységéből milyen esetekben következtethetünk arra, hogy a közelítés hibája kicsi.

4.18. tétel. *Legyen* $\mathbf{A}$ *egy nemszinguláris négyzetes mátrix,* $\mathbf{x}$ *az* $\mathbf{A}\mathbf{x}=\mathbf{b}$ *egyenlet pontos megoldása,* $\tilde{\mathbf{x}}$ *egy közelítő megoldása, és legyen* $\mathbf{r}\equiv\mathbf{b}-\mathbf{A}\tilde{\mathbf{x}}$. *Ekkor*

$$\|\mathbf{x}-\tilde{\mathbf{x}}\|\le\|\mathbf{A}^{-1}\|\|\mathbf{r}\|, \tag{4.21}$$

és

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|}\le\|\mathbf{A}\|\|\mathbf{A}^{-1}\|\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}. \tag{4.22}$$

Bizonyítás. Az $\mathbf{A}\mathbf{x}=\mathbf{b}$ és $\mathbf{A}\tilde{\mathbf{x}}=\mathbf{b}-\mathbf{r}$ összefüggésből kapjuk, hogy $\mathbf{A}(\mathbf{x}-\tilde{\mathbf{x}})=\mathbf{r}$, és így $\mathbf{x}-\tilde{\mathbf{x}}=\mathbf{A}^{-1}\mathbf{r}$. Ebből az $\|\mathbf{A}^{-1}\mathbf{r}\|\le\|\mathbf{A}^{-1}\|\|\mathbf{r}\|$ egyenlőtlenséget felhasználva következik (4.21).

A (4.21) és a $\|\mathbf{b}\|\le\|\mathbf{A}\|\|\mathbf{x}\|$ egyenlőtlenségekből

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|}\le\frac{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|\|\mathbf{r}\|}{\|\mathbf{A}\|\|\mathbf{x}\|}\le\|\mathbf{A}\|\|\mathbf{A}^{-1}\|\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}.$$

□

Az előbbi tétel ad választ a 4.17. példában is vizsgált kérdésre. Abból, hogy a közelítő megoldás reziduális vektora kicsi, akkor következik csak, hogy a közelítés relatív hibája kicsi, ha az $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ szorzat nem „túl nagy". Vezessük be a következő elnevezést: az $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ számot az $\mathbf{A}$ mátrix ($\|\cdot\|$ normára vonatkozó) *kondíciószámának* nevezzük és $\mathrm{cond}(\mathbf{A})$-val jelöljük. Megjegyezzük, hogy a kondíciószám a használt mátrixnormától függ. A $\|\cdot\|_p$ mátrixnormához tartozó kondíciószámot $\mathrm{cond}_p(\mathbf{A})$-val jelöljük. Ha egy $\mathbf{A}$ mátrix kondíciószáma „nagy", akkor a mátrixot *rosszul kondícionált*, vagy *gyengén meghatározott* mátrixnak nevezzük. Arra, hogy mekkora kell legyen a kondíciószám ahhoz, hogy rosszul kondícionált mátrixról beszéljünk, nem adunk pontos definíciót. Általában 100–1000 feletti kondíciószám esetén szokás rosszul

kondícionált mátrixról beszélni. Rosszul kondícionált mátrixokra tehát nem megbízható a 3. megállási feltétel.

4.19. példa. Tekintsük a 4.17. példa $\mathbf{A}$ együtthatómátrixát! Könnyen ellenőrizhető, hogy

$$\mathbf{A}^{-1} = \begin{pmatrix} -33.33 & 33.33 \\ 134.3 & -133.3 \end{pmatrix},$$

és így $\|\mathbf{A}\|_\infty = 5.03$, $\|\mathbf{A}^{-1}\|_\infty = 267.6$. Ebből kapjuk, hogy $\mathrm{cond}_\infty(\mathbf{A}) = 1346$. A 4.18. tétel szerint ez magyarázza azt, hogy $(2,-3)^T$ nem jó közelítése az egyenlet megoldásának, bár a reziduális vektor kicsi. □

Tegyük fel, hogy az $\mathbf{Ax} = \mathbf{b}$ egyenletet Gauss-eliminációval oldjuk meg, t-jegyű aritmetikát használva. Legyen $\tilde{\mathbf{x}}$ a kapott közelítő megoldás, amely kerekítési hibával terhelt. Számítsuk ki az $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ reziduális vektort, de az értékes számjegyek megőrzése érdekében most használjunk $2t$-jegyű aritmetikát (dupla pontosságot) $\mathbf{r}$ számolásához. Megmutatható, hogy

$$\|\mathbf{r}\| \approx 10^{-t}\|\mathbf{A}\|\|\tilde{\mathbf{x}}\|.$$

Ezt az összefüggést felhasználhatjuk $\mathbf{A}$ kondíciószámának becslésére a következőképpen: Tekintsük az $\mathbf{Ay} = \mathbf{r}$ egyenletet, és legyen $\tilde{\mathbf{y}}$ ennek numerikus megoldása t-jegyű aritmetikát használva. Megjegyezzük, hogy az $\mathbf{Ay} = \mathbf{r}$ egyenletet hatékonyan meg tudjuk oldani, ha az első Gauss-elimináció során a sorcseréket és az l_{ij} faktorokat, és a Gauss-elimináció végén kapott együtthatómátrixot megjegyezzük. Így csak az $\mathbf{r}$ vektoron kell újra eliminációt végezni, az együtthatómátrixon nem. (Az 5.1. szakaszban egy hasonlóan hatékony módszert fogunk bemutatni olyan lineáris egyenletrendszerek megoldására LU-faktorizáció segítségével, ahol az együtthatómátrix azonos.) Ekkor

$$\tilde{\mathbf{y}} \approx \mathbf{A}^{-1}\mathbf{r} = \mathbf{A}^{-1}(\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}) = \mathbf{A}^{-1}\mathbf{b} - \tilde{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{x}},$$

tehát $\|\tilde{\mathbf{y}}\|$ becslése az $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ hibának, és

$$\|\tilde{\mathbf{y}}\| \approx \|\mathbf{A}^{-1}\mathbf{r}\| \le \|\mathbf{A}^{-1}\|\|\mathbf{r}\| \approx \|\mathbf{A}^{-1}\|\|\mathbf{A}\|10^{-t}\|\tilde{\mathbf{x}}\| = 10^{-t}\mathrm{cond}(\mathbf{A})\|\tilde{\mathbf{x}}\|.$$

Ebből kapjuk, hogy a

$$\mathrm{cond}(\mathbf{A}) \approx 10^t \frac{\|\tilde{\mathbf{y}}\|}{\|\tilde{\mathbf{x}}\|} \tag{4.23}$$

képletet használhatjuk a kondíciószám becslésére. Legyen $\tilde{\mathbf{r}} = \mathbf{r} - \mathbf{A}\tilde{\mathbf{y}}$ az $\tilde{\mathbf{y}}$-hoz tartozó reziduális vektor. Általában $\|\tilde{\mathbf{r}}\|$ sokkal kisebb, mint $\|\mathbf{r}\|$, ezért ha $\tilde{\mathbf{x}}$ helyett $\bar{\mathbf{x}} \equiv \tilde{\mathbf{x}} + \tilde{\mathbf{y}}$-t tekintjük $\mathbf{x}$ közelítésének, akkor az $\bar{\mathbf{x}}$-hez tartozó reziduális vektorra

$$\|\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}\| = \|\mathbf{b} - \mathbf{A}(\tilde{\mathbf{x}} + \tilde{\mathbf{y}})\| = \|\mathbf{r} - \mathbf{A}\tilde{\mathbf{y}}\| = \|\tilde{\mathbf{r}}\| \ll \|\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}\|,$$

azaz $\bar{\mathbf{x}}$ sokkal pontosabb közelítése $\mathbf{x}$-nek, mint $\tilde{\mathbf{x}}$. Ha ezt az eljárást iterációs eljárásként ismételjük, akkor az ún. *iteratív finomítás* vagy más néven *reziduális korrekció* módszerét kapjuk. Ez a módszer rosszul kondícionált mátrixok esetén is az egyenlet megoldásának jó közelítését adja néhány lépésben.

4.20. algoritmus. Iteratív finomítás

INPUT: $\mathbf{A}$, $\mathbf{b}$
N - maximális iterációszám
TOL - tolerancia
t - a számábrázolás pontossága
OUTPUT: $\mathbf{z}$ - az egyenlet megoldásának közelítése
$COND$ - $\mathrm{cond}_\infty(\mathbf{A})$ közelítése

Az $\mathbf{Ax} = \mathbf{b}$ egyenletet megoldjuk Gauss-eliminációval
for $k = 1, 2, \ldots, N$ **do**
  Az $\mathbf{r} = \mathbf{b} - \mathbf{Ax}$ reziduális vektort kétszeres pontossággal kiszámoljuk.
  Az $\mathbf{Ay} = \mathbf{r}$ egyenletet megoldjuk $\mathbf{y}$-ra
  $\mathbf{z} \leftarrow \mathbf{x} + \mathbf{y}$
  if $k = 1$ **do**
    $COND \leftarrow 10^t \frac{\|\mathbf{y}\|_\infty}{\|\mathbf{x}\|_\infty}$
    output$(COND)$
  end do
  if $\|\mathbf{y}\|_\infty < TOL$ **do**
    output$(\mathbf{z})$
    stop
  end do
  $\mathbf{x} \leftarrow \mathbf{z}$
end do
output(A maximális iterációszámot túlléptük)

4.21. példa. Tekintsük a (4.20) egyenletet. Ennek pontos megoldása $\mathbf{x} = (1,\ 1)^T$. Gauss-eliminációval négyjegyű aritmetikát használva az $\tilde{\mathbf{x}} = (0.9375,\ 1.2500)^T$ közelítő megoldást kapjuk. Az ehhez tartozó reziduális vektor (dupla pontossággal számolva): $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}} = (0,\ 0.001875)^T$, így $\|\mathbf{r}\|_\infty = 0.001875$.

Az $\mathbf{Ay} = \mathbf{r}$ egyenletet Gauss-eliminációval megoldva (négyjegyű aritmetikát használva) kapjuk $\tilde{\mathbf{y}} = (0.0586,\ -0.2344)^T$. Ezért a (4.23) becslés szerint

$$\mathrm{cond}_\infty(\mathbf{A}) \approx 10^4 \frac{\|\tilde{\mathbf{y}}\|_\infty}{\|\tilde{\mathbf{x}}\|_\infty} = 10^4 \frac{0.2344}{1.25} = 1875. \tag{4.24}$$

A 4.19. példában láttuk, hogy a kondíciószám pontos értéke: $\mathrm{cond}_\infty(\mathbf{A}) = 1346$, tehát (4.24) valóban közelítése a pontos kondíciószámnak. Az $\tilde{\mathbf{x}}$ köxelítő megoldás relatív hibája

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} = 0.25,$$

ami elég nagy ($\mathbf{A}$ rosszul kondícionált). A 4.18. tétel szerint a

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} \le \mathrm{cond}_\infty(\mathbf{A}) \frac{\|\mathbf{r}\|_\infty}{\|\mathbf{b}\|_\infty} = 0.5017$$

hibakorlátot kapjuk az elkövetett relatív hibára. Az iteratív finomítás egy lépését alkalmazva az $\mathbf{x}^{(2)} = \mathbf{x} + \mathbf{y} = (0.9961, 1.016)^T$ közelítő megoldást kapjuk, ami közel van az egyenlet pontos megoldásához. □

Feladatok

1. Számítsa ki az

$$\text{(a)} \quad \begin{pmatrix} 1 & 2 \\ 4 & -1 \end{pmatrix}, \qquad \text{(b)} \quad \begin{pmatrix} 1 & 0 & 2 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix}$$

   mátrixok cond_∞ és cond_1 kondíciószámát!

2. Becsülje meg a $\text{cond}_\infty(\mathbf{A})$ kondíciószámot, ha

$$\mathbf{A} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix}.$$

3. Négyjegyű aritmetikát használva oldja meg az

$$0.009x_1 - 0.52x_2 = -5.191$$
$$9211x_1 + 21.1x_2 = 9422$$

   egyenletrendszert az iteratív finomítás módszerének két lépését használva! (A pontos megoldás: $(1,\ 10)$.)

4.5. Lineáris egyenletrendszerek perturbációja

Tekintsük az

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{4.25}$$

lineáris egyenletrendszert. Tegyük fel, hogy a (4.25) egyenlet jobb oldala helyett annak egy kis perturbációja, $\tilde{\mathbf{b}} = \mathbf{b} + \Delta\mathbf{b}$ adott, és a hozzá tartozó

$$\mathbf{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}} \tag{4.26}$$

egyenletet oldjuk meg, aminek a megoldását $\tilde{\mathbf{x}}$-mal jelöltük.

4.22. tétel. *Legyen A nemszinguláris, $\mathbf{x}$ és $\tilde{\mathbf{x}}$ megoldása a (4.25) ill. a (4.26) egyenletnek. Ekkor*

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \text{cond}(\mathbf{A}) \frac{\|\mathbf{b} - \tilde{\mathbf{b}}\|}{\|\mathbf{b}\|}.$$

Bizonyítás. A (4.25) és (4.26) egyenleteket kivonva egymásból $\mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}}) = \mathbf{b} - \tilde{\mathbf{b}}$ adódik, amiből $\mathbf{x} - \tilde{\mathbf{x}} = \mathbf{A}^{-1}(\mathbf{b} - \tilde{\mathbf{b}})$, azaz $\|\mathbf{x} - \tilde{\mathbf{x}}\| \le \|\mathbf{A}^{-1}\|\|\mathbf{b} - \tilde{\mathbf{b}}\|$. Ezt és az $\|\mathbf{b}\| = \|\mathbf{A}\mathbf{x}\| \le \|\mathbf{A}\|\|\mathbf{x}\|$ egyenlőtlenséget felhasználva

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \frac{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|\|\mathbf{b} - \tilde{\mathbf{b}}\|}{\|\mathbf{A}\|\|\mathbf{x}\|} \le \text{cond}(\mathbf{A}) \frac{\|\mathbf{b} - \tilde{\mathbf{b}}\|}{\|\mathbf{b}\|}.$$

□

A tétel szerint egy nagyságrendi növekedés $\mathrm{cond}(\mathbf{A})$-ban eredményezheti a megoldás relatív hibájának egy nagyságrenddel való növekedését, azaz egy értékes számjegy elvesztését.

Tekintsük most az általános esetet, az együtthatómátrixot és az egyenlet jobb oldalát is perturbáljuk:

$$\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}, \tag{4.27}$$

ahol $\|\mathbf{b} - \tilde{\mathbf{b}}\|$ és $\|\mathbf{A} - \tilde{\mathbf{A}}\|$ „kicsi”.

4.23. tétel. *Legyen A nemszinguláris, $\tilde{\mathbf{A}}$ olyan hogy $\|\mathbf{A}-\tilde{\mathbf{A}}\| < 1/\|\mathbf{A}^{-1}\|$. Legyen $\mathbf{x}$ megoldása (4.25)-nek és $\tilde{\mathbf{x}}$ megoldása (4.27)-nek. Ekkor*

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \frac{\mathrm{cond}(\mathbf{A})}{1 - \mathrm{cond}(\mathbf{A})\frac{\|\mathbf{A}-\tilde{\mathbf{A}}\|}{\|\mathbf{A}\|}} \left(\frac{\|\mathbf{A} - \tilde{\mathbf{A}}\|}{\|\mathbf{A}\|} + \frac{\|\mathbf{b} - \tilde{\mathbf{b}}\|}{\|\mathbf{b}\|} \right).$$

Bizonyítás. Induljunk ki az $\tilde{\mathbf{A}} = \mathbf{A} - (\mathbf{A} - \tilde{\mathbf{A}}) = \mathbf{A}(\mathbf{I} - \mathbf{A}^{-1}(\mathbf{A} - \tilde{\mathbf{A}}))$ azonosságból. Mivel a feltétel szerint $\|\mathbf{A}^{-1}(\mathbf{A} - \tilde{\mathbf{A}})\| \le \|\mathbf{A}^{-1}\|\|\mathbf{A} - \tilde{\mathbf{A}}\| < 1$, ezért a 4.4. állítás szerint $\tilde{\mathbf{A}}$ invertálható, és

$$\begin{aligned} \|(\tilde{\mathbf{A}})^{-1}\| &\le \|(\mathbf{I} - \mathbf{A}^{-1}(\mathbf{A} - \tilde{\mathbf{A}}))^{-1}\|\|\mathbf{A}^{-1}\| \\ &\le \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}(\mathbf{A} - \tilde{\mathbf{A}})\|} \\ &\le \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\|\|\mathbf{A} - \tilde{\mathbf{A}}\|}. \end{aligned}$$

A (4.26) és (4.25) egyenletekből kapjuk

$$\mathbf{x} - \tilde{\mathbf{x}} = \mathbf{x} - (\tilde{\mathbf{A}})^{-1}\tilde{\mathbf{b}} = (\tilde{\mathbf{A}})^{-1}(\tilde{\mathbf{A}}\mathbf{x} - \tilde{\mathbf{b}}) = (\tilde{\mathbf{A}})^{-1}(\mathbf{b} - \tilde{\mathbf{b}} - (\mathbf{A} - \tilde{\mathbf{A}})\mathbf{x}).$$

Ebből

$$\begin{aligned} \|\mathbf{x} - \tilde{\mathbf{x}}\| &\le \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\|\|\mathbf{A} - \tilde{\mathbf{A}}\|}(\|\mathbf{b} - \tilde{\mathbf{b}}\| + \|\mathbf{A} - \tilde{\mathbf{A}}\|\|\mathbf{x}\|) \\ &= \frac{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\|\|\mathbf{A}\|\frac{\|\mathbf{A}-\tilde{\mathbf{A}}\|}{\|\mathbf{A}\|}} \left(\frac{\|\mathbf{b} - \tilde{\mathbf{b}}\|}{\|\mathbf{A}\|} + \frac{\|\mathbf{A} - \tilde{\mathbf{A}}\|}{\|\mathbf{A}\|}\|\mathbf{x}\| \right). \end{aligned}$$

Leosztva az egyenlőtlenséget $\|\mathbf{x}\|$-val és a $\|\mathbf{b}\| \le \|\mathbf{A}\|\|\mathbf{x}\|$ egyenlőtlenséget alkalmazva

$$\begin{aligned} \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} &\le \frac{\mathrm{cond}(\mathbf{A})}{1 - \mathrm{cond}(\mathbf{A})\frac{\|\mathbf{A}-\tilde{\mathbf{A}}\|}{\|\mathbf{A}\|}} \left(\frac{\|\mathbf{b} - \tilde{\mathbf{b}}\|}{\|\mathbf{A}\|\|\mathbf{x}\|} + \frac{\|\mathbf{A} - \tilde{\mathbf{A}}\|}{\|\mathbf{A}\|} \right) \\ &\le \frac{\mathrm{cond}(\mathbf{A})}{1 - \mathrm{cond}(\mathbf{A})\frac{\|\mathbf{A}-\tilde{\mathbf{A}}\|}{\|\mathbf{A}\|}} \left(\frac{\|\mathbf{b} - \tilde{\mathbf{b}}\|}{\|\mathbf{b}\|} + \frac{\|\mathbf{A} - \tilde{\mathbf{A}}\|}{\|\mathbf{A}\|} \right). \end{aligned}$$

□

Könnyen igazolhatók a kondíciószám következő tulajdonságai:

4.24. tétel. *Legyen $\|\cdot\|$ egy tetszőleges mátrixnorma és $\mathrm{cond}(\cdot)$ a hozzá tartozó kondíciószám függvény. Ekkor*

1. $\mathrm{cond}(\mathbf{A}) \geq 1$,

2. $\rho(\mathbf{A})\rho(\mathbf{A}^{-1}) \leq \mathrm{cond}(\mathbf{A})$

teljesül minden invertálható $\mathbf{A}$*-ra.*

A $\mathrm{cond}_*(\mathbf{A}) \equiv \rho(\mathbf{A})\rho(\mathbf{A}^{-1})$ számot az $\mathbf{A}$ mátrix *spektrál kondíciószámának* nevezzük. Az előző tétel szerint a mátrix spektrál kondíciószáma kisebb, mint bármely normához tartozó kondíciószáma. Hátránya, hogy nehéz kiszámolni, mivel a mátrix sajátértékeit kell hozzá meghatározni.

Bizonyítás nélkül közöljük a következő eredményt:

4.25. tétel (Gastinel). *Legyen* $\|\cdot\|$ *egy tetszőleges mátrixnorma,* $\mathbf{A}$ *invertálható mátrix. Ekkor*

$$\frac{1}{\mathrm{cond}(\mathbf{A})} = \min\left\{\frac{\|\mathbf{A}-\mathbf{B}\|}{\|\mathbf{A}\|} :\ \mathbf{B}\ \textit{szinguláris}\right\}.$$

A tételből következik, hogy ha az $\mathbf{A}$ mátrix kondíciószáma nagy, akkor $\mathbf{A}$-hoz „közel" van egy szinguláris mátrix.

Rosszul kondícionált mátrixok klasszikus példája az ún. *Hilbert-mátrix:*

$$\mathbf{H}_n = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \vdots & & & & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{pmatrix}.$$

A 4.3. táblázatban feltüntettük a Hilbert-mátrix spektrál kondíciószámát néhány n-re. Látható, hogy milyen gyorsan növekszik a spektrál kondíciószám n növekedésével.

4.3. táblázat. A Hilbert-mátrix spektrál kondíciószáma

n	$\mathrm{cond}_*(\mathbf{H}_n)$	n	$\mathrm{cond}_*(\mathbf{H}_n)$
3	$5.24 \cdot 10^2$	7	$7.45 \cdot 10^8$
4	$1.55 \cdot 10^4$	8	$1.53 \cdot 10^{10}$
5	$4.77 \cdot 10^5$	9	$4.93 \cdot 10^{11}$
6	$1.50 \cdot 10^6$	10	$1.60 \cdot 10^{13}$

Feladatok

1. Számítsa ki az
$$\begin{pmatrix} 1 & 4 \\ 2 & -1 \end{pmatrix}$$
mátrix spektrál kondíciószámát!
2. Bizonyítsa be a 4.24. tételt!
3. Igazolja, hogy
$$\rho(\mathbf{A})\rho(\mathbf{A}^{-1}) = \frac{\max\{|\lambda_1|, \ldots, |\lambda_n|\}}{\min\{|\lambda_1|, \ldots, |\lambda_n|\}},$$
ahol $\lambda_1, \ldots, \lambda_n$ az $\mathbf{A}$ mátrix sajátértékei!

5. fejezet
Mátrix faktorizáció

A következő lineáris algebrai feladatot vizsgáljuk: alakítsuk az $\mathbf{A}$ mátrixot $\mathbf{A} = \mathbf{BC}$ alakú szorzattá, ahol $\mathbf{B}$ és $\mathbf{C}$ speciális mátrixok. Először az LU-faktorizációt, majd annak speciális esetét, a Cholesky-faktorizációt tanulmányozzuk, ahol alulról és felülről trianguláris mátrixok szorzatára szeretnénk felbontani az adott mátrixot, majd a QR-faktorizációt tekintjük, ahol ortogonális és felülről trianguláris faktorokat keresünk. Ez utóbbi módszer lesz az alapja a sajátérték keresés egyik módszerének, amelyet a következő fejezetben fogunk definiálni.

5.1. LU-faktorizáció

Legyen $\mathbf{A}$ egy $n \times n$-es mátrix. Az $\mathbf{A} = \mathbf{LU}$ szorzatot, ahol $\mathbf{L}$ alulról trianguláris mátrix, amelynek főátlójában csupa egyes áll, az $\mathbf{U}$ mátrix pedig felülről trianguláris, az $\mathbf{A}$ mátrix *trianguláris felbontásának* vagy *LU-faktorizációjának* vagy *Doolittle-faktorizációjának* nevezzük.

5.1. tétel. *Legyen $\mathbf{A}$ egy nemszinguláris mátrix. Ha az $\mathbf{A}$ mátrix LU-faktorizációja létezik, akkor az egyértelmű.*

Bizonyítás. Tegyük fel, hogy $\mathbf{A} = \mathbf{L}_1\mathbf{U}_1 = \mathbf{L}_2\mathbf{U}_2$ az $\mathbf{A}$ mátrix két LU-felbontása. Mivel $\det(\mathbf{A}) = \det(\mathbf{L}_1)\det(\mathbf{U}_1) = \det(\mathbf{L}_2)\det(\mathbf{U}_2) \neq 0$, ezért $\mathbf{L}_1$, $\mathbf{L}_2$, $\mathbf{U}_1$ és $\mathbf{U}_2$ nemszinguláris mátrixok. Így $\mathbf{L}_2^{-1}\mathbf{L}_1 = \mathbf{U}_2\mathbf{U}_1^{-1}$. A 3.6. tétel szerint a $\mathbf{L}_2^{-1}\mathbf{L}_1$ szorzat alulról trianguláris, a $\mathbf{U}_2\mathbf{U}_1^{-1}$ pedig felülről trianguláris mátrix. Mivel a két mátrix megegyezik, ezért ennek diagonálisnak kell lennie. Könnyen látható, hogy az $\mathbf{L}_2^{-1}\mathbf{L}_1$ mátrix főátlójában csupa egyes áll, tehát $\mathbf{L}_2^{-1}\mathbf{L}_1 = \mathbf{U}_2\mathbf{U}_1^{-1} = \mathbf{I}$, amiből kapjuk, hogy $\mathbf{L}_1 = \mathbf{L}_2$ és $\mathbf{U}_1 = \mathbf{U}_2$. □

Térjünk vissza a 3.3. szakaszban bevezetett Gauss-elimináció definíciójához. Legyen $l_{i1} = a_{i1}/a_{11}$, $i = 2, 3, \ldots, n$, mint a 3.3. szakaszban, és definiáljuk az

$$\mathbf{L}_1 \equiv \begin{pmatrix} 1 & & & & \\ -l_{21} & 1 & & & \\ -l_{31} & & 1 & & \\ \vdots & & & \ddots & \\ -l_{n1} & & & & 1 \end{pmatrix}$$

alulról trianguláris mátrixot, amelynek az első oszlopa és a főátlója kivételével minden eleme 0. Könnyen ellenőrizhető, hogy az $\mathbf{L}_1\mathbf{A}$ mátrixszorzat pontosan a Gauss-elimináció első lépésekor kapott $\mathbf{A}^{(1)}$ mátrixot adja vissza: $\mathbf{A}^{(1)} = \mathbf{L}_1\mathbf{A}$. Hasonlóan, legyen $l_{i2} = a_{i2}^{(1)}/a_{22}^{(1)}$, $i = 3, 4, \ldots, n$, és legyen

$$\mathbf{L}_2 \equiv \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & -l_{32} & 1 & & \\ & \vdots & & \ddots & \\ & -l_{n2} & & & 1 \end{pmatrix},$$

amelynél a főátlóban csupa egyes, a második oszlopban a főátló alatt a $-l_{32}$, $-l_{42}$, ..., $-l_{n2}$ számok állnak, a többi elem pedig 0. Ekkor $\mathbf{A}^{(2)} = \mathbf{L}_2\mathbf{A}^{(1)}$ teljesül. Hasonlóan definiáljuk az $\mathbf{L}_3, \ldots, \mathbf{L}_{n-1}$ alulról trianguláris mátrixokat. Egyszerű számolással kapjuk, hogy

$$\mathbf{L}_{n-1}\mathbf{L}_{n-2}\cdots\mathbf{L}_1 = \begin{pmatrix} 1 & & & & \\ -l_{21} & 1 & & & \\ -l_{31} & -l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ -l_{n1} & -l_{n2} & \cdots & -l_{n,n-1} & 1 \end{pmatrix}, \tag{5.1}$$

és

$$\begin{aligned} \mathbf{L} &\equiv (\mathbf{L}_{n-1}\mathbf{L}_{n-2}\cdots\mathbf{L}_1)^{-1} \\ &= \mathbf{L}_1^{-1}\cdots\mathbf{L}_{n-2}^{-1}\mathbf{L}_{n-1}^{-1} \\ &= \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & 0 & 1 & & \\ \vdots & 0 & \ddots & \ddots & \\ l_{n1} & 0 & \cdots & 0 & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & \vdots & \ddots & \ddots & \\ 0 & 0 & \cdots & l_{n,n-1} & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{n,n-1} & 1 \end{pmatrix}. \end{aligned} \tag{5.2}$$

Legyen $\mathbf{U} \equiv \mathbf{A}^{(n-1)}$, azaz a Gauss-eliminációval kapott felülről trianguláris mátrix. Ekkor $\mathbf{U} = \mathbf{L}_{n-1}\cdots\mathbf{L}_1\mathbf{A}$, amiből $\mathbf{A} = \mathbf{LU}$. Beláttuk tehát a következő tételt:

5.2. tétel. *Ha a Gauss-elimináció végrehajtható egy* $\mathbf{A}$ *mátrixon, akkor az* $\mathbf{A} = \mathbf{LU}$ *faktorizáció létezik. Ekkor* $\mathbf{U}$ *a Gauss-eliminációval kapott felülről trianguláris mátrix,* $\mathbf{L}$ *pedig az (5.2) képlettel definiált alulról trianguláris mátrix, ahol* l_{ij} *jelöli a Gauss-eliminációban használt faktorokat.*

5.3. példa. Vegyük a 3.22. példában szereplő együtthatómátrixot:

$$\mathbf{A} = \begin{pmatrix} 1 & -2 & -2 & -2 \\ 2 & -1 & 2 & 4 \\ -1 & 2 & 3 & -4 \\ -2 & 1 & 4 & -2 \end{pmatrix}.$$

A 3.22. példában már láttuk, hogy a Gauss-elimináció végrehajtható $\mathbf{A}$-n, és $l_{21} = 2$, $l_{31} = -1$, $l_{41} = -2$, $l_{32} = 0$, $l_{42} = -1$ és $l_{43} = 6$. Az LU faktorizáció céljából végzett Gauss-eliminációt úgy szokás leírni, hogy az l_{ij} elemeket a kinullázott elemek helyére írjuk le:

$$\begin{pmatrix} 1 & -2 & -2 & -2 \\ 2 & -1 & 2 & 4 \\ -1 & 2 & 3 & -4 \\ -2 & 1 & 4 & -2 \end{pmatrix} \sim \begin{pmatrix} 1 & -2 & -2 & -2 \\ 2 & 3 & 6 & 8 \\ -1 & 0 & 1 & -6 \\ -2 & -3 & 0 & -6 \end{pmatrix} \sim$$

$$\begin{pmatrix} 1 & -2 & -2 & -2 \\ 2 & 3 & 6 & 8 \\ -1 & 0 & 1 & -6 \\ -2 & -1 & 6 & 2 \end{pmatrix} \sim \begin{pmatrix} 1 & -2 & -2 & -2 \\ 2 & 3 & 6 & 8 \\ -1 & 0 & 1 & -6 \\ -2 & -1 & 6 & 38 \end{pmatrix}.$$

Az utolsó mátrixban ekkor a főátlóban és fölötte $\mathbf{U}$ elemei, alatta pedig $\mathbf{L}$ elemei állnak. Azaz

$$\begin{pmatrix} 1 & -2 & -2 & -2 \\ 2 & -1 & 2 & 4 \\ -1 & 2 & 3 & -4 \\ -2 & 1 & 4 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -2 & -1 & 6 & 1 \end{pmatrix} \begin{pmatrix} 1 & -2 & -2 & -2 \\ 0 & 3 & 6 & 8 \\ 0 & 0 & 1 & -6 \\ 0 & 0 & 0 & 38 \end{pmatrix},$$

amit beszorzással ellenőrizhetünk. □

Könnyen beláthatók a következő tételek (4. feladat):

5.4. tétel. *Ha az* $\mathbf{A}$ *mátrix összes bal felső főminorjai 0-tól különböznek, akkor a Gauss-elimináció sorcsere nélkül végrehajtható, és így az* $\mathbf{A} = \mathbf{LU}$ *faktorizáció létezik.*

5.5. tétel. *Tetszőleges* $n \times n$*-es invertálható* $\mathbf{A}$ *mátrixhoz létezik olyan* $\mathbf{P}$ *permutációs mátrix, hogy a* $\mathbf{PA} = \mathbf{LU}$ *faktorizáció létezik.*

Ha ismerjük egy $\mathbf{A}$ mátrix $\mathbf{A} = \mathbf{LU}$ felbontását, akkor annak segítségével hatékonyan tudunk lineáris egyenletrendszereket megoldani. Tekintsük az $\mathbf{Ax} = \mathbf{b}$ egyenletet. Vezessük be az $\mathbf{y} = \mathbf{Ux}$ új változót. Ekkor az eredeti egyenletrendszer ekvivalens az

$$\begin{aligned} \mathbf{Ly} &= \mathbf{b} \\ \mathbf{Ux} &= \mathbf{y} \end{aligned}$$

trianguláris együtthatójú egyenletekkel. Először az elsőt oldjuk meg a visszahelyettesítés algoritmusával analóg módszerrel, majd $\mathbf{y}$ ismeretében a másodikat a visszahelyettesítés módszerével. Könnyen ellenőrizhető, hogy a két egyenlet megoldásához $n^2 + \mathcal{O}(n)$, az LU-faktorizációhoz pedig $n^3/3 + \mathcal{O}(n^2)$ számú osztásra ill. szorzásra van szükség. Különösen előnyös ezt a módszert használni abban az esetben, amikor különböző jobb oldalra de azonos együtthatómátrixra kell megoldani több lineáris lineáris egyenletrendszert.

Feladatok

1. Számítsa ki a következő mátrixok LU-felbontását:

(a) $\begin{pmatrix} 2 & 3 & -1 \\ -1 & -2 & -1 \\ 0 & 2 & 4 \end{pmatrix}$ (b) $\begin{pmatrix} 4 & -1 & 2 \\ -12 & 0 & -1 \\ 8 & -17 & 26 \end{pmatrix}$

(c) $\begin{pmatrix} 1 & 3 & -1 & 2 \\ -2 & -4 & 5 & -5 \\ 0 & 6 & 6 & -2 \\ 2 & 4 & -14 & 16 \end{pmatrix}$ (d) $\begin{pmatrix} 2 & -1 & 3 & -2 \\ -8 & 5 & -7 & 7 \\ 2 & -4 & -14 & 0 \\ -4 & 7 & 23 & 4 \end{pmatrix}$

2. Mutassa meg, hogy a

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

mátrixnak nem létezik az LU-faktorizációja!

3. Mutassa meg, hogy az

$$\begin{pmatrix} 1 & 1 & -1 \\ 2 & 2 & 2 \\ 3 & 3 & -4 \end{pmatrix}$$

mátrixnak végtelen sok LU-felbontása van! Nem mond ez ellent az 5.1. tételnek?

4. Bizonyítsa be az 5.4. tételt! (Útmutatás: használja, hogy az elimináció során az $\mathbf{A}^{(k-1)}$ és $\mathbf{A}^{(k)}$ mátrixok megfelelő főminorjai megegyeznek. Miért?)
5. Bizonyítsa be az 5.5. tételt!
6. Oldja meg a 3.3. szakasz 1. feladatában szereplő lineáris egyenletrendszereket LU-faktorizációt használva!

5.2. Cholesky-faktorizáció

Legyen $\mathbf{A}$ egy szimmetrikus mátrix. Az $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ szorzatot, ahol $\mathbf{L}$ egy alulról trianguláris mátrix, az $\mathbf{A}$ mátrix *Cholesky-faktorizációjának* nevezzük.

Megjegyezzük, hogy a Cholesky-faktorizáció, ha létezik, nem egyértelmű. A következő tétel elégséges feltételt biztosít a Cholesky-faktorizáció létezésére.

5.6. tétel. *Ha $\mathbf{A}$ pozitív definit, akkor az $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ Cholesky-faktorizáció létezik, az $\mathbf{L}$ mátrix valós, és a főátlójában pozitív elemeket választhatunk.*

Bizonyítás. Az $\mathbf{A}$ mátrix dimenziója szerinti teljes indukcióval látjuk be az állítást. 1×1-es mátrixokra az állítás nyilvánvaló. Tegyük fel, hogy $(n-1) \times (n-1)$-es mátrixokra teljesül az állítás, és legyen $\mathbf{A}$ $n \times n$-es mátrix. Az $\mathbf{A}$ mátrixot partícionáljuk a következő alakba:

$$\mathbf{A} = \begin{pmatrix} \mathbf{X} & \mathbf{y} \\ \mathbf{y}^T & a_{nn} \end{pmatrix},$$

ahol $\mathbf{X}$ egy $(n-1) \times (n-1)$-es mátrix, $\mathbf{y}$ egy $n-1$-dimenziós oszlopvektor. A 3.10. tételből következik, hogy $\mathbf{X}$ pozitív definit. Keressük az $\mathbf{A}$ mátrix Cholesky-felbontását az

$$\mathbf{A} = \begin{pmatrix} \mathbf{X} & \mathbf{y} \\ \mathbf{y}^T & a_{nn} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{L}} & \mathbf{0} \\ \mathbf{c}^T & d \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{L}}^T & \mathbf{c} \\ \mathbf{0}^T & d \end{pmatrix} \tag{5.3}$$

alakban. Itt $\tilde{\mathbf{L}}$ egy $(n-1) \times (n-1)$-es alulról trianguláris mátrix, $\mathbf{c}$ egy $n-1$-dimenziós oszlopvektor, $d \in \mathbb{R}$. Ha a mátrixszorzást elvégezzük a partícionált mátrixokon, akkor az

$$\mathbf{X} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T, \qquad \tilde{\mathbf{L}}\mathbf{c} = \mathbf{y}, \quad \text{és} \quad \mathbf{c}^T\mathbf{c} + d^2 = a_{nn}$$

egyenleteket kapjuk. Az indukciós hipotézis szerint az $\mathbf{X} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T$ egyenletnek létezik $\tilde{\mathbf{L}} \in \mathbb{R}^{(n-1)\times(n-1)}$ alulról trianguláris megoldása, amelynek főátlójában pozitív elemeket választhatunk. Ebből következik, hogy $\tilde{\mathbf{L}}$ nemszinguláris mátrix, így az $\tilde{\mathbf{L}}\mathbf{c} = \mathbf{y}$ egyenletnek is létezik egyértelmű megoldása. Legyen d egy (esetleg komplex) gyöke a $\mathbf{c}^T\mathbf{c} + d^2 = a_{nn}$ egyenletnek. Ekkor az (5.3) összefüggés teljesül. d pontosan akkor választható pozitív valós számnak, ha $d^2 = a_{nn} - \mathbf{c}^T\mathbf{c} > 0$. Ha az (5.3) egyenlet bal és jobb oldalának determinánsát vesszük, akkor a $\det(\mathbf{A}) = \det(\tilde{\mathbf{L}})^2 d^2$ összefüggést kapjuk. Mivel $\mathbf{A}$ pozitív definit, így $\det(\mathbf{A}) > 0$ (lásd a 3.10. tételt). Ebből következik, hogy d^2 pozitív, azaz d választható pozitív valós számnak. □

5.7. példa. Keressük meg a

$$\begin{pmatrix} 4 & -8 & 4 \\ -8 & 17 & -11 \\ 4 & -11 & 22 \end{pmatrix}.$$

mátrix Cholesky-felbontását! Írjuk fel a keresett felbontást:

$$\begin{pmatrix} 4 & -8 & 4 \\ -8 & 17 & -11 \\ 4 & -11 & 22 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{pmatrix}$$

A módszer a következő: először tekintsük az első sor első elemére vonatkozó egyenletet: $4 = l_{11}^2$. Ezt megoldhatjuk l_{11}-re: a pozitív megoldás $l_{11} = 2$. Ezután sorra írjuk fel az első oszlopban a főátló alatti elemekre az egyenleteket: $-8 = l_{21}l_{11}$, $4 = l_{31}l_{11}$. Ezeket meg tudjuk oldani egyértelműen l_{21} és l_{31}-re: $l_{21} = -4$, $l_{31} = 2$. Most nézzük a második oszlop főátlójában levő elemet: $17 = l_{21}^2 + l_{22}^2$. Ennek pozitív megoldása $l_{22} = 1$. Ezután a második oszlop főátlója alatti elemeket nézzük: $-11 = l_{31}l_{21} + l_{32}l_{22}$. Ez megoldható l_{32}-re: $l_{32} = -3$. Végül a harmadik oszlop főátlójában levő elemét tekintjük: $22 = l_{31}^2 + l_{32}^2 + l_{33}^2$. Ebből $l_{33} = 3$. Kaptuk tehát, hogy

$$\begin{pmatrix} 4 & -8 & 4 \\ -8 & 17 & -11 \\ 4 & -11 & 22 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ -4 & 1 & 0 \\ 2 & -3 & 3 \end{pmatrix} \begin{pmatrix} 2 & -4 & 2 \\ 0 & 1 & -3 \\ 0 & 0 & 3 \end{pmatrix}.$$

□

Az előző példa általánosítását a következőképpen fogalmazhatjuk meg:

5.8. algoritmus. Cholesky-faktorizáció

INPUT: **A**
OUTPUT: **L**

$l_{11} \leftarrow \sqrt{a_{11}}$
for $i = 2, \ldots, n$ **do**
  $l_{i1} \leftarrow a_{i1}/l_{11}$
end do
for $j = 2, \ldots, n-1$ **do**
  $l_{jj} \leftarrow \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}$
  for $i = j+1, \ldots, n$ **do**
    $l_{ij} \leftarrow \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}\right) / l_{jj}$
  end do
end do
$l_{nn} \leftarrow \sqrt{a_{nn} - \sum_{k=1}^{n-1} l_{nk}^2}$
output$(l_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, i)$

Az 5.8. algoritmus műveletigénye $n^3/6 + n^2/2 - 2n/3$ osztás ill. szorzás, $n^3/6 - n/6$ összeadás ill. kivonás és n db gyökvonás.

Feladatok

1. Számítsa ki a következő mátrixoknak azt a Cholesky-faktorizációját, amelynél a főátlóban pozitív elemek állnak:

$$\text{(a)} \begin{pmatrix} 16 & -8 & -12 \\ -8 & 8 & 4 \\ -12 & 4 & 35 \end{pmatrix}, \qquad \text{(b)} \begin{pmatrix} 4 & -2 & -4 \\ -2 & 26 & 7 \\ -4 & 7 & 6 \end{pmatrix},$$

$$\text{(c)} \begin{pmatrix} 1 & -1 & -2 & 1 \\ -1 & 10 & 2 & 2 \\ -2 & 2 & 29 & 8 \\ 1 & 2 & 8 & 7 \end{pmatrix}, \qquad \text{(d)} \begin{pmatrix} 16 & -8 & 0 & -4 \\ -8 & 5 & 1 & 3 \\ 0 & 1 & 10 & -5 \\ -4 & 3 & -5 & 7 \end{pmatrix}.$$

2. Mutasson példát arra, hogy a Cholesky-faktorizáció nem egyértelmű!
3. Mutassa meg, hogy a
$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
mátrixnak nem létezik a Cholesky-felbontása!
4. Igazolja, hogy a Cholesky-faktorizáció műveletigénye $n^3/6 + n^2/2 - 2n/3$ osztás ill. szorzás és $n^3/6 - n/6$ összeadás ill. kivonás!
5. Lássa be a 3.10. tételre való hivatkozás nélkül, hogy az 5.6. tétel bizonyításában szereplő **X** mátrix pozitív definit!

6. fejezet

Interpoláció

Adottak x_0, x_1, …, $x_n \in [a,b]$ páronként különböző pontok, ún. alappontok vagy osztópontok, és hozzá tartozó y_0, y_1, …, y_n függvényértékek. Az interpoláció alapfeladata a következő: keresünk olyan, valamely adott függvényosztálybeli g függvényt, amely *interpolálja* a megadott pontokat, azaz teljesíti a

$$g(x_i) = y_i, \qquad i = 0, 1, \ldots, n$$

egyenleteket. Az interpolációs feladat geometriai jelentése az, hogy olyan g függvényt keresünk, amely valamely megadott tulajdonságokkal rendelkezik, és a grafikonja átmegy az (x_i, y_i) pontokon.

Ebben a fejezetben először ennek az általános feladatnak elméleti és gyakorlati szempontból talán legfontosabb esetével, a polinom interpolációval foglalkozunk, azaz feltesszük, hogy g polinom. A 6.4. szakaszban ennek az interpolációs feladatnak egy általánosítását, az ún. Hermite-interpolációt vizsgáljuk, ahol nem csak függvényértékeket, hanem derivált értékeket is interpolálunk. Tárgyaljuk továbbá a spline függvényekkel (szakaszonként polinomokkal) történő interpolációt is.

6.1. Lagrange-interpoláció

Tegyük fel most, hogy a bevezetésben leírt interpolációs alapfeledatban $g(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_m x^m$ alakú. Ebben a képletben $m+1$ ismeretlen szerepel, és az interpolációs feltételek $n+1$ egyenletet határoznak meg. Természetes azt várni, hogy a feladatnak az $m = n$ esetben lesz egyértelmű megoldása. Fogalmazzuk újra a feladatot: Keresünk egy olyan L_n legfeljebb n-edfokú polinomot, amelyre

$$L_n(x_i) = y_i, \qquad i = 0, 1, \ldots, n. \tag{6.1}$$

Ez a *Lagrange-féle interpolációs feladat*. Megmutatjuk, hogy ennek a feladatnak mindig létezik egyértelmű megoldása. A feladatot teljesítő L_n polinomot *Lagrange-féle interpolációs polinomnak*, vagy röviden Lagrange-polinomnak nevezzük. Azt, hogy ilyen polinom létezik, könnyű belátni: megadjuk L_n explicit képletét az alappontok és az adott függvényértékek segítségével. Definiáljuk $k = 0, 1, \ldots, n$-re az

$$l_k(x) \equiv \frac{(x-x_0)(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)} \tag{6.2}$$

n-edfokú polinomokat. Az $l_0, \ldots, l_n$ polinomokat *Lagrange-féle n-edfokú alappolinomoknak* nevezzük. A polinom definíciójából nyilvánvaló, hogy

$$l_k(x_i) = \begin{cases} 1, & \text{ha } k = i, \\ 0, & \text{ha } k \neq i. \end{cases}$$

Ebből következik, hogy az $L_n(x) \equiv \sum_{k=0}^{n} y_k l_k(x)$ függvény egy legfeljebb n-edfokú polinom, és megoldása a (6.1) interpolációs problémának.

Most belátjuk, hogy a Lagrange-féle interpolációs feladatnak csak egy megoldása van. Tegyük fel, hogy L_n és $\tilde{L}_n$ mindketten legfeljebb n-edfokú polinomok és teljesítik a (6.1) egyenleteket. Definiáljuk a $P(x) \equiv L_n(x) - \tilde{L}_n(x)$ függvényt. Ekkor P is legfeljebb n-edfokú polinom, és $P(x_i) = 0$ minden $i = 0, 1, \ldots, n$-re, azaz P-nek $n+1$ különböző gyöke van. Ekkor viszont az algebra alaptételéből következik, hogy P azonosan 0 polinom, azaz $L_n = \tilde{L}_n$. Beláttuk tehát a következő állítást:

6.1. tétel. *A Lagrange-féle interpolációs feladatnak létezik egyértelmű megoldása, amely az*

$$L_n(x) = \sum_{k=0}^{n} y_k \frac{(x-x_0)(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)} \tag{6.3}$$

alakban adható meg.

6.2. példa. Tekintsük az

x_i	-1	1	2	3
y_i	-3	1	3	29

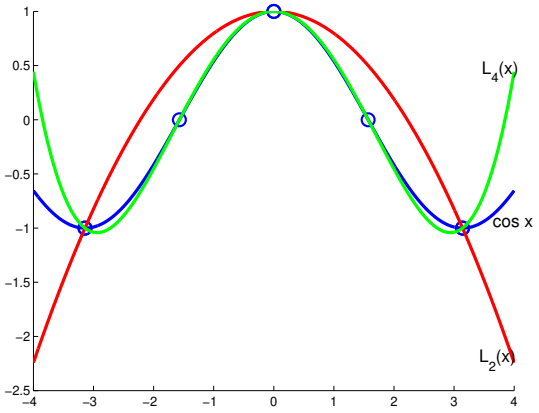
alappontokat és a hozzá tartozó függvényértékeket. Határozzuk meg az adatokhoz tartozó Lagrange-féle interpolációs polinomot! Mivel négy alappont van, ezért harmadfokú Lagrange-polinomot keresünk. A (6.3) képlet szerint

$$\begin{aligned} L_3(x) = & -3\frac{(x-1)(x-2)(x-3)}{(-1-1)(-1-2)(-1-3)} + \frac{(x+1)(x-2)(x-3)}{(1+1)(1-2)(1-3)} \\ & + 3\frac{(x+1)(x-1)(x-3)}{(2+1)(2-1)(2-3)} + 29\frac{(x+1)(x-1)(x-2)}{(3+1)(3-1)(3-2)} \\ = & \, 3x^3 - 6x^2 - x + 5. \end{aligned}$$

□

Az x_i értékekhez hozzárendelt y_i értékeket általában természetes módon tekinthetjük egy f függvény értékeinek az alappontokban, azaz $y_i = f(x_i)$. Például lehet f egy fizikai mennyiség, amelyet véges sok időpontban mértünk. Vagy lehet f egy matematikai modell megoldása, amelyet csak numerikus módszerekkel tudunk megoldani, és a megoldást, azaz az f függvény értékét csak véges sok pontban tudjuk megkapni, pontosabban a közelítő értékét megkapni. Vagy lehet, hogy f egy olyan függvény, amelynek képlete ill. kiszámítási szabálya ismert, csak „túl sok" számolást igényel f-et kiértékelni, így csak néhány pontban számoljuk ki f pontos értékét. Mindhárom esetben igény lehet arra, hogy f értékét kiszámoljuk, pontosabban megbecsüljük a már ismert véges sok függvényérték segítségével egy alapponton kívüli pontban is. Erre egyszerű módszer az, ha interpoláljuk a véges sok megadott pontot, és az interpolációs polinom adott pontbeli értékével (amit könnyű kiszámolni) közelítjük a kívánt függvényértéket. Az interpoláció kifejezést használjuk abban az értelemben, hogy az interpoláló függvényt (polinomot) számítjuk ki, de szokás interpoláción az interpolációs polinom segítségével történő függvényérték közelítést is érteni. Ez utóbbi esetben ha az a pont, amelyben az f függvényt akarjuk becsülni az alappontok által meghatározott intervallumon kívül esik, akkor *extrapolációról* szokás beszélni, interpoláción szigorúan véve azt értjük, amikor a megadott pont az alappontok között helyezkedik el.

6.3. példa. Tekintsük az $f(x) = \cos x$ függvényt a $[-\pi, \pi]$ intervallumon. A π, 0 és π illetve $-\pi$,$-\pi/2$, 0, $\pi/2$ és π osztópontokat használva meghatároztuk az L_2 és L_4 másod- ill. negyedfokú Lagrange-féle interpolációs polinomokat. A polinomok és az f függvény grafikonja a 6.1. ábrán látható. Az ábrából megállapíthatjuk, hogy az 5 osztópontot használva f jobb közelítését kapjuk, mint akkor, ha csak 3 pontot használunk. Az is nyilvánvaló ebben az esetben, hogy a $[-\pi, \pi]$ intervallumon kívül a polinomok



6.1. ábra. A $\cos x$ függvény interpolációja a $-\pi, 0, \pi$ ill. a $-\pi, -\pi/2, 0, \pi/2, \pi$ alappontokat használva

nem jó közelítései az eredeti függvénynek. □

A 6.5. tétel bizonyításához szükségünk lesz a következő segédtételre.

6.4. tétel (Általánosított Rolle-tétel). *Legyen $f \in C^n(a,b)$, $a \le x_0 < x_1 \cdots < x_n \le b$, és tegyük fel, hogy $f(x_0) = f(x_1) = \cdots = f(x_n) = 0$. Ekkor létezik olyan $\xi \in (x_0, x_n)$, hogy $f^{(n)}(\xi) = 0$.*

Bizonyítás. A feltételek szerint $f(x_0) = f(x_1) = 0$, így a Rolle-tétel (2.3 tétel) szerint létezik olyan $\eta_1 \in (x_0, x_1)$, hogy $f'(\eta_1) = 0$. Hasonlóan az $[x_1, x_2]$, …, $[x_{n-1}, x_n]$ intervallumokra alkalmazva a Rolle-tételt kapjuk, hogy léteznek olyan $\eta_2 \in (x_1, x_2)$, …, $\eta_n \in (x_{n-1}, x_n)$ számok, amelyekre $f'(\eta_2) = \cdots = f'(\eta_n) = 0$. Tekintsük ezután az $[\eta_1, \eta_2]$, …, $[\eta_{n-1}, \eta_n]$ intervallumokat. Mivel ezek végpontjaiban $f'(\eta_i) = 0$, ezért a Rolle-tétel szerint léteznek olyan $\theta_2 \in (\eta_1, \eta_2)$, …, $\theta_n \in (\eta_{n-1}, \eta_n)$ számok, amelyekre $f''(\theta_2) = \cdots = f''(\theta_n) = 0$. Ismételten alkalmazva a Rolle-tételt, kapjuk, hogy f harmadik deriváltja $n-2$ pontban, f negyedik deriváltja $n-3$ pontban stb., $f^{(n)}$ pedig egy pontban egyenlő nullával. □

6.5. tétel. *Legyen $f \in C^{n+1}(a,b)$, $x_i \in [a,b]$ $(i = 0, \dots, n)$ páronként különböző alappontok és $y_i = f(x_i)$ $(i = 0, \dots, n)$. Legyen $L_n(x)$ az adatokhoz tartozó n-edfokú Lagrange-polinom. Ekkor bármely $x \in [a,b]$-hez létezik olyan $\xi = \xi(x) \in \langle x, x_0, x_1, \dots, x_n \rangle$ szám, hogy*

$$f(x) = L_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n).$$

Bizonyítás. Ha $x = x_i$ valamely i-re, akkor az állítás nyilvánvalóan teljesül. Rögzítsünk egy $x \in (a,b)$ számot, amelyre $x \ne x_i$ minden $i = 0, \dots, n$-re, és tekintsük a

$$g(t) \equiv f(t) - L_n(t) - \frac{(t - x_0)\cdots(t - x_n)}{(x - x_0)\cdots(x - x_n)}(f(x) - L_n(x))$$

függvényt. Nyilvánvalóan $g \in C^{n+1}$, és $g(x) = g(x_0) = g(x_1) = \cdots = g(x_n) = 0$. Ekkor alkalmazva az általánosított Rolle-tételt (6.4. tétel), kapjuk, hogy létezik olyan $\xi \in \langle x, x_0, \dots, x_n \rangle$ szám, hogy $g^{(n+1)}(\xi) = 0$. Mivel L_n n-edfokú polinom, ezért $(n+1)$-edik deriváltja nulla, így

$$g^{(n+1)}(t) = f^{(n+1)}(t) - \frac{(n+1)!}{(x - x_0)\cdots(x - x_n)}(f(x) - L_n(x)).$$

Ebből a $t = \xi$ értéket véve adódik a tétel állítása. □

Most tekintsük azt a speciális esetet, amikor ekvidisztáns osztópontokat használunk, azaz $x_i = x_0 + ih$. A 6.5. tétel szerint az interpoláció képlethibája az

$$|f(x) - L_n(x)| \le \frac{M_{n+1}}{(n+1)!}|(x - x_0)\cdots(x - x_n)| \tag{6.4}$$

kifejezéssel becsülhető, ahol $M_{n+1} = \sup\{|f^{(n+1)}(t)| : t \in [x_0, x_n]\}$. Tegyük fel, hogy $x \in (x_k, x_{k+1})$ valamilyen $0 \le k < n$-re. Ekkor könnyen ellenőrizhető, hogy

$$|(x - x_k)(x - x_{k+1})| \le \frac{h^2}{4},$$

és így

$$\begin{aligned}
\prod_{i=0}^{n}|x - x_i| &\le \frac{h^2}{4}\prod_{i=0}^{k-1}(x - x_i)\prod_{i=k+2}^{n}(x_i - x) \\
&\le \frac{h^2}{4}\prod_{i=0}^{k-1}(x_{k+1} - x_i)\prod_{i=k+2}^{n}(x_i - x_k) \\
&= \frac{h^{n+1}}{4}\prod_{i=0}^{k-1}(k+1-i)\prod_{i=k+2}^{n}(i-k) \\
&= \frac{h^{n+1}}{4}(k+1)!(n-k)! \\
&\le \frac{h^{n+1}}{4}n!
\end{aligned}$$

(Lásd a 4. feladatot!) Ebből és a (6.4) egyenlőtlenségből következik:

6.6. tétel. *Legyen $f \in C^{n+1}(a,b)$, $x_i = a+i(b-a)/n$ $(i = 0,\dots,n)$ és $y_i = f(x_i)$ $(i = 0,\dots,n)$. Legyen $x \in [a,b]$. Ekkor*

$$|f(x) - L_n(x)| \le \frac{M_{n+1}}{4(n+1)}\left(\frac{b-a}{n}\right)^{n+1},$$

ahol $M_{n+1} \equiv \sup\{|f^{(n+1)}(x)|\colon x \in [a,b]\}$.

6.7. példa. Térjünk vissza a 6.3. példához! Az előző tétel szerint minden $x \in [-\pi,\pi]$-re

$$|f(x) - L_2(x)| \le \frac{1}{12}\pi^3 \approx 2.5839, \quad \text{és} \qquad |f(x) - L_4(x)| \le \frac{1}{20}\left(\frac{\pi}{2}\right)^5 \approx 0.4782.$$

Természetesen a 6.6. tétellel csak felső korlátot kapunk a hibára. A 6.1. ábrán látható, hogy a tényleges hiba ennél jelen esetben lényegesen kisebb. □

A következő eredményre szükségünk lesz a 7. fejezetben. A bizonyítást nem közöljük itt.

6.8. tétel. *Tegyük fel, hogy $f \in C^{n+2}(a,b)$, $a = x_0 < \cdots < x_n = b$, és legyen*

$$\frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)\cdots(x-x_n)$$

az n-edfokú Lagrange-interpoláció maradéktagja. Ekkor az $x \mapsto f^{(n+1)}(\xi(x))$ függvény folytonosan kiterjeszthető $x = x_i$-re, differenciálható minden $x \neq x_i$-re, és

$$\frac{d}{dx}f^{(n+1)}(\xi(x)) = \frac{1}{n+2}f^{(n+2)}(\eta(x))$$

alakú, ahol $\eta(x) \in \langle x_0, \ldots, x_n, x\rangle$, továbbá $\dfrac{d}{dx}f^{(n+1)}(\xi(x))$ is folytonosan kiterjeszthető $x = x_i$-re $(i = 0, 1, \ldots, n)$.

Most kétváltozós függvények interpolációjával foglalkozunk röviden, annak is csak a legegyszerűbb esetével: feltesszük, hogy f egy téglalapon definiált. Legyen $f\colon [a,b] \times [c,d] \to \mathbb{R}$, és tekintsük az $[a,b]$ és $[c,d]$ intervallumok $a = x_0 < x_1 < \ldots < x_n = b$ és $c = y_0 < y_1 < \ldots < y_m = d$ beosztásait. Legyen $z_{ij} = f(x_i, y_j)$, $i = 0, \ldots, n$, $j = 0, \ldots, m$. Ezen adatok interpolációjára a következő függvényt használhatjuk:

$$L_{n,m}(x,y) \equiv \sum_{i=0}^{n}\sum_{j=0}^{m} z_{ij} l_i(x)\tilde{l}_j(y), \tag{6.5}$$

ahol l_i ill. $\tilde{l}_j$ az $a = x_0 < x_1 < \ldots < x_n = b$ ill. $c = y_0 < y_1 < \ldots < y_m = d$ alappontokhoz tartozó (6.2) képlettel definiált n ill. m-edrendű polinomok. Az így definiált $L_{n,m}$ függvény teljesíti az $L_{n,m}(x_i, y_j) = z_{ij}$ összefüggést minden i, j-re. Ha x-et rögzítjük, akkor $L_{n,m}(x, \cdot)$ egy legfeljebb m-edrendű polinom, és fordítva, ha y-t rögzítjük, akkor $L_{n,m}(\cdot, y)$ egy legfeljebb n-edrendű polinom.

6.9. példa. Tekintsük a következő függvényértékeket:

(x_i, y_j)	$(0,0)$	$(1,0)$	$(2,0)$	$(0,2)$	$(1,2)$	$(2,2)$
z_{ij}	2	-1	1	1	0	2

Alkalmazva az adatokra a (6.5) formulát kapjuk az

$$\begin{aligned} L_{2,1}(x,y) &= 2\frac{(x-1)(x-2)}{(0-1)(0-2)}\frac{y-2}{0-2} - \frac{x(x-2)}{1(1-2)}\frac{y-2}{0-2} + \frac{x(x-1)}{2(2-1)}\frac{y-2}{0-2} \\ &\quad + \frac{(x-1)(x-2)}{(0-1)(0-2)}\frac{y}{2} + 0\frac{x(x-2)}{1(1-2)}\frac{y}{2} + 2\frac{x(x-1)}{2(2-1)}\frac{y}{2} \\ &= -\frac{1}{2}x^2y + \frac{5}{2}x^2 + \frac{3}{2}xy - \frac{11}{2}x - \frac{1}{2}y + 2 \end{aligned}$$

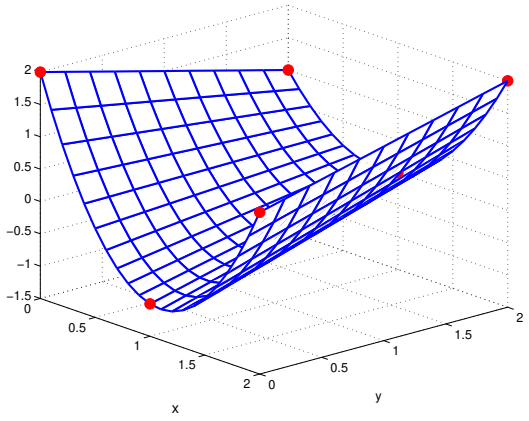
kétváltozós polinomot. Ez x-ben másodfokú, y-ban pedig elsőfokú polinom. Az interpolációs polinom grafikonja a 6.2. ábrán látható. □

Feladatok

1. Számítsa ki és ábrázolja az alábbi adatokhoz tartozó Lagrange-féle interpolációs polinomokat:

(a)

x_i	-1	0	2	4
y_i	3	-2	4	-2



6.2. ábra. Kétváltozós Lagrange-interpoláció

(b)

x_i	0.1	0.4	1.3	2.5	2.8
y_i	1.2	0.2	-2.2	3.1	1.3

(c)

x_i	-0.5	0.0	1.5	2.0	3.0	3.5
y_i	-0.5	1.5	3.5	2.0	2.5	6.5

2. Lássa be a Lagrange-polinom képletének megadása nélkül, hogy a (6.1) egyenletrendszernek létezik egyértelmű megoldása!
3. Legyen $l_i(x)$ $(i = 0, 1, \ldots, n)$ a (6.2) képlettel definiált n-edfokú polinom. Mutassa meg, hogy bármely x-re
$$\sum_{i=0}^{n} l_i(x) = 1.$$
4. Igazolja, hogy $(k+1)!(n-k)! \leq n!$ minden $k = 0, 1, \ldots, n-1$-re!
5. Mi az a legkisebb n, amelyre a $\cos x$ függvényt minden $x \in [-\pi, \pi]$-re 0.001-nél kisebb hibával lehet közelíteni az $L_n(x)$ interpolációs értékkel, ha ekvidisztáns osztópontokat használunk a $[-\pi, \pi]$ intervallumon?
6. Számítsa ki és ábrázolja az alábbi adatokhoz tartozó $L_{2,2}$ kétváltozós interpolációs polinomot:

(x_i, y_j)	$(0,0)$	$(0,1)$	$(0,2)$	$(1,0)$	$(1,1)$	$(1,2)$	$(2,0)$	$(2,1)$	$(2,2)$
z_{ij}	3	1	0	2	-1	0	2	3	1

6.2. Osztott differenciák

Adott egy $f\colon [a,b] \to \mathbb{R}$ függvény és $x_i \in [a,b]$ $(i = 0, \ldots, n)$ páronként különböző alappontok. Ekkor az f függvény x_0 pontbeli *nulladrendű osztott differenciáján* az $f[x_0] \equiv f(x_0)$ számot értjük. Az f függvény x_0, x_1 pontokra felírt *elsőrendű osztott differenciáján* az

$$f[x_0, x_1] \equiv \frac{f[x_1] - f[x_0]}{x_1 - x_0}$$

számot értjük, (azaz $f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$). Általában pedig, az f függvény $x_0, x_1, \ldots, x_n$ pontokra felírt *n-edrendű osztott differenciáján* az

$$f[x_0, x_1, \ldots, x_n] \equiv \frac{f[x_1, x_2, \ldots, x_n] - f[x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}$$

számot értjük. Megjegyezzük, hogy nem tettük fel, hogy az alappontok növekvő sorrendben rendezettek.

6.10. tétel. *Legyenek x_i $(i = 0, 1, \dots, n)$ páronként különböző alappontok. Ekkor*

$$f[x_0, x_1, \dots, x_n] = \sum_{i=0}^{n} \frac{f(x_i)}{(x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)}.$$

Bizonyítás. n-szerinti teljes indukcióval bizonyítjuk az állítást. $n = 0$-ra az állítás nyilvánvaló. (Ebben az esetben a nevezőben „üres szorzat" áll, ez definíció szerint 1-gyel egyezik meg.) Tegyük fel, hogy n-re teljesül az állítás, és tekintsük $f[x_0, x_1, \dots, x_{n+1}]$-et. Az osztott differenciák definíciója, az indukciós hipotézis és egy kis számolás alapján:

$$\begin{aligned}
&f[x_0, x_1, \dots, x_{n+1}] \\
&\quad= \frac{f[x_1, x_2, \dots, x_{n+1}] - f[x_0, x_1, \dots, x_n]}{x_{n+1} - x_0} \\
&\quad= \frac{1}{x_{n+1} - x_0}\left\{\sum_{i=1}^{n+1} \frac{f(x_i)}{(x_i - x_1)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_{n+1})}\right. \\
&\qquad\left. - \sum_{i=0}^{n} \frac{f(x_i)}{(x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)}\right\} \\
&\quad= \frac{1}{x_{n+1} - x_0}\left\{\frac{f(x_{n+1})}{(x_{n+1} - x_1)\cdots(x_{n+1} - x_n)} - \frac{f(x_0)}{(x_0 - x_1)\cdots(x_0 - x_n)}\right. \\
&\qquad + \sum_{i=1}^{n} \frac{f(x_i)}{(x_i - x_1)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)} \\
&\qquad\left. \cdot \left(\frac{1}{x_i - x_{n+1}} - \frac{1}{x_i - x_0}\right)\right\} \\
&\quad= \sum_{i=0}^{n+1} \frac{f(x_i)}{(x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_{n+1})},
\end{aligned}$$

amiből, a teljes indukció elve szerint, következik a tétel állítása. □

Az előző tétel állításából következnek:

6.11. következmény. *Az osztott differenciák az alappontok sorrendjétől függetlenek.*

6.12. következmény. *Ha f folytonos, akkor az osztott differencia az alappontoktól folytonosan függ.*

Tegyük fel, hogy f differenciálható függvény. Az utóbbi következmény szerint az $x_1 \mapsto f[x_0, x_1]$ függvény folytonos ha $x_1 \neq x_0$. Vizsgáljuk meg, hogy létezik-e a $\lim_{x_1 \to x_0} f[x_0, x_1]$ határérték! Az elsőrendű osztott differencia definícióját és f differenciálhatóságát használva

$$\lim_{x_1 \to x_0} f[x_0, x_1] = \lim_{x_1 \to x_0} \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f'(x_0).$$

Ezért az elsőrendű osztott differenciákat egyenlő alappontokra a következőképpen definiáljuk:

$$f[x_0, x_0] \equiv f'(x_0).$$

Ezzel a definícióval az $x_1 \mapsto f[x_0, x_1]$ függvényt folytonosan terjesztettük ki $x_1 = x_0$-ra. Magasabbrendű osztott differenciák egyenlő alappontokra kiterjesztésével a következő szakasz 6. és 7. feladatai foglalkoznak.

Feladatok

1. Számítsa ki a következő osztott differenciákat:
   (a) $f[x_0, x_1, x_2, x_3]$, ahol $x_i = i$, $f(x) = x^2$,
   (b) $f[x_0, x_1, x_2]$, ahol $x_i = 0.2i$, $f(x) = \sin x$,
   (c) $f[x_0, x_0]$, ahol $x_0 = 0$, $f(x) = \sin x$.
2. Legyen $f \in C^1(a, b)$, és $x_0, x_1 \in (a, b)$, $x_0 \neq x_1$. Bizonyítsa be, hogy létezik olyan $\xi \in \langle x_0, x_1 \rangle$, hogy
   $$f[x_0, x_1] = f'(\xi)!$$
3. Legyen $x_0 < x_1 < x_2 < x_3$ és
   $$P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + a_3(x - x_0)(x - x_1)(x - x_2).$$
   Lássa be, hogy
   $$a_0 = P[x_0], \quad a_1 = P[x_0, x_1], \quad a_2 = P[x_0, x_1, x_2], \quad \text{és} \quad a_3 = P[x_0, x_1, x_2, x_3]!$$

6.3. A Lagrange-féle interpolációs polinom Newton-féle alakja

A (6.3) képletnek van egy kellemetlen hátránya: új osztópont felvételekor teljesen újra kell számolni a (6.3) kifejezést. Ezt a hiányosságot küszöböli ki a Lagrange-polinom egy másik alakja, az ún. Newton-féle alak. Tegyük fel, hogy az f függvényt akarjuk interpolálni, azaz $y_i = f(x_i)$. A Lagrange-féle interpolációs polinom Newton-féle alakjának levezetéséhez induljunk ki az

$$L_n(x) = L_0(x) + (L_1(x) - L_0(x)) + (L_2(x) - L_1(x)) + \cdots + (L_n(x) - L_{n-1}(x))$$

összefüggésből. Definíció szerint $L_0(x) = f(x_0)$ konstans függvény. Vizsgáljuk most az $L_i(x) - L_{i-1}(x)$ különbséget! $L_i - L_{i-1}$ egy legfeljebb i-edfokú polinom, és mivel L_i és L_{i-1} is teljesítik az interpolációs egyenletetet x_0, …, x_{i-1}-ben, ezért $L_i(x_j) - L_{i-1}(x_j) = f(x_j) - f(x_j) = 0$ $(j = 0, 1, \ldots, i-1)$. De ekkor az algebra alaptétele szerint $L_i - L_{i-1}$ alakja:

$$L_i(x) - L_{i-1}(x) = a_i(x - x_0)(x - x_1)\cdots(x - x_{i-1}),$$

ahol $a_i \in \mathbb{R}$. Ha ebbe a relációba $x = x_i$-t helyettesítünk és használjuk $L_{i-1}(x_i)$-re a (6.3) képletet, kapjuk, hogy

$$f(x_i) - \sum_{k=0}^{i-1} f(x_k) \frac{(x_i - x_0)\cdots(x_i - x_{k-1})(x_i - x_{k+1})\cdots(x_i - x_{i-1})}{(x_k - x_0)\cdots(x_k - x_{k-1})(x_k - x_{k+1})\cdots(x_k - x_{i-1})} = a_i(x_i - x_0)\cdots(x_i - x_{i-1}).$$

Ebből a_i-t kifejezve

$$\begin{aligned} a_i &= \frac{f(x_i)}{(x_i-x_0)\cdots(x_i-x_{i-1})} - \frac{1}{(x_i-x_0)\cdots(x_i-x_{i-1})} \\ &\qquad \cdot \sum_{k=0}^{i-1} f(x_k) \frac{(x_i-x_0)\cdots(x_i-x_{k-1})(x_i-x_{k+1})\cdots(x_i-x_{i-1})}{(x_k-x_0)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_{i-1})} \\ &= \sum_{k=0}^{i} \frac{f(x_k)}{(x_k-x_0)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_i)} \\ &= f[x_0,x_1,\ldots,x_i]. \end{aligned}$$

Összefoglalva az eddigieket, a Lagrange-féle interpolációs polinomot megadhatjuk az

$$\begin{aligned} L_n(x) &= f[x_0] + f[x_0,x_1](x-x_0) + f[x_0,x_1,x_2](x-x_0)(x-x_1) + \cdots \\ &\qquad + f[x_0,x_1,\ldots,x_n](x-x_0)(x-x_1)\cdots(x-x_{n-1}) \end{aligned} \tag{6.6}$$

képlettel is. Hangsúlyozzuk, hogy ez ugyanaz a polinom, mint (6.3), csak egy másik alakban felírva. A (6.6) formulával definiált polinomot nevezzük a *Lagrange-féle interpolációs polinom Newton-féle alakjának* vagy röviden *Newton-polinomnak*.

A (6.6) képletből leolvasható ennek a formulának az előnye a (6.3) képlethez viszonyítva. Először is, új osztópont hozzávételével a képlet kényelmesen bővíthető egy új taggal:

$$L_{n+1}(x) = L_n(x) + f[x_0,x_1,\ldots,x_{n+1}](x-x_0)\cdots(x-x_n).$$

Fontos előny még az is, hogy a (6.6) alakban felírt polinomot könnyen kiértékelhetjük a Horner-elrendezés segítségével. Ebből az alakból rögtön leolvasható a polinom fokszáma is. Ha pl. $f[x_0,x_1,\ldots,x_n] \neq 0$, akkor a polinom n-edfokú. A 6.13. algoritmusban megadtuk a Newton-féle interpolációs polinom együtthatóinak, azaz az $a_i = f[x_0,\ldots,x_i]$ értékek kiszámítását, a 6.14. algoritmusban pedig a Newton-polinom kiértékelését Horner-eljárással.

6.13. algoritmus. A Newton-polinom együtthatóinak generálása

INPUT: n - az alappontok száma − 1
x_i, $(i = 0,1,\ldots,n)$ - alappontok
y_i, $(i = 0,1,\ldots,n)$ - függvényértékek
OUTPUT: a_i, $(i = 0,1,\ldots,n)$ - a Newton-polinom együtthatói, ahol a_i az i-edfokú tag együtthatója

for $i = 0,1,\ldots,n$ **do**
  $a_i \leftarrow y_i$
end do
for $j = 1,2,\ldots,n$ **do**
  for $i = n,n-1,\ldots,j$ **do**
    $a_i \leftarrow (a_i - a_{i-1})/(x_i - x_{i-j})$
  end do
end do
output$(a_0,a_1,\ldots,a_n)$

Megjegyezzük, hogy a 6.13. algoritmust úgy szerveztük, hogy a Newton-polinom felírása közben számolt osztott differenciák közül csak az együtthatókhoz szükségeseket őrizzük meg a számolás végéig.

6.14. algoritmus. A Newton-polinom kiértékelése

INPUT: n - az alappontok száma $-$ 1
x_i, $(i = 0, 1, \ldots, n)$ - alappontok
a_i, $(i = 0, 1, \ldots, n)$ - a Newton-polinom együtthatói
x - a pont, ahol kiértékeljük a Newton-polinomot
OUTPUT: y - a Newton-polinom értéke x-ben

$y \leftarrow a_n$
for $i = n-1, n-2, \ldots, 0$ **do**
$\quad y \leftarrow y(x - x_i) + a_i$
end do
output(y)

Kézi számoláskor az osztópontokat, a megadott függvényértékeket és a számított osztott differenciákat érdemes a 6.1. táblázatban látható módon egy háromszög alakú táblázatban elrendezni. A táblázat első két oszlopában szereplő számok input adatok, a táblázat többi elemét számoljuk a tőle balra álló és az a fölötti eggyel kisebb rendű osztott differenciák különbségét osztva megfelelő x_k értékek különbségének hányadosaként. A táblázatban a bekeretezett számok fogják adni a (6.6) képletben szereplő együtthatókat.

6.1. táblázat. Osztott differenciák elrendezése kézi számoláskor

x_0	$\boxed{f(x_0)}$				
x_1	$f(x_1)$	$\boxed{f[x_0, x_1]}$			
x_2	$f(x_2)$	$f[x_1, x_2]$	$\boxed{f[x_0, x_1, x_2]}$		
x_3	$f(x_3)$	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	$\ddots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$		
x_n	$f(x_n)$	$f[x_{n-1}, x_n]$	$f[x_{n-2}, x_{n-1}, x_n]$	$\cdots$	$\boxed{f[x_0, x_1, \ldots, x_n]}$

6.15. példa. Tekintsük újra a 6.2. példát. Adjuk meg $L_3(x)$ Newton-féle alakját, majd számítsuk ki $L_3(0)$-t! Képezzük a Newton-polinom felírásához szükséges osztott differenciák táblázatát:

$$\begin{array}{rrrrr} -1 & -3 & & & \\ 1 & 1 & 2 & & \\ 2 & 3 & 2 & 0 & \\ 3 & 29 & 26 & 12 & 3 \end{array}$$

Ebből kapjuk, hogy

$$L_3(x) = -3 + 2(x+1) + 3(x+1)(x-1)(x-2),$$

és így $L_3(0) = -3 + 2 \cdot 1 + 3 \cdot 1(-1)(-2) = 5$. Természetesen egyszerűsítve L_3 képletét visszakapjuk a 6.2. példában kiszámolt $L_3(x) = 3x^3 - 6x^2 - x + 5$ képletet. □

Most az interpoláció képlethibájával foglalkozunk újra. A 6.1. szakaszban megállapítottuk, hogy a közelítés hibája az $\frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n)$ alakban írható fel. Ez a képlet természetesen érvényes a Newton-alakban felírt interpolációs polinomot használva is, de itt megadjuk a képlethiba egy másik alakját is.

6.16. tétel. *Legyenek $x_i \in (a,b)$ $(i=0,\dots,n)$ páronként különböző alappontok és $y_i = f(x_i)$ $(i=0,\dots,n)$. Legyen $L_n(x)$ az adatokhoz tartozó n-edfokú Lagrange-polinom. Ekkor*

$$f(x) = L_n(x) + f[x_0, x_1, \dots, x_n, x](x-x_0)(x-x_1)\cdots(x-x_n).$$

Bizonyítás. Rögzítsünk egy $x \in (a,b)$ számot amely nem egyezik meg egyik alapponttal sem. (Ha $x = x_i$ valamely i-re, akkor az állítás nyilvánvaló.) Vegyük x-et az alappontokhoz és rendeljük hozzá az $f(x)$ függvényértéket. Legyen L_{n+1} a kibővitett adatokhoz tartozó Lagrange-polinom. A Newton-polinom definíciója szerint

$$L_{n+1}(t) = L_n(t) + f[x_0, x_1, \dots, x_n, x](t-x_0)\cdots(t-x_n).$$

Ebből $t = x$-et véve következik az állítás, hiszen $f(x) = L_{n+1}(x)$. □

Az interpoláció képlethibájának a 6.16. tételben közölt alakja elsősorban elméleti jelentőségű, hiszen $f[x_0,\dots,x_n,x]$ kiszámításához $f(x)$ ismerete is kell. Fontos viszont a tétel következménye. Ha összehasonlítjuk az előző tétel állítását a 6.5. tételével, akkor rögtön kapjuk a következő eredményt:

6.17. következmény. *Ha $f \in C^n(a,b)$ és x_i $(i=0,\dots,n)$ páronként különböző alappontok, akkor létezik olyan $\xi \in \langle x_0, x_1, \dots, x_n\rangle$, hogy*

$$f[x_0, x_1, \dots, x_n] = \frac{1}{n!} f^{(n)}(\xi).$$

Feladatok

1. Ismételje meg a 6.1. szakasz 1. feladatát a Lagrange-polinom Newton-féle alakját használva!
2. Igazolja, hogy ha P egy n-edfokú polinom, akkor
$$P(x) = \sum_{i=0}^{n} P[x_0,\dots,x_i] \prod_{k=0}^{i-1}(x-x_k).$$
3. Legyenek $x_0,\dots,x_n$ páronként különböző számok. Igazolja, hogy ha P egy n-edfokú polinom, akkor $P[x_0,\dots,x_m] = 0$ minden $m > n$-re!
4. Mutassa meg, hogy ha $f(x) = c_0 + c_1 x + \cdots + c_n x^n$, akkor $c_n = f[x_0, x_1, \dots, x_n]$!
5. Bizonyítsa be, hogy
$$f[x_0,x_1,\dots,x_n] = \frac{\begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^{n-1} & f(x_0) \\ 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} & f(x_1) \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} & f(x_n) \end{vmatrix}}{\begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^{n-1} & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} & x_1^n \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} & x_n^n \end{vmatrix}}!$$
6. Mutassa meg, hogy
$$\lim_{(x_1,x_2,\dots,x_n)\to(x_0,x_0\dots,x_0)} f[x_0,x_1,\dots,x_n] = \frac{f^{(n)}(x_0)}{n!}!$$
(Útmutatás: Használja a 6.17. következményt!)

7. Legyen $f \in C^2$. Definiáljuk a következő osztott differenciákat:

$$f[x_0, x_0, x_1] \equiv \lim_{x_2 \to x_0} f[x_0, x_2, x_1], \quad f[x_0, x_1, x_0] \equiv \lim_{x_2 \to x_0} f[x_0, x_1, x_2],$$

és

$$f[x_1, x_0, x_0] \equiv \lim_{x_2 \to x_0} f[x_1, x_0, x_2], \qquad f[x_0, x_0, x_0] = \frac{f''(x_0)}{2}!$$

Mutassa meg, hogy az előbbi határértékek léteznek, és az így definiált másodrendű osztott differenciák megőrzik a páronként különböző alappontokra felírt osztott differenciák szokásos tulajdonságait:

(a) $f[x_0, x_0, x_1] = \dfrac{f[x_0, x_1] - f[x_0, x_0]}{x_1 - x_0}$,

(b) $f[x_1, x_0, x_0] = \dfrac{f[x_0, x_0] - f[x_1, x_0]}{x_0 - x_1}$,

(c) $f[x_0, x_0, x_1] = f[x_0, x_1, x_0] = f[x_1, x_0, x_0]$,

(d) $\lim\limits_{(x_1, x_2) \to (x_0, x_0)} f[x_0, x_1, x_2] = f[x_0, x_0, x_0]$,

(e) Létezik olyan $\xi \in \langle x_0, x_1 \rangle$, hogy $f[x_0, x_0, x_1] = f''(\xi)/2$.

8. Ellenőrizze, hogy a 6.13. algoritmus valóban visszaadja a Newton-polinom együtthatóit!

6.4. Hermite-interpoláció

Ebben a szakaszban az interpoláció alapfeladatát módosítjuk. Legyen adott egy f differenciálható függvény, és osztópontoknak egy x_i $(i = 0, \ldots, n)$ véges sorozata. Az ún. *Hermite-féle interpolációs feladatban* azon kívül, hogy az $y_i = f(x_i)$ függvényértékeket interpoláljuk, az $y_i' \equiv f'(x_i)$ derivált értékeket is szeretnénk interpolálni. Keresünk tehát egy olyan $g(x) = c_0 + c_1 x + \cdots + c_m x^m$ polinomot, amelyre

$$g(x_i) = y_i, \qquad g'(x_i) = y_i', \qquad i = 0, 1, \ldots, n$$

teljesül. A feladat geometriai jelentése az, hogy olyan polinomot keresünk, amelynek grafikonja a megadott irányokban megy át az adott (x_i, y_i) pontokon, azaz az érintőjének iránytangense megegyezik az y_i' értékekkel. A g függvény képletében $m + 1$ db paraméter szerepel, az előző feltételek $2(n+1)$ egyenletet határoznak meg, így azt várjuk, hogy $m = 2n+1$-edfokú polinomok között találunk egyértelmű megoldását az Hermite-féle interpolációs problémának. A következő tételben ezt be is látjuk. Az Hermite-féle interpolációs probléma megoldását *Hermite-féle interpolációs polinomnak* vagy röviden *Hermite-polinomnak* nevezzük, és H_{2n+1}-gyel jelöljük.

A következő tételben szükségünk lesz olyan magasabbrendű speciális osztott differenciákra, ahol az egymás után következő két alappontok megegyeznek: $f[x_0, x_0, x_1, x_1, \ldots, x_n, x_n]$, ahol $x_0, x_1, \ldots, x_n$ páronként különböznek. Ezeket az osztott differenciákat a szokásos rekurzív definícióval értelmezhetjük eggyel alacsonyabb fokú osztott differenciák segítségével:

$$f[x_0, x_0, x_1, x_1, \ldots, x_n, x_n] = \frac{f[x_0, x_1, x_1, \ldots, x_n, x_n] - f[x_0, x_0, x_1, x_1, \ldots, x_n]}{x_n - x_0}$$

Az alacsonyabb fokú osztott differenciákat is ehhez hasonlóan definiáljuk, és ezt folytathatjuk egészen addig, amíg vagy különböző vagy egyenlő alappontokra felírt elsőrendű osztott differenciákig nem jutunk vissza, amelyeket már definiáltuk a 6.2. szakaszban.

6.18. tétel. *Az Hermite-féle interpolációs feladatnak létezik egyértelmű megoldása a legfeljebb $(2n+1)$-edfokú polinomok körében, amelyet a*

$$\begin{aligned} H_{2n+1}(x) \\ = \; & f[x_0] + f[x_0,x_0](x-x_0) + f[x_0,x_0,x_1](x-x_0)^2 \\ & + f[x_0,x_0,x_1,x_1](x-x_0)^2(x-x_1) + f[x_0,x_0,x_1,x_1,x_2](x-x_0)^2(x-x_1)^2 \\ & + f[x_0,x_0,x_1,x_1,x_2,x_2](x-x_0)^2(x-x_1)^2(x-x_2) + \cdots \\ & + f[x_0,x_0,x_1,x_1,\ldots,x_n,x_n](x-x_0)^2(x-x_1)^2\cdots(x-x_{n-1})^2(x-x_n) \end{aligned} \tag{6.7}$$

alakban adhatunk meg. Továbbá a közelítés képlethibája

$$f(x) - H_{2n+1}(x) = f[x_0,x_0,\ldots,x_n,x_n,x](x-x_0)^2\cdots(x-x_n)^2. \tag{6.8}$$

Bizonyítás. Először vizsgáljuk az Hermite-polinom egyértelműségét. Tegyük fel, hogy H_{2n+1} és $\tilde{H}_{2n+1}$ legyfeljebb $(2n+1)$-edfokú polinomok, amelyek teljesítik az Hermite-féle interpolációs feltételeket. Ekkor $P \equiv H_{2n+1} - \tilde{H}_{2n+1}$ is egy legfeljebb $(2n+1)$-edfokú polinom, amelyre $P(x_i) = H_{2n+1}(x_i) - \tilde{H}_{2n+1}(x_i) = f(x_i) - f(x_i) = 0$, és $P'(x_i) = H'_{2n+1}(x_i) - \tilde{H}'_{2n+1}(x_i) = f'(x_i) - f'(x_i) = 0$, azaz x_i kétszeres gyöke P-nek minden $i = 0,1,\ldots,n$-re. P-nek van tehát $2(n+1) = 2n+2$ gyöke, amiből következik az algebra alaptétele szerint, hogy P azonosan 0 polinom, hiszen P legfeljebb $(2n+1)$-edfokú. Ebből következik, hogy az Hermite-féle interpolációs feladatnak legfeljebb egy $(2n+1)$-edfokú megoldása lehet.

Most belátjuk, hogy a (6.7) képlettel definiált H_{2n+1} polinom megoldása az Hermite-féle interpolációs feladatnak, és teljesíti a (6.9) hibaformulát. Direkt számolással rögtön kapjuk, hogy $H_{2n+1}(x_0) = f(x_0)$ és $H'_{2n+1}(x_0) = f[x_0,x_0] = f'(x_0)$. Következő lépésként belátjuk, hogy $H_{2n+1}(x_1) = f(x_1)$ és $H'_{2n+1}(x_1) = f'(x_1)$ is teljesül. Ehhez válasszunk olyan x_i-hez közeli $\tilde{x}_i$ számokat, hogy $\{x_i, \tilde{x}_i : \; i = 0,1,\ldots,n\}$ páronként különbözőek legyenek, és legyen L_{2n+1} ezekhez az alappontokhoz tartozó, f-et interpoláló Lagrange-féle interpolációs polinom. Ekkor

$$\begin{aligned} L_{2n+1}(x) \;=\; & f[x_0] + f[x_0,x'_0](x-x_0) + f[x_0,x'_0,x_1](x-x_0)(x-x'_0) \\ & + f[x_0,x'_0,x_1,x'_1](x-x_0)(x-x'_0)(x-x_1) + \cdots \\ & + f[x_0,x'_0,x_1,x'_1,\ldots,x_n,x'_n](x-x_0)(x-x'_0)\cdots(x-x_{n-1}) \\ & \quad \cdot (x-x'_{n-1})(x-x_n), \end{aligned}$$

és

$$f(x) = L_{2n+1}(x) + f[x_0,x'_0,\ldots,x_n,x'_n,x](x-x_0)(x-x'_0)\cdots(x-x_n)(x-x'_n).$$

L_{2n+1} és H_{2n+1} definíciójából és az osztott differencia folytonosságából (lásd a 3. feladatot) kapjuk, hogy minden x-re

$$L_{2n+1}(x) \to H_{2n+1}(x) \quad \text{ha } (x'_0,x'_1,\ldots,x'_n) \to (x_0,x_1,\ldots,x_n), \tag{6.9}$$

és így

$$f(x) = H_{2n+1}(x) + f[x_0,x_0,x_1,x_1,\ldots,x_n,x_n,x](x-x_0)^2(x-x_1)^2\cdots(x-x_n)^2.$$

Ez igazolja a (6.8) összefüggést. A Lagrange-féle interpolációs polinom egyértelműségéből következik, hogy ha x_0, x'_0 és x_1, x'_1 sorrendjét felcseréljük, az interpolációs polinom nem fog változni, azaz

$$\begin{aligned} L_{2n+1}(x) \\ = \; & f[x_1] + f[x_1,x'_1](x-x_1) + f[x_1,x'_1,x_0](x-x_1)(x-x'_1) \\ & + f[x_1,x'_1,x_0,x'_0](x-x_1)(x-x'_1)(x-x_0) + \cdots \\ & + f[x_1,x'_1,x_0,x'_0,x_2,x'_2\ldots,x_n,x'_n](x-x_1)(x-x'_1)(x-x_0)(x-x'_0) \\ & \quad \cdot (x-x_2)(x-x'_2)\cdots\ (x-x_{n-1})(x-x'_{n-1})(x-x_n). \end{aligned}$$

Ebből viszont kapjuk, mindkét oldal határértékét véve, ha $(x_0', x_1', \dots, x_n') \to (x_0, x_1, \dots, x_n)$, és használva a (6.9) összefüggést és a határérték egyértelműségét, hogy

$$\begin{aligned} & H_{2n+1}(x) \\ &= f[x_1] + f[x_1, x_1](x - x_1) + f[x_1, x_1, x_0](x - x_1)^2 \\ &\quad + f[x_1, x_1, x_0, x_0](x - x_1)^2(x - x_0) + f[x_1, x_1, x_0, x_0, x_2](x - x_1)^2(x - x_0)^2 \\ &\quad + f[x_1, x_1, x_0, x_0, x_2, x_2](x - x_1)^2(x - x_0)^2(x - x_2) + \cdots \\ &\quad + f[x_1, x_1, x_0, x_0, x_2, x_2, \dots, x_n, x_n](x - x_1)^2(x - x_0)^2(x - x_2)^2 \\ &\qquad \cdots (x - x_{n-1})^2(x - x_n) \end{aligned}$$

alakban *is* felírható. Ebből viszont nyilvánvaló, hogy $H_{2n+1}(x_1) = f(x_1)$ és $H'_{2n+1}(x_1) = f'(x_1)$. Ehhez hasonlóan látható be, hogy $H_{2n+1}(x_i) = f(x_i)$ és $H'_{2n+1}(x_i) = f'(x_i)$ teljesül $i = 2, 3, \dots, n$-re is. □

6.19. tétel. *Legyen $f \in C^{2n+2}$. Ekkor létezik olyan $\xi \in \langle x_0, x_1, \dots, x_n, x\rangle$, hogy*

$$f(x) - H_{2n+1}(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!}(x - x_0)^2 \dots (x - x_n)^2.$$

Bizonyítás. A bizonyítás hasonló a 6.5. tétel bizonyításához. Legyen x egy osztópontoktól különböző rögzített szám, és definiáljuk a

$$g(z) = f(z) - H_{2n+1}(z) - \frac{(z - x_0)^2 \cdots (z - x_n)^2}{(x - x_0)^2 \cdots (x - x_n)^2}(f(x) - H_{2n+1}(x))$$

függvényt. Nyilván $g \in C^{2n+2}$, és $x_0, \dots, x_n$ kétszeres gyökei, x pedig egyszeres gyöke g-nek. Ezért az általánosított Rolle-tétel (6.4 tétel) szerint létezik olyan $\xi \in \langle x_0, x_1, \dots, x_n, x\rangle$, hogy $g^{(2n+2)}(\xi) = 0$. Ebből pedig következik a tétel állítása.

□

A (6.8) összefüggést és a 6.19. tételt összehasonlítva rögtön kapjuk:

6.20. következmény. *Tegyük fel, hogy $f \in C^{2n+2}$ és $x, x_0, \dots, x_n$ páronként különböző számok. Ekkor létezik olyan $\xi \in \langle x_0, x_1, \dots, x_n, x\rangle$, hogy*

$$f[x_0, x_0, \dots, x_n, x_n, x] = \frac{f^{(2n+2)}(\xi)}{(2n+2)!}.$$

Kézi számoláskor a (6.8) képlethez szükséges osztott differenciákat a 6.2. táblázat segítségével számolhatjuk ki. Megjegyezzük, hogy ez a táblázat nagyon hasonlít a 6.1. táblázathoz. A különbség az, hogy minden alappont és a hozzá tartozó függvényérték kétszer szerepel benne, és a harmadik oszlopban az azonos alappontokra felírt elsőrendű osztott differenciák is előre adottak, a megadott derivált értékkel egyeznek meg. A táblázat többi elemét ugyanúgy számítjuk, mint a 6.1. táblázatban. A bekeretezett számok fogják adni a (6.8) képletben szereplő együtthatókat.

6.2. táblázat. Osztott differenciák elrendezése kézi számoláskor

$$
\begin{array}{llllll}
x_0 & \boxed{f(x_0)} & & & & \\
x_0 & f(x_0) & \boxed{f[x_0,x_0]} & & & \\
x_1 & f(x_1) & f[x_0,x_1] & \boxed{f[x_0,x_0,x_1]} & & \\
x_1 & f(x_1) & f[x_1,x_1] & f[x_0,x_1,x_1] & \ddots & \\
\vdots & \vdots & \vdots & \vdots & & \\
x_n & f(x_n) & f[x_{n-1},x_n] & f[x_{n-1},x_{n-1},x_n] & \cdots & \\
x_n & f(x_n) & f[x_n,x_n] & f[x_{n-1},x_n,x_n] & \cdots & \boxed{f[x_0,x_0,x_1,x_1\ldots,x_n,x_n]}
\end{array}
$$

6.21. példa. Tekintsük a következő adatokat:

x_i	-1	1	2
y_i	2	4	11
y_i'	3	-5	30

Keressük meg az adatokat interpoláló Hermit-féle interpolációs polinomot! Készítsük el a következő táblázatot:

$$
\begin{array}{llllllll}
-1 & 2 & & & & & & \\
-1 & 2 & \boxed{3} & & & & & \\
1 & 4 & 1 & -1 & & & & \\
1 & 4 & \boxed{-5} & -3 & -1 & & & \\
2 & 11 & 7 & 12 & 5 & 2 & & \\
2 & 11 & \boxed{30} & 23 & 11 & 2 & 0
\end{array}
$$

(A harmadik oszlopban bekereteztük az inputként megadott derivált értékeket.) Az Hermite-polinom tehát

$$H_5(x) = 2 + 3(x+1) - (x+1)^2 - (x+1)^2(x-1) + 2(x+1)^2(x-1)^2 = 2x^4 - x^3 - 6x^2 + 2x + 7,$$

azaz H_5 jelen esetben egy negyedfokú polinom. □

Feladatok

1. Számítsa ki és ábrázolja az alábbi adatokhoz tartozó Hermite-féle interpolációs polinomokat:

(a)

x_i	-2	-1	0	1
y_i	4	1	14	-35
y_i'	-1	-2	43	-394

(b)

x_i	-1	0	2	3
y_i	1	2	64	-19
y_i'	3	-1	111	-301

2. Bizonyítsa be, hogy ha P egy legfeljebb $(2n+2)$-edfokú polinom, x_i $(i = 0, 1, \ldots, n)$ páronként különböző alappontok, és H_{2n+1} a P-hez és az alappontokhoz tartozó Hermite-polinom, akkor $P(x) = H_{2n+1}(x)$ minden x-re!
3. Legyen $f \in C^1$. Bizonyítsa be, hogy

$$\lim_{(x_0',x_1',\ldots,x_n')\to(x_0,x_1,\ldots,x_n)} f[x_0,x_0',x_1,x_1',\ldots,x_n,x_n'] = f[x_0,x_0,x_1,x_1,\ldots,x_n,x_n]$$

és

$$\lim_{(x_0',\ldots,x_{n-1}')\to(x_0,\ldots,x_{n-1})} f[x_0,x_0',x_1,x_1',\ldots,x_{n-1},x_{n-1}',x_n] = f[x_0,x_0,x_1,x_1,\ldots,x_{n-1},x_{n-1},x_n]!$$

4. Legyen $i_0, i_1, \ldots, i_n$ a $0, 1, \ldots, n$ véges számsorozatnak egy átrendezése. Lássa be, hogy ekkor
$$f[x_0, x_0, x_1, x_1, \ldots, x_n, x_n] = f[x_{i_0}, x_{i_0}, x_{i_1}, x_{i_1}, \ldots, x_{i_n}, x_{i_n}]!$$

5. Az Hermite-interpolációs feladatot általánosabban is meg lehet fogalmazni: az i-edik osztópontban a függvényérték és az első k_i derivált érték adott, amelyeket interpolálni szeretnénk. Erre a feladatra könnyen általánosítható az ebben a szakaszban tárgyalt módszer. Illusztrálásként tekintsünk most egy konkrét, egyszerű feladatot: adott két osztópont, x_0 és x_1, és egy $f \in C^3$ függvény. Keresünk egy olyan minimális fokszámú polinomot, amelyre
$$H(x_0) = f(x_0), \quad H'(x_0) = f'(x_0), \quad H''(x_0) = f''(x_0), \quad \text{és} \quad H(x_1) = f(x_1).$$
(Itt $k_0 = 2$ és $k_1 = 0$.) Lássa be, hogy a feladat megoldása a
$$H(x) \equiv f[x_0] + f[x_0, x_0](x - x_0) + f[x_0, x_0, x_0](x - x_0)^2 + f[x_0, x_0, x_0, x_1](x - x_0)^3$$
legfeljebb harmadfokú polinom!

6.5. Spline interpoláció

Legyen $a = x_0 < x_1 < \ldots < x_n = b$ az $[a, b]$ intervallumnak egy felosztása. Az $S\colon [a, b] \to \mathbb{R}$ folytonos függvényt az $\{x_i\}$ osztópontokhoz tartozó *k-adrendű spline* függvénynek nevezzük, ha $S \in C^{k-1}(a, b)$, és S megszorítása minden $[x_i, x_{i+1}]$ intervallumra egy k-adrendű polinom. Az elsőrendű, másodrendű ill. harmadrendű spline függvényeket *lineáris, kvadratikus, ill. kubikus spline függvényeknek* is nevezzük.

A legegyszerűbb, és így a gyakorlatban igen gyakran használt interpolációs módszer lineáris spline-okkal interpolálja a megadott adatokat. Geometriailag ez azt jelenti, hogy a megadott (x_i, y_i) pontokat szakaszokkal kötjük össze. A lineáris spline interpolációval elkövetett hiba becslésével a 2. feladat foglalkozik.

A lineáris spline interpoláció hátránya az, hogy az interpolációs függvény nem sima, azaz nem differenciálható. Ezt a hátrányt kiküszöböli a harmadrendű spline interpoláció. Ekkor az interpolációs függvény kétszer folytonosan differenciálható lesz, ami a gyakorlati alkalmazásoknál többnyire elegendő. A szakasz hátralevő részében a harmadrendű spline interpolációval foglalkozunk.

Adott osztópontoknak egy $a = x_0 < x_1 < \ldots < x_n = b$, és hozzá tartozó $y_0, y_1, \ldots, y_n$ függvényértékek véges sorozata. Keresünk egy olyan S harmadrendű spline függvényt, amely interpolálja a megadott adatokat, azaz
$$S(x_i) = y_i, \qquad i = 0, 1, \ldots, n.$$

Jelöljük S_i-vel az S függvény $[x_i, x_{i+1}]$ intervallumra vett megszorítását ($i = 0, 1, \ldots, n-1$). A feltevés szerint S interpolálja az (x_i, y_i) pontokat és kétszer folytonosan differenciálható, ezért az S_i függvények teljesítik a következő feltételeket:
$$\begin{aligned}
S_i(x_i) &= y_i, && i = 0, 1, \ldots, n-1, && (6.10)\\
S_i(x_{i+1}) &= y_{i+1}, && i = 0, 1, \ldots, n-1, && (6.11)\\
S_i'(x_{i+1}) &= S_{i+1}'(x_{i+1}), && i = 0, 1, \ldots, n-2, && (6.12)\\
S_i''(x_{i+1}) &= S_{i+1}''(x_{i+1}), && i = 0, 1, \ldots, n-2. && (6.13)
\end{aligned}$$

Mivel minden egyes S_i függvényt 4 paraméter határoz meg, így összesen $4n$ paraméter definiálja S-t. A (6.10)–(6.13) feltételek száma viszont csak $4n-2$, ezért a feladatnak így nem egyértelmű a megoldása. Ezért várhatóan még két feltételt megadhatunk, és ettől remélhetően egyértelmű megoldást kapunk. Egy gyakran használt feltétel a következő:
$$S_0''(x_0) = 0 \qquad \text{és} \qquad S_{n-1}''(x_n) = 0. \tag{6.14}$$

A (6.10)–(6.14) feltételekkel definiált kubikus spline függvényt *természetes spline* függvénynek nevezzük. Belátjuk, hogy az interpolációs feladatnak pontosan egy természetes spline függvény megoldása van. Vegyük fel S_i-t a következő alakban:

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

ahol a_i, b_i, c_i és d_i $(i = 0, 1, \ldots, n-1)$ meghatározandó paraméterek. Ekkor

$$\begin{aligned} S_i'(x) &= b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2, \\ S_i''(x) &= 2c_i + 6d_i(x - x_i). \end{aligned}$$

Ezekből az összefüggésekből rögtön következik

$$a_i = S_i(x_i) = y_i, \quad b_i = S_i'(x_i) \quad \text{és} \quad c_i = S_i''(x_i)/2, \quad i = 0, 1, \ldots, n-1. \tag{6.15}$$

A (6.15) összefüggések segítségével definiálhatjuk az a_n, b_n és c_n konstansokat is (amelyekre később szükségünk lesz):

$$a_n \equiv y_n, \qquad b_n \equiv S'(x_n) \qquad \text{és} \qquad c_n \equiv S''(x_n)/2. \tag{6.16}$$

(A (6.16) képletekben a deriváltak bal oldali deriváltakat jelentenek.) $x = x_{i+1}$-t behelyettesítve S_i képletébe és a (6.11) egyenletet, valamint az $a_i = y_i$ összefüggést használva kapjuk

$$y_i + b_i(x_{i+1} - x_i) + c_i(x_{i+1} - x_i)^2 + d_i(x_{i+1} - x_i)^3 = y_{i+1}.$$

Vezessük be a $\Delta x_i \equiv x_{i+1} - x_i$ és a $\Delta y_i \equiv y_{i+1} - y_i$ jelöléseket. Így

$$b_i \Delta x_i + c_i(\Delta x_i)^2 + d_i(\Delta x_i)^3 = \Delta y_i, \qquad i = 0, 1, \ldots, n-1. \tag{6.17}$$

A (6.12) feltételből és a $b_{i+1} = S_{i+1}'(x_{i+1})$ összefüggésből

$$b_i + 2c_i\Delta x_i + 3d_i(\Delta x_i)^2 = b_{i+1} \tag{6.18}$$

minden $i = 0, 1, \ldots, n-2$-re. Használva b_n definícióját kapjuk, hogy (6.18) teljesül $i = n-1$-re is. Hasonlóan, a (6.13) egyenletből és c_n definíciójából következik

$$2c_i + 6d_i\Delta x_i = 2c_{i+1}, \qquad i = 0, 1, \ldots, n-1,$$

amiből

$$d_i = \frac{c_{i+1} - c_i}{3\Delta x_i}, \qquad i = 0, 1, \ldots, n-1. \tag{6.19}$$

Ezt behelyettesítjük a (6.17) és (6.18) egyenletekbe:

$$\begin{aligned} b_i\Delta x_i + c_i(\Delta x_i)^2 + \frac{c_{i+1} - c_i}{3}(\Delta x_i)^2 &= \Delta y_i, \qquad i = 0, 1, \ldots, n-1, & (6.20) \\ b_i + 2c_i\Delta x_i + (c_{i+1} - c_i)\Delta x_i &= b_{i+1}, \qquad i = 0, 1, \ldots, n-1. & (6.21) \end{aligned}$$

Az első egyenletből kifejezve b_i-t

$$b_i = \frac{\Delta y_i}{\Delta x_i} - \frac{2c_i + c_{i+1}}{3}\Delta x_i,$$

és behelyettesítve a másodikba $i = 0, 1, \ldots, n-2$-re kis számolással adódik

$$c_i\Delta x_i + 2c_{i+1}(\Delta x_i + \Delta x_{i+1}) + c_{i+2}\Delta x_{i+1} = 3\frac{\Delta y_{i+1}}{\Delta x_{i+1}} - 3\frac{\Delta y_i}{\Delta x_i}, \quad i = 0, 1, \ldots, n-2. \tag{6.22}$$

Megjegyezzük, hogy a (6.22) egyenlet levezetéséhez nem használtuk a (6.14) feltételt, így ez tetszőleges harmadrendű spline interpolációra teljesül.

A (6.22) egyenlet $n-1$ db, c_i-re nézve lineáris egyenletet ír le. Ehhez hozzávéve a (6.14) feltételből adódó $c_0 = 0$ és $c_n = 0$ egyenleteket $n+1$ egyenletből álló $\mathbf{Ax} = \mathbf{b}$ alakú lineáris egyenletrendszert kapunk, ahol $\mathbf{x} = (c_0, c_1, \ldots, c_n)^T$,

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \Delta x_0 & 2(\Delta x_0 + \Delta x_1) & \Delta x_1 & 0 & 0 & \cdots & 0 \\ 0 & \Delta x_1 & 2(\Delta x_1 + \Delta x_2) & \Delta x_2 & 0 & \cdots & 0 \\ & & \ddots & \ddots & \ddots & & \\ 0 & \cdots & & & \Delta x_{n-2} & 2(\Delta x_{n-2} + \Delta x_{n-1}) & \Delta x_{n-1} \\ 0 & \cdots & & & 0 & 0 & 1 \end{pmatrix}$$

tridiagonális mátrix és

$$\mathbf{b} = \begin{pmatrix} 0 \\ 3\frac{\Delta y_1}{\Delta x_1} - 3\frac{\Delta y_0}{\Delta x_0} \\ \vdots \\ 3\frac{\Delta y_{n-1}}{\Delta x_{n-1}} - 3\frac{\Delta y_{n-2}}{\Delta x_{n-2}} \\ 0 \end{pmatrix}.$$

Mivel $\mathbf{A}$ diagonálisan domináns, az $\mathbf{Ax} = \mathbf{b}$ egyenletnek létezik egyértelmű megoldása. A c_i-k ismeretében pedig a d_i és b_i együtthatókat is meghatározhatjuk. Ezzel beláttuk, hogy a feladatnak létezik egyértelmű megoldása. Megjegyezzük, hogy a gyakorlatban az $\mathbf{Ax} = \mathbf{b}$ egyenletrendszert a tridiagonális lineáris egyenletre vonatkozó Gauss-eliminációval (3.37. algoritmus) oldhatjuk meg hatékonyan. Beláttuk tehát:

6.22. tétel. *A harmadrendű spline interpoláció feladatának létezik pontosan egy természetes harmadrendű spline függvény megoldása.*

6.23. példa. Illesszünk természetes harmadrendű spline függvényt az

x_i	0.0	1.0	1.5	2.0	3.0	4.0
y_i	0.5	0.1	2.5	-1.0	-0.5	0.0

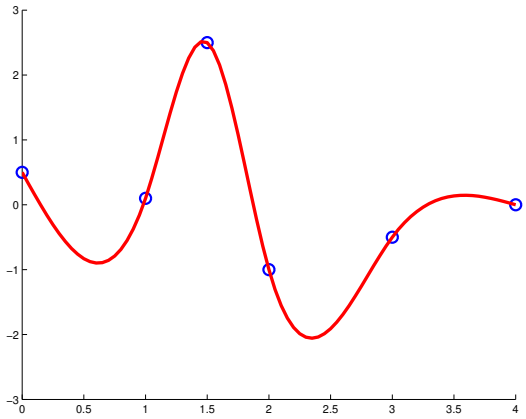
adatokra! Az előző jelölést követve a c_i együtthatókra felírt lineáris egyenletrendszer az adott adatokra a következő lesz:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 3 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 2 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 3 & 1 & 0 \\ 0 & 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 15.6 \\ -35.4 \\ 22.5 \\ 0 \\ 0 \end{pmatrix}.$$

Ezt megoldva kapjuk a c_i értékeket, amit visszahelyettesítve a (6.19) és (6.20) egyenletekbe kiszámíthatók a d_i és b_i együtthatók értékei. A számolást elvégezve a következő harmadrendű polinomokat kapjuk az egyes intervallumokon:

$$\begin{aligned} S_0(x) &= 0.5 - 3.4141079x + 3.0141079x^3, \\ S_1(x) &= 0.1 + 5.6282158(x-1) + 9.04232365(x-1)^2 - 21.3975104(x-1)^3, \\ S_2(x) &= 2.5 - 1.3775934(x-1.5) - 23.0539419(x-1.5)^2 + 23.6182573(x-1.5)^3, \\ S_3(x) &= -1.0 - 6.7178423(x-2) + 12.3734440(x-2)^2 - 5.1556017(x-2)^3, \\ S_4(x) &= -0.5 + 2.5622407(x-3) - 3.0933610(x-3)^2 + 1.0311203(x-3)^3. \end{aligned}$$

A kapott spline függvény és az adatok grafikonja a 6.3. ábrán látható. □



6.3. ábra. Spline interpoláció

A (6.14) feltétel helyett számos más, S végpontjaira vonatkozó feltételt is kiköthetünk. Itt most csak az

$$S'(x_0) = y'_0 \qquad \text{és } S'(x_n) = y'_n \tag{6.23}$$

feltételt vizsgáljuk, ahol y'_0 és y'_n adott számok. Ez azt jelenti, hogy ismerjük az S függvény érintőjét a grafikon végpontjaiban. A (6.23) feltételt teljesítő spline függvényt *teljes spline függvénynek* nevezzük. Ebben az esetben is ugyanúgy kapjuk a (6.22) egyenleteket. Még két egyenletet kell felírni, hogy az egyenletrendszer jól meghatározott legyen. Használva a $b_0 = S'(x_0) = y'_0$ összefüggést, a (6.20) egyenletből következik

$$y'_0\Delta x_0 + c_0(\Delta x_0)^2 + \frac{c_1 - c_0}{3}(\Delta x_0)^2 = \Delta y_0,$$

azaz

$$2c_0\Delta x_0 + c_1\Delta x_0 = 3\frac{\Delta y_0}{\Delta x_0} - 3y'_0. \tag{6.24}$$

b_{n-1}-et kifejezve a (6.20) egyenletből és behelyettesítve a (6.21) egyenletbe, és a $b_n = y'_n$ összefüggést használva kapjuk

$$\frac{\Delta y_{n-1}}{\Delta x_{n-1}} - \frac{2c_{n-1} + c_n}{3}\Delta x_{n-1} + \Delta x_{n-1}(c_{n-1} + c_n) = y'_n,$$

ill. átrendezve

$$c_{n-1}\Delta x_{n-1} + 2c_n\Delta x_{n-1} = 3y'_n - 3\frac{\Delta y_{n-1}}{\Delta x_{n-1}}. \tag{6.25}$$

Ha a természetes spline interpolációnál kapott $\mathbf{Ax} = \mathbf{b}$ egyenlet első egyenletét kicseréljük a (6.24) egyenletre, és az utolsó egyenletet a (6.25) egyenletre, akkor könnyen látható, hogy az együtthatómátrix továbbra is diagonálisan domináns marad, azaz a módosított egyenletrendszernek is van egyértelmű megoldása. Így a (6.23) feltétellel kiegészített interpolációs problémának van teljes spline függvény megoldása, és a megoldás egyértelmű.

A harmadrendű természetes spline interpolációs függvények a következő minimum tulajdonsággal rendelkeznek, ami bizonyos értelemben azt jelenti, hogy spline függvénnyel lehet a legsimábban interpolálni adott pontokat.

6.24. tétel. *Legyen $a = x_0 < x_1 < \ldots < x_n = b$ és $y_0, y_1, \ldots, y_n$ osztópontoknak és hozzátartozó függvényértékeknek egy véges sorozata, és legyen S az ezeket interpoláló természetes kubikus spline függvény. Ekkor*

$$\int_a^b (S''(x))^2\,dx \le \int_a^b (f''(x))^2\,dx \tag{6.26}$$

minden olyan $f \in C^2(a,b)$-re, amely szintén interpolálja az adatokat, azaz $f(x_i) = y_i$ minden $i = 0, 1, \ldots, n$-re.

Bizonyítás. Vezessük be a $g(x) \equiv f(x) - S(x)$ függvényt. Ekkor $f''(x) = S''(x) + g''(x)$, és így

$$\int_a^b (f''(x))^2\,dx = \int_a^b (S''(x))^2\,dx + 2\int_a^b S''(x)g''(x)\,dx + \int_a^b (g''(x))^2\,dx.$$

Mivel $\int_a^b (g''(x))^2\,dx \ge 0$, így a tétel állítása következik ebből az egyenlőségből, ha belátjuk, hogy $\int_a^b S''(x)g''(x)\,dx = 0$. Az integrált felbontva és parciálisan integrálva kapjuk

$$\begin{aligned}
\int_a^b S''(x)g''(x)\,dx &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} S''(x)g''(x)\,dx \\
&= \sum_{i=1}^n \left[S''(x)g'(x)\right]_{x_{i-1}}^{x_i} - \sum_{i=1}^n \int_{x_{i-1}}^{x_i} S'''(x)g'(x)\,dx \\
&= S''(b)g'(b) - S''(a)g'(a) - \sum_{i=1}^n \int_{x_{i-1}}^{x_i} S'''(x)g'(x)\,dx.
\end{aligned}$$

S természetes spline függvény, így $S''(a) = S''(b) = 0$. Mivel S harmadfokú polinom minden $[x_{i-1}, x_i]$ intervallumon, ezért ott S''' konstans függvény, így az integrál elé kivihető. Viszont $\int_{x_{i-1}}^{x_i} g'(x)\,dx = g(x_i) - g(x_{i-1}) = 0$, mivel $g(x_i) = 0$ minden $i = 0, 1, \ldots, n$-re. Ezzel a tételt beláttuk. □

A következő tétel a teljes spline interpoláció hibáját vizsgálja. Bizonyítás nélkül közöljük az eredményt.

6.25. tétel. *Legyen $f \in C^4(a,b)$, $a = x_0 < x_1 < \ldots < x_n = b$ osztópontok, $y_i = f(x_i)$, $i = 0, 1, \ldots, n$ függvényértékek, valamint $y_0' = f'(a)$ és $y_n' = f'(b)$ derivált értékek, és legyen S az ezekhez tartozó teljes spline függvény. Ekkor $x \in [a,b]$-re*

$$\begin{aligned}
|f(x) - S(x)| &\le \frac{5}{384} M_4 h^4, \\
|f'(x) - S'(x)| &\le \left(\frac{\sqrt{3}}{216} + \frac{1}{24}\right) M_4 h^3, \\
|f''(x) - S''(x)| &\le \left(\frac{1}{12} + \frac{h}{3k}\right) M_4 h^2,
\end{aligned}$$

ahol $M_4 \equiv \max\{|f^{(4)}(x)|\colon\ x \in [a,b]\}$, $h \equiv \max\{x_{i+1} - x_i\colon\ i = 0, 1, \ldots, n-1\}$, $k \equiv \min\{x_{i+1} - x_i\colon\ i = 0, 1, \ldots, n-1\}$.

Megjegyezzük, hogy a természetes spline interpoláció hibája ehhez hasonló módon becsülhető.

Feladatok

1. Adja meg az (x_i, y_i), $i = 0, 1, \ldots, n$ adatokat interpoláló lineáris spline függvény képletét az $[x_i, x_{i+1}]$ intervallumon!
2. Adott egy $f : [a, b] \to \mathbb{R}$ folytonos függvény, és legyen S_h az $[a, b]$ intervallum ekvidisztáns, h lépésközű osztópontjaihoz tartozó f-et interpoláló lineáris spline függvény.
   (a) Mutassa meg, hogy $\max\{|f(x) - S_h(x)| \colon x \in [a, b]\} \to 0$, ha $h \to 0$.
   (b) Legyen $f \in C^1[a, b]$. Mutassa meg, hogy
   $$|f(x) - S_h(x)| \le M_1 h, \qquad x \in [a, b],$$
   ahol $M_1 \equiv \max\{|f'(x)| \colon x \in [a, b]\}$.
3. Számítsa ki és ábrázolja a 6.1. szakasz 1. feladatában szereplő adatokhoz tartozó természetes kubikus spline interpolációs függvényeket!
4. Mutassa meg, hogy kvadratikus spline-interpolációnál az
   $$S'(x_0) = f'(x_0) \qquad \text{vagy} \qquad S'(x_n) = f'(x_n)$$
   feltételek egyike teljesülése egyértelműen meghatározza a spline interpolációs függvényt!
5. Mutassa meg, hogy ha S adott $a = x_0 < x_1 < \ldots < x_n = b$ osztópontokhoz és $y_0, y_1, \ldots, y_n$ függvényértékekhez, valamint y'_0 és y'_n derivált értékekhez tartozó teljes spline függvény, akkor S teljesíti a (6.26) egyenlőtlenséget minden olyan $f \in C^2(a, b)$ függvényre, amelyre $f(x_i) = y_i$ minden i-re, $f'(a) = y'_0$ és $f'(b) = y'_n$!

7. fejezet

Numerikus differenciálás és integrálás

Ebben a fejezetben először a numerikus differenciálás különböző képleteit vizsgáljuk, majd a Richardson-extrapolációt definiáljuk, mellyel egy adott rendű numerikus módszer képletéből magasabbrendű formulákat nyerhetünk. Ezután határozott integrálok közelítésének két népszerű módszerét tanulmányozzuk: Newton-Cotes- és Gauss-féle kvadratúra formulák. A Gauss-féle kvadratúra formula levezetése kapcsán az ortogonális polinomok elméletének elemeit is ismertetjük.

7.1. Numerikus differenciálás

Ebben a szakaszban függvények deriváltjait közelítő képletek levezetésének két módszerét és az egyszerűbb közelítő képleteket ismertetjük. A derivált a függvény differenciahányadosának határértéke:

$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h) - f(x_0)}{h}.$$

Így nyilvánvalóan ha $|h|$ kicsi, akkor a differenciahányados, $\frac{f(x_0+h)-f(x_0)}{h}$ közel van a derivált értékéhez. A numerikus analízisben ennél többre van szükség: ismerni szeretnénk a közelítés hibáját. A következőkben kétféleképpen vezetjük le ugyanezt a közelítő képletet, de úgy, hogy közben a közelítés hibáját is megkapjuk.

Tegyük fel, hogy $f \in C^3(a,b)$, és $x_0 \in (a,b)$. Az első megközelítés alapötlete a következő: Helyettesítsük az f függvényt x_0 egy környezetében valamilyen $L_n(x)$ Lagrange-féle közelítő polinommal. Használjuk $L_n'(x_0)$-t az $f'(x_0)$ érték közelítésére! Ezt a módszert *Lagrange-módszernek* nevezzük. Nézzük a legegyszerűbb esetet: Legyen $n = 1$, $x_1 = x_0 + h \in (a,b)$ (és $x_0 \neq x_1$), és tekintsük az f függvény x_0, x_1 osztópontokhoz tartozó elsőfokú Lagrange-polinom közelítését:

$$\begin{aligned} f(x) &= L_1(x) + E_1(x) \\ &= \frac{f(x_0)(x-x_0-h)}{-h} + \frac{f(x_0+h)(x-x_0)}{h} + \frac{f''(\xi(x))}{2}(x-x_0)(x-x_0-h). \end{aligned}$$

Ezt differenciálva kapjuk:

$$\begin{aligned} f'(x) &= \frac{f(x_0+h)-f(x_0)}{h} + \frac{f''(\xi(x))}{2}(2(x-x_0)+h) \\ &\quad + \frac{d}{dx}\Big(f''(\xi(x))\Big)\frac{(x-x_0)(x-x_0-h)}{2}. \end{aligned} \tag{7.1}$$

A 6.8. tétel szerint $f''(\xi(x))$ differenciálható $x \neq x_0, x_0+h$-ra, de a deriváltat nem tudjuk explicit módon kiszámolni. Viszont az $x \to x_0$ határértéket véve a (7.1) képletben kapjuk az

$$f'(x_0) = \frac{f(x_0+h)-f(x_0)}{h} - \frac{h}{2}f''(\xi) \tag{7.2}$$

összefüggést, ahol $\xi \in \langle x_0, x_0 + h\rangle$. Azaz, ha az

$$f'(x_0) \approx \frac{f(x_0 + h) - f(x_0)}{h} \tag{7.3}$$

közelítést használjuk, a közelítés hibája $-\frac{h}{2}f''(\xi)$ alakban írható fel. A (7.3) képletet az f függvény *jobb oldali elsőrendű differenciájának* nevezzük, ha $h > 0$, illetve *bal oldali elsőrendű differenciájának* nevezzük, ha $h < 0$ (mert ekkor az $x_0 + h$ pont az x_0-tól jobbra, ill. balra helyezkedik el). A (7.2) képlet mutatja, hogy a (7.3) közelítés hibája h-ban elsőrendű.

Ugyanezt az eredményt (de egy kicsit enyhébb feltételek mellett) levezethetjük a következőképpen is: Legyen $f \in C^2(a, b)$, és tekintsük az f függvény elsőrendű x_0-körüli Taylor-közelítését:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(\xi(x))}{2}(x - x_0)^2.$$

Behelyettesítve $x = x_0 + h$-t, következik, hogy

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{f''(\xi)}{2}h^2,$$

azaz

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} - \frac{h}{2}f''(\xi),$$

ahol $\xi = \xi(x_0 + h)$.

7.1. példa. Tekintsük az $f(x) = e^{x^2+x}$ függvényt. $f'(x) = e^{x^2+x}(2x + 1)$, így $f'(0) = 1$. Számítsuk ki az $f'(0)$ egy közelítő értékét jobb oldali (pozitív h) és bal oldali (negatív h) elsőrendű differencia képletet ((7.3) képlet) használva! A 7.1. táblázatban feltüntettük a derivált közelítő értékeket és a fellépő hibát különböző h értékekre. A numerikus eredmények igazolják, hogy ha egy nagyságrenddel csökkentjük a lépésközt, akkor a hiba egy nagyságrenddel csökken. □

7.1. táblázat. Elsőrendű differencia képlet, $f(x) = e^{x^2+x}$, $x_0 = 0$

$\lvert h\rvert$	jobb oldali	hiba	bal oldali	hiba
0.100	1.1627807	1.6278e-01	0.8606881	1.3931e-01
0.010	1.0151177	1.5118e-02	0.9851156	1.4884e-02
0.001	1.0015012	1.5012e-03	0.9985012	1.4988e-03

Az előbb említett két módszer magasabbrendű (azaz pontosabb) közelítő képletek levezetésére is használható. Tekintsük az n-edfokú Lagrange-polinom közelítést használó módszert: legyen $f \in C^{n+1}$, és tekintsük az

$$f(x) = \sum_{k=0}^{n} f(x_k)l_k(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n) \tag{7.4}$$

összefüggést, ahol $l_k(x)$ a (6.2) képlettel definiált n-edfokú Lagrange-féle alappolinom. Differenciálva (7.4)-et és az $x = x_i$ helyettesítést alkalmazva kis számolás után kapjuk

$$f'(x_i) = \sum_{j=0}^{n} f(x_j)l_j'(x_i) + \frac{f^{(n+1)}(\xi(x_i))}{(n+1)!}\prod_{\substack{j=0\\ j\neq i}}^{n}(x_i - x_j). \tag{7.5}$$

A (7.5) összefüggést ekvidisztáns alappontokra szokás felírni, azaz feltesszük, hogy $x_j = x_0 + jh$, ahol $h > 0$. A (7.5) képletet $n+1$ alappontot használó *differencia képletnek* nevezzük. Belátható, hogy a (7.5) képletben szereplő hibatag h-ban n-edrendű.

Tekintsük most az $n = 2$ esetet, azaz a három pontra illeszkedő formulákat. Tekintsük az $x_0, x_0 + h, x_0 + 2h$ osztópontokat. Ekkor

$$\begin{aligned} l_0(x) &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-x_1)(x-x_2)}{2h^2}, \\ l_1(x) &= \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-x_0)(x-x_2)}{-h^2}, \\ l_2(x) &= \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-x_0)(x-x_1)}{2h^2}, \end{aligned}$$

ezért

$$\begin{aligned} l_0'(x) &= \frac{2x-x_1-x_2}{2h^2}, \\ l_1'(x) &= \frac{2x-x_0-x_2}{-h^2}, \\ l_2'(x) &= \frac{2x-x_0-x_1}{2h^2}. \end{aligned}$$

Ezt alkalmazva $x = x_0$, $x = x_0 + h$ ill. $x = x_0 + 2h$-ra, a (7.5) képletből kapjuk, hogy

$$f'(x_0) = \frac{1}{h}\left(-\frac{3}{2}f(x_0) + 2f(x_0+h) - \frac{1}{2}f(x_0+2h)\right) + \frac{h^2}{3}f'''(\xi_0), \tag{7.6}$$

$$f'(x_0+h) = \frac{1}{h}\left(-\frac{1}{2}f(x_0) + \frac{1}{2}f(x_0+2h)\right) - \frac{h^2}{6}f'''(\xi_1), \tag{7.7}$$

$$f'(x_0+2h) = \frac{1}{h}\left(\frac{1}{2}f(x_0) - 2f(x_0+h) + \frac{3}{2}f(x_0+2h)\right) + \frac{h^2}{3}f'''(\xi_2). \tag{7.8}$$

Az $x_0 \leftarrow x_0 - 2h$ és $h \leftarrow -h$ helyettesítéssel a (7.8) a (7.6) alakban írható fel, (7.7) pedig az $x_0 \leftarrow x_0 - h$ és $h \leftarrow -h$ helyettesítéssel

$$f'(x_0) = \frac{1}{h}\left(-\frac{1}{2}f(x_0-h) + \frac{1}{2}f(x_0+h)\right) - \frac{h^2}{6}f'''(\xi_1) \tag{7.9}$$

alakú lesz. A (7.9) képlet egy *centrális másodrendű differencia képlet*, (7.6) pedig *jobb oldali* ill. *bal oldali másodrendű differencia*, attól függően, hogy h pozitív vagy negatív.

7.2. példa. Az $f(x) = e^{x^2+x}$ függvény $x = 0$ pontjában vett deriváltját közelítettük jobb oldali, bal oldali és centrális másodrendű differencia képletekkel ((7.6) és (7.9) képletek). Az eredményeket a 7.2. táblázatban adtuk meg különböző h-ra, amelyekből látható, hogy a képletek másodrendű hibával rendelkeznek. □

7.2. táblázat. Másodrendű differencia képlet, $f(x) = e^{x^2+x}$, $x_0 = 0$

h	jobb oldali	hiba	bal oldali	hiba	centrális	hiba
0.100	0.9693157	3.0684e-02	0.9820952	1.7905e-02	1.0117344	1.1734e-02
0.010	0.9997603	2.3968e-04	0.9997728	2.2718e-04	1.0001167	1.1667e-04
0.001	0.9999977	2.3396e-06	0.9999977	2.3271e-06	1.0000012	1.1667e-06

Bizonyítás nélkül közöljük az 5 pontra felírt egyoldali és centrális negyedrendű képleteket:

$$f'(x_0) = \frac{1}{12h}\Big(-25f(x_0)+48f(x_0+h)-36f(x_0+2h)+16f(x_0+3h) - 3f(x_0+4h)\Big)+\frac{h^4}{5}f^{(5)}(\xi_0), \tag{7.10}$$

$$f'(x_0) = \frac{1}{12h}\Big(f(x_0-2h)-8f(x_0-h)+8f(x_0+h)-f(x_0+2h)\Big) + \frac{h^4}{30}f^{(5)}(\xi_1). \tag{7.11}$$

A (7.10) egyoldali, (7.11) pedig centrális differencia képlet.

7.3. példa. Alkalmazzuk a (7.10) és (7.11) képleteket az $f(x) = e^{x^2+x}$ függvény deriváltjának közelítésére $x = 0$-ban! A 7.3. táblázatban láthatók a numerikus eredmények. □

7.3. táblázat. Negyedrendű differencia képlet, $f(x) = e^{x^2+x}$, $x_0 = 0$

h	jobb oldali	hiba	bal oldali	hiba	centrális	hiba
0.100	0.9967110	3.2890e-03	0.9991793	8.2070e-04	0.9997248	2.7523e-04
0.010	0.9999998	1.7345e-07	0.9999998	1.5136e-07	1.0000000	2.7005e-08
0.001	1.0000000	1.6311e-11	1.0000000	1.6090e-11	1.0000000	2.7000e-12

Magasabbrendű deriváltak közelítésére a Lagrange-módszernél kényelmesebben használható a *Taylor-módszer*. Legyen $f \in C^4$, és tekintsük az f függvény x_0 körüli harmadrendű Taylor-képletét:

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \frac{f'''(x_0)}{6}(x-x_0)^3 + \frac{f^{(4)}(\xi)}{24}(x-x_0)^4.$$

Ha ebbe $x = x_0 - h$-t és $x = x_0 + h$-t helyettesítünk, akkor az

$$f(x_0-h) = f(x_0) - f'(x_0)h + \frac{f''(x_0)}{2}h^2 - \frac{f'''(x_0)}{6}h^3 + \frac{f^{(4)}(\xi_1)}{24}h^4$$

és

$$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2}h^2 + \frac{f'''(x_0)}{6}h^3 + \frac{f^{(4)}(\xi_2)}{24}h^4$$

összefüggéseket kapjuk. Ezt a két egyenletet összeadva

$$f(x_0-h)+f(x_0+h) = 2f(x_0) + f''(x_0)h^2 + \frac{f^{(4)}(\xi_1)+f^{(4)}(\xi_2)}{24}h^4$$

adódik, amiből

$$f''(x_0) = \frac{f(x_0-h)-2f(x_0)+f(x_0+h)}{h^2} + \frac{f^{(4)}(\xi_1)+f^{(4)}(\xi_2)}{24}h^4.$$

Ebből látszik, hogy az

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

közelítő képlet h^2 nagyságrendű hibával rendelkezik. Az $\frac{f^{(4)}(\xi_1)+f^{(4)}(\xi_2)}{24}h^4$ hibatagot egyszerűbb alakra hozhatjuk. A feltételek szerint $f^{(4)}$ folytonos, ezért a 2.2. tétel szerint valamely ξ_1 és ξ_2 közötti ξ pontban

$$f^{(4)}(\xi) = \frac{f^{(4)}(\xi_1) + f^{(4)}(\xi_2)}{2}.$$

Ezért

$$f''(x_0) = \frac{f(x_0 - h) - 2f(x_0) + f(x_0 + h)}{h^2} + \frac{f^{(4)}(\xi)}{12}h^2. \tag{7.12}$$

7.4. példa. Számítsuk ki az $f(x) = e^{x^2+x}$ függvény második deriváltjának közelítő értékét $x = 0$-ban! A 7.4. táblázatban láthatók a numerikus eredmények. □

7.4. táblázat. Másodrendű derivált közelítése, $f(x) = e^{x^2+x}$, $x_0 = 0$

h	közelítés	hiba
0.100	3.0209256	2.0926e-02
0.010	3.0002083	2.0834e-04
0.001	3.0000021	2.0833e-06

A numerikus differenciálás egy instabil feladat. Ennek igazolására tekintsünk egy $f(x)$ függvényt és annak egy

$$g(x) = f(x) + \frac{1}{n}\sin(n^2x)$$

perturbációját. Ha f helyett a g függvény numerikus deriváltját számoljuk ki, akkor a differencia képletekben használt függvényértékek nagy n esetén csak kicsit változnak, a derivált értéke viszont jelentősen megváltozik, hiszen $g'(x) = f'(x) + n\cos(n^2x)$.

Vizsgáljuk most a kerekítési hiba hatását a numerikus differenciálási képletekre. Tekintsük pl. a legegyszerűbb numerikus differenciálási képletet, a (7.2) formulát. Ebben $f(x_0)$ és $f(x_0+h)$ pontos értékei helyett f_0 ill. f_1 közelítő értékekkel számolunk, ahol

$$f(x_0) = f_0 + e_0 \qquad \text{és} \qquad f(x_0 + h) = f_1 + e_1.$$

Ekkor

$$f'(x_0) \approx \frac{f_1 - f_0}{h},$$

és az elkövetett hiba

$$\begin{aligned} f'(x_0) - \frac{f_1 - f_0}{h} &= f'(x_0) - \frac{f(x_0 + h) - f(x_0)}{h} + \frac{f(x_0 + h) - f(x_0)}{h} - \frac{f_1 - f_0}{h} \\ &= -\frac{h}{2}f''(\xi) + \frac{e_1 - e_0}{h}. \end{aligned} \tag{7.13}$$

A (7.13) összefüggésből látszik, hogy a tényleges hiba két részből adódik. Az egyik a képlethiba, a másik pedig a kerekítési hiba. Ha a lépésköz kicsi, akkor a képlethiba kicsi lesz, viszont a kerekítési hiba tart a végtelenbe, ha $h \to 0$.

7.5. példa. Tekintsük az $f(x) = e^x$ függvényt. Számítsuk ki $f'(1)$ közelítését elsőrendű jobb oldali differencia képlettel. Hogy a kerekítési hibák hatását vizsgáljuk, a számításokat 6- illetve 4-jegyű aritmetikát használva végeztük el. A 7.5. táblázatból látható, hogy 4-jegyű aritmetika használata esetén a lépéshossz

7.5. táblázat. Kerekítési hibák hatása, $f(x) = e^x$, $x_0 = 1$

h	6-jegyű aritmetikával `differencia`	`hiba`	4-jegyű aritmetikával `differencia`	`hiba`
0.100	2.8589000	1.4062e-01	2.8600000	1.4172e-01
0.010	2.7320000	1.3718e-02	2.8000000	8.1718e-02
0.001	2.7200000	1.7182e-03	3.0000000	2.8172e-01

0.01-ről 0.001-re csökkentésekor az elkövetett hiba növekszik. □

Az itt megismert módszereket könnyen átfogalmazhatjuk többváltozós függvények parciális deriváltjai közelítésére. A következő egyoldali ill. centrális közelítő képletek levezetését az olvasóra hagyjuk.

$$\frac{\partial f(x_0,y_0)}{\partial x} \approx \frac{f(x_0+h,y_0)-f(x_0,y_0)}{h}, \tag{7.14}$$

$$\frac{\partial f(x_0,y_0)}{\partial x} \approx \frac{f(x_0,y_0+h)-f(x_0,y_0)}{h}, \tag{7.15}$$

$$\frac{\partial^2 f(x_0,y_0)}{\partial x^2} \approx \frac{f(x_0+h,y_0)-2f(x_0,y_0)+f(x_0-h,y_0)}{h^2} \tag{7.16}$$

$$\frac{\partial^2 f(x_0,y_0)}{\partial y^2} \approx \frac{f(x_0,y_0+h)-2f(x_0,y_0)+f(x_0,y_0-h)}{h^2} \tag{7.17}$$

$$\frac{\partial^2 f(x_0,y_0)}{\partial x\,\partial y} \approx \frac{f(x_0+h,y_0+h)-f(x_0+h,y_0)-f(x_0,y_0+h)+f(x_0,y_0)}{h^2} \tag{7.18}$$

$$\frac{\partial^2 f(x_0,y_0)}{\partial x^2} \approx \frac{f(x_0+2h,y_0)-2f(x_0+h,y_0)+f(x_0,y_0)}{h^2} \tag{7.19}$$

Feladatok

1. Számítsa ki $f'(x_0)$ közelítő értékét elsőrendű jobb és bal oldali differencia képletek segítségével a $h = 0.1$ és 0.01 lépésközt használva, ha

   (a) $f(x) = x^4 - 6x^2 + 3x$, $x_0 = 1$, (b) $f(x) = e^x \sin x$, $x_0 = 0$,

   (c) $f(x) = \cos x^2$, $x_0 = 1$, (d) $f(x) = x \ln x$, $x_0 = 1$.

2. Ismételje meg az előző feladatot másodrendű differencia képleteket használva!
3. Számítsa ki $f''(x_0)$ közelítő értékét az 1. feladatban felsorolt függvényekre!
4. Vezesse le a (7.6) és (7.9) közelítő képleteket Taylor-módszerrel!
5. Vezesse le a (7.10) és (7.11) közelítő képleteket!
6. Vezesse le a következő közelítő képleteket:

$$f'''(x_0) \approx \frac{1}{2h^3}\Big(f(x_0+2h)-2f(x_0+h)+2f(x_0-h)-f(x_0-2h)\Big),$$

$$f^{(4)}(x_0) \approx \frac{1}{h^4}\Big(f(x_0+2h)-4f(x_0+h)+6f(x_0)-4f(x_0-h)+f(x_0+2h)\Big)$$

7. Vezesse le a (7.14)–(7.19) közelítéseket
   (a) egyváltozós függvényekre vonatkozó közelítő deriválási képletek,
   (b) kétváltozós Lagrange-módszer,

(c) kétváltozós Taylor-módszer

segítségével! Határozza meg a képlethiba rendjét!

7.2. Richardson-extrapoláció

Tegyük fel, hogy adott egy M mennyiség, amelynek ismerjük egy $K(h)$ közelítését, ahol h a közelítő módszer paramétere (lépésköze), és ismerjük a közelítés képlethibáját is. Feltesszük, hogy a hiba speciális alakú, h-szerint páros hatványú véges (vagy végtelen) hatványsorba fejthető:

$$M = K(h) + a_2h^2 + a_4h^4 + a_6h^6 + \cdots + a_{2m}h^{2m} + b(h), \tag{7.20}$$

ahol $|b(h)| \leq Bh^{2m+2}$ valamely $B > 0$ konstanssal. Ez a hiba h-ban másodrendű. Most egy általános módszert ismertetünk, amelynek segítségével magasabbrendű hibával rendelkező közelítő képletet nyerhetünk a $K(h)$ képletből kiindulva. Írjuk fel $h/2$-re az előző közelítő képletet és a hozzá tartozó hibát:

$$M = K(h/2) + a_2\frac{h^2}{4} + a_4\frac{h^4}{16} + a_6\frac{h^6}{64} + \cdots + a_{2m}\frac{h^{2m}}{2^{2m}} + b(h/2). \tag{7.21}$$

A (7.21) egyenlet 4-szereséből kivonva a (7.20) egyenletet a h-ban másodrendű hibatag kiesik. A kapott egyenletből M-et kifejezhetjük:

$$\begin{aligned} M &= \frac{4K(h/2) - K(h)}{3} - \frac{1}{4}a_4h^4 - \frac{5}{16}a_6h^6 \\ &\quad - \cdots - \frac{2^{2m-2}-1}{2^{2m-2}\cdot 3}a_{2m}h^{2m} + \frac{4b(h/2) - b(h)}{3}. \end{aligned} \tag{7.22}$$

Ezt az összefüggést felírhatjuk a következő alakban:

$$M = K^{(1)}(h) + a_4^{(1)}h^4 + a_6^{(1)}h^6 + \cdots + a_{2m}^{(1)}h^{2m} + b^{(1)}(h), \tag{7.23}$$

ahol

$$K^{(1)} \equiv \frac{4K(h/2) - K(h)}{3}, \quad b^{(1)}(h) \equiv \frac{4b(h/2) - b(h)}{3}, \quad a_{2i}^{(1)} \equiv \frac{1 - 4^{i-1}}{4^{i-1}\cdot 3}a_{2i},$$

$i = 2, \ldots, m$. A (7.23) egyenlet azt mutatja, hogy ha a $K^{(1)}(h)$ képletet M közelítésének tekintjük, akkor a közelítés hibája h-ban már negyedrendű. Az előbbi ötletet újra alkalmazhatjuk: A (7.23) egyenletbe $h/2$-t helyettesítünk, majd a kapott egyenlet 16-szorosából kivonjuk a (7.23) egyenletet, és a kapott egyenletet megoldjuk M-re. Ekkor a h^4 tagok kiesnek, és az

$$M = K^{(2)}(h) + a_6^{(2)}h^6 + \cdots + a_{2m}^{(2)}h^{2m} + b^{(2)}(h), \tag{7.24}$$

egyenletet kapjuk, ahol

$$\begin{aligned} K^{(2)} &\equiv \frac{16K^{(1)}(h/2) - K^{(1)}(h)}{15}, \quad b^{(2)}(h) \equiv \frac{16b^{(1)}(h/2) - b^{(1)}(h)}{15}, \\ a_{2i}^{(2)} &\equiv \frac{1 - 4^{i-2}}{4^{i-2}\cdot 15}a_{2i}^{(1)}, \quad i = 3, \ldots, m. \end{aligned}$$

A (7.24) képlet szerint $K^{(2)}(h)$ hatodrendű hibával közelíti M-et. Az eljárást folytatva definiálhatjuk a

$$K^{(i+1)} \equiv K^{(i)}(h/2) + \frac{K^{(i)}(h/2) - K^{(i)}(h)}{4^{i+1} - 1}, \qquad i = 0, 1, \ldots, m-1, \tag{7.25}$$

közelítő képleteket, ahol $K^{(0)}(h) \equiv K(h)$. Az ebben a szakaszban leírt módszert egy közelítő képlet pontosságának növelésére *Richardson-extrapolációnak* nevezzük. A módszer természetesen akkor is alkalmazható, ha a hiba h-nak nem csak páros hatványait tartalmazza (lásd a 2. és 3. feladatokat), de a későbbiekben az itt bemutatott esetre lesz majd szükségünk.

7.6. példa. Az előző szakaszban láttuk, hogy a (7.9) centrális differencia másodrendben közelíti a függvény első deriváltját. A Taylor-módszert alkalmazva megkaphatjuk a képlethiba pontosabb alakját. Tegyük fel, hogy $f \in C^{2m+3}$, és induljunk ki a következő Taylor-képletből:

$$f(x_0+h) = f(x_0) + f'(x_0)h + \cdots + \frac{f^{(2m+2)}(x_0)}{(2m+2)!}h^{2m+2} + \frac{f^{(2m+3)}(\xi_1)}{(2m+3)!}h^{2m+3}.$$

Az előző képletet h helyett $-h$-ra felírva és a két egyenletet kivonva, majd $f'(x_0)$-at kifejezve kapjuk:

$$\begin{aligned} f'(x_0) &= \frac{f(x_0+h)-f(x_0-h)}{2h} - \frac{f'''(x_0)}{3!}h^2 - \frac{f^{(5)}(x_0)}{5!}h^4 \\ &\quad - \cdots - \frac{f^{(2m+1)}(x_0)}{(2m+1)!}h^{2m} - \frac{f^{(2m+3)}(\xi_1)+f^{(2m+3)}(\xi_2)}{(2m+3)!}h^{2m+2}, \end{aligned}$$

azaz a centrális differencia képlete teljesíti a (7.20) összefüggést. Magasabbrendű képletet kaphatunk tehát a centrális differencia képletből kiindulva a Richardson-extrapolációval. Negyedrendű közelítő képlet ad például a

$$\begin{aligned} K^{(1)}(h) &= \frac{4\dfrac{f(x_0+h/2)-f(x_0+h/2)}{h} - \dfrac{f(x_0+h)-f(x_0-h)}{2h}}{3} \\ &= \frac{f(x_0-h)-8f(x_0-h/2)+8f(x_0+h/2)-f(x_0+h)}{6h} \end{aligned}$$

formula. Vegyük észre, hogy a kapott képlet lényegében megegyezik a (7.11) formulával. □

Feladatok

1. Vezessen le egy hatodrendű képletet első derivált közelítésére a centrális differencia képletből kiindulva Richardson-extrapolációval! Alkalmazza a képletet az $f(x) = e^{x^2+x}$ függvény deriváltjának közelítésére $x = 0$-ban a $h = 0.25$ lépésközt alkalmazva!
2. Fogalmazza meg a Richardson-extrapolációt arra az esetre, ha a közelítés képlethibája h minden hatványát tartalmazhatja, azaz
$$M = K(h) + a_1h + a_2h^2 + \cdots + a_mh^m + b(x)$$
alakú, ahol $|b(h)| \le Bh^{m+1}$ valamely $B > 0$-ra!
3. Fogalmazza meg a Richardson-extrapolációt arra az általános esetre, amikor
$$M = K(h) + a_1h^{\alpha_1} + a_2h^{\alpha_2} + \cdots + a_mh^{\alpha_m} + b(x)$$
alakú, ahol $1 \le \alpha_1 < \alpha_2 < \cdots < \alpha_m$ egész számok és $|b(h)| \le Bh^{\alpha_m+1}$ valamely $B > 0$-ra!
4. Készítsen harmadrendű képletet első derivált közelítésére Richardson-extrapolációval az egyoldali differencia formulából kiindulva!

7.3. Newton–Cotes-formulák

Legyen $f \in C(a,b)$. A határozott integrált is, a deriválthoz hasonlóan, határérték segítségével definiáljuk. Riemann-összeg segítségével ez a következő alakban adható meg: vegyük az $[a,b]$

intervallum egy $a = x_0 < x_1 < \cdots < x_n = b$ beosztását, és minden $[x_{i-1}, x_i]$ részintervallumból válasszunk ki egy ξ_i pontot. Ekkor az $\int_a^b f(x)\,dx$ integrál a $\sum_{i=1}^n f(\xi_i)(x_i - x_{i-1})$ alakú Riemann-féle közelítő összeg határértéke, ha a beosztás normája, azaz $\max\{x_i - x_{i-1} : i = 1, \ldots, n\}$ nullához tart. Egy ilyen Riemann-összeg például

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}\left(f\left(\frac{x_0+x_1}{2}\right) + f\left(\frac{x_1+x_2}{2}\right) + \cdots + f\left(\frac{x_{n-1}+x_n}{2}\right)\right), \tag{7.26}$$

ahol $x_i = a + i(b-a)/n$, $i = 0, 1, \ldots, n$. Ezt a közelítő képletet *érintőformulának* nevezzük. (Az érintőformulával kapcsolatban lásd az 5. és 6. feladatokat!)

A numerikus differenciáláshoz hasonlóan integrál közelítő képletek levezetésére is alkalmazhatjuk a Lagrange-módszert: Az $[a, b]$ intervallumon vegyünk (többnyire ekvidisztáns) osztópontokat és legyen L_n a választott alappontokhoz és az f függvényhez tartozó interpolációs polinom. Tekintsük $\int_a^b L_n(x)\,dx$-et mint $\int_a^b f(x)\,dx$ közelítését. Feltéve, hogy $f \in C^{n+1}(a, b)$, a közelítés hibáját megkapjuk a 6.5. tétel felhasználásával:

$$\begin{aligned}\int_a^b f(x)\,dx &= \sum_{k=0}^n f(x_k)\int_a^b l_k(x)\,dx \\ &\quad + \int_a^b \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)(x-x_1)\cdots(x-x_n)\,dx,\end{aligned} \tag{7.27}$$

ahol $l_k(x)$ a (6.2) egyenlettel definiált (az alappontoktól függő) n-edfokú polinom. Ezzel egy

$$\int_a^b f(x)\,dx \approx \sum_{k=0}^n c_k f(x_k) \tag{7.28}$$

alakú integrál közelítő képletet kaptunk, ahol a c_k súlyokat a

$$c_k = \int_a^b l_k(x)\,dx \tag{7.29}$$

integrálok adják. A (7.28) alakú közelítő képleteket *kvadratúra képleteknek* nevezzük, azokat a kvadratúra képleteket pedig, ahol a c_k súlyokat a (7.29) integrálok adják, *Newton–Cotes-formuláknak* hívjuk. Ha az alappontokhoz az a és b pontok is hozzá tartoznak, akkor a (7.28)–(7.29) képletet *zárt Newton–Cotes-formuláknak*, ha az összes alappont az (a, b) nyílt intervallumból van, akkor *nyílt Newton–Cotes-formuláknak* nevezzük. Egy kvadratúra formula *pontossági foka* n, ha a képlet az integrál pontos értékét adja vissza minden legfeljebb n-edfokú polinomra, de van olyan $n+1$-edfokú polinom, amelyre a képlet nem egyezik meg az integrál pontos értékével. Az $n+1$ pontra felírt Newton–Cotes-formulák pontossági rendje tehát legalább n, hiszen az n-edfokú polinomot interpoláló Lagrange-polinom hibája 0. Megmutatható azonban, hogy páros n-re a Newton–Cotes-formula $(n+1)$-edrendű polinomokra is pontos értéket ad vissza.

Vizsgáljuk meg $n = 1$-re a zárt Newton–Cotes-képletet. Legyen $x_0 = a$, $x_1 = b$, $h = b - a$. Ekkor

$$L_1(x) = f(x_0)\frac{x-x_1}{x_0-x_1} + f(x_1)\frac{x-x_0}{x_1-x_0},$$

így

$$\begin{aligned}\int_{x_0}^{x_1} L_1(x)\,dx &= f(x_0)\int_{x_0}^{x_1}\frac{x-x_1}{x_0-x_1}dx + f(x_1)\int_{x_0}^{x_1}\frac{x-x_0}{x_1-x_0}dx \\ &= \left[f(x_0)\frac{(x-x_1)^2}{2(x_0-x_1)} + f(x_1)\frac{(x-x_0)^2}{2(x_1-x_0)}\right]_{x_0}^{x_1} \\ &= \frac{h}{2}(f(x_0)+f(x_1)).\end{aligned}$$

Ennek a formulának a hibáját (7.27) szerint az

$$\int_{x_0}^{x_1} f(x)\,dx - \frac{h}{2}(f(x_0)+f(x_1)) = \int_{x_0}^{x_1} \frac{f''(\xi(x))}{2}(x-x_0)(x-x_1)\,dx$$

képlet adja. A hibatag átalakításához használjuk, hogy $(x-x_0)(x-x_1)<0$, ha $x\in(x_0,x_1)$, ezért alkalmazható a 2.6. tétel. Létezik tehát olyan $\eta\in(x_0,x_1)$ konstans, hogy

$$\int_{x_0}^{x_1} \frac{f''(\xi(x))}{2}(x-x_0)(x-x_1)\,dx = \frac{f''(\eta)}{2}\int_{x_0}^{x_1}(x-x_0)(x-x_1)\,dx,$$

tehát

$$\begin{aligned}\int_{x_0}^{x_1} f(x)\,dx - \frac{h}{2}(f(x_0)+f(x_1)) &= \frac{f''(\eta)}{2}\int_{x_0}^{x_1}(x-x_0)^2 - h(x-x_0)\,dx\\ &= \frac{f''(\eta)}{2}\left[\frac{(x-x_0)^3}{3} - h\frac{(x-x_0)^2}{2}\right]_{x_0}^{x_1}\\ &= -\frac{h^3}{12}f''(\eta).\end{aligned}$$

Kaptuk tehát az ún. *elemi trapézformulát:*

$$\int_a^b f(x)\,dx = \frac{h}{2}(f(a)+f(b)) - \frac{h^3}{12}f''(\xi), \qquad \xi\in(a,b). \tag{7.30}$$

A képlet a nevét a geometriai jelentéséből kapta: a $\frac{h}{2}(f(a)+f(b))$ kifejezés az f függvény grafikonjának a és b x-koordinátájú pontjához tartozó szelő alatti területet, azaz a trapéz területét adja vissza.

Az elemi trapéz formula akkor alkalmazható sikeresen, ha az intervallum hossza kicsi. Ha az intervallum hossza nem kicsi, akkor osszuk fel az $[a,b]$ intervallumot n egyenlő hosszú részintervallumra az x_i $(i=0,1,\dots,n)$ osztópontokkal, ahol $x_i=a+ih$, $h=(b-a)/n$, és minden részintervallumra alkalmazzuk az elemi trapézformulát:

$$\begin{aligned}\int_a^b f(x)\,dx &= \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} f(x)\,dx\\ &= \sum_{i=1}^{n}\frac{h}{2}(f(x_{i-1})+f(x_i)) - \frac{h^3}{12}\sum_{i=1}^{n} f''(\xi_i)\\ &= \frac{h}{2}\left(f(x_0)+2\sum_{i=1}^{n-1}f(x_i)+f(x_n)\right) - \frac{nh^3}{12}\frac{1}{n}\sum_{i=1}^{n} f''(\xi_i).\end{aligned}$$

Feltéve, hogy $f\in C^2(a,b)$, a 2.2. tétel szerint az $\frac{1}{n}\sum_{i=1}^{n} f''(\xi_i)$ átlagérték helyettesíthető egy $f''(\xi)$ alakú függvényértékkel. Ezért, használva még a $hn=b-a$ összefüggést,

$$\int_a^b f(x)\,dx = \frac{h}{2}\left(f(x_0)+2\sum_{i=1}^{n-1}f(x_i)+f(x_n)\right) - \frac{(b-a)h^2}{12}f''(\xi), \qquad \xi\in(a,b). \tag{7.31}$$

Ezt a képletet *összetett trapézformulának* nevezzük.

7.7. példa. Számítsuk ki az $\int_0^1 x^2e^x\,dx$ integrál közelítő értékét a trapézformulával $h = 1$ (elemi trapézformula), $h = 0.5$ és $h = 0.25$ lépésközt használva! Könnyen ellenőrizhető, hogy a pontos integrál $\int_0^1 x^2e^x\,dx = e - 2 = 0.7182818$. Az első esetben

$$\int_0^1 x^2e^x\,dx \approx \frac{1}{2}(0 + e) = 1.3591409.$$

A hiba ekkor 0.6408591. Ha $h = 0.5$-re alkalmazzuk az összetett trapézformulát, akkor

$$\int_0^1 x^2e^x\,dx \approx \frac{0.5}{2}(0 + 0.5^2e^{0.5} + e) = 0.8856606.$$

Ennek hibája 0.1673788. Végül $h = 0.25$-re

$$\int_0^1 x^2e^x\,dx \approx \frac{0.25}{2}(0 + 0.25^2e^{0.25} + 0.5^2e^{0.5} + 0.75^2e^{0.75} + e) = 0.7605963,$$

aminek a hibája 0.0423145. Látható, hogy felezve a lépésközt a hiba körülbelül a negyedrészére csökken, azaz a hiba h-ban másodrendű. □

Számítsuk most ki a (7.27) képletet $n = 2$-re, ekvidisztáns osztópontokat használva, azaz $x_0 = a$, $x_1 = x_0 + h$, $x_2 = b$, $h = (b - a)/2$.

$$\begin{aligned}
&\int_{x_0}^{x_2} L_2(x)\,dx \\
&= f(x_0)\int_{x_0}^{x_2}\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}dx + f(x_1)\int_{x_0}^{x_2}\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}dx \\
&\qquad + f(x_2)\int_{x_0}^{x_2}\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}dx \\
&= \frac{f(x_0)}{2h^2}\int_{x_0}^{x_2}(x-x_2+h)(x-x_2)\,dx - \frac{f(x_1)}{h^2}\int_{x_0}^{x_2}(x-x_0)(x-x_0-2h)\,dx \\
&\qquad + \frac{f(x_2)}{2h^2}\int_{x_0}^{x_2}(x-x_0)(x-x_0-h)\,dx \\
&= \frac{f(x_0)}{2h^2}\left[\frac{(x-x_2)^3}{3} + h\frac{(x-x_2)^2}{2}\right]_{x_0}^{x_2} - \frac{f(x_1)}{h^2}\left[\frac{(x-x_0)^3}{3} - 2h\frac{(x-x_0)^2}{2}\right]_{x_0}^{x_2} \\
&\qquad + \frac{f(x_2)}{2h^2}\left[\frac{(x-x_0)^3}{3} - h\frac{(x-x_0)^2}{2}\right]_{x_0}^{x_2} \\
&= \frac{h}{3}(f(x_0) + 4f(x_1) + f(x_2)).
\end{aligned}$$

A közelítés képlethibája

$$\int_{x_0}^{x_2}\frac{f'''(\xi(x))}{6}(x-x_0)(x-x_1)(x-x_2)\,dx.$$

A különbség az előző esethez képest az, hogy most az $(x-x_0)(x-x_1)(x-x_2)$ szorzat különböző előjelű az (x_0, x_1) és az (x_1, x_2) intervallumokon, tehát nem alkalmazható a 2.6. tétel az (x_0, x_2)

intervallumon. Másképp járunk tehát el. Legyen

$$\begin{aligned} p(x) &\equiv \int_{x_0}^{x} (t-x_0)(t-x_1)(t-x_2)\,dt \\ &= \int_{x_0}^{x} (t-x_1+h)(t-x_1)(t-x_1-h)\,dt \\ &= \left[\frac{(t-x_1)^4}{4} - h^2\frac{(t-x_1)^2}{2}\right]_{x_0}^{x} \\ &= \frac{(x-x_1)^4}{4} - \frac{h^2(x-x_1)^2}{2} + \frac{h^4}{4} \\ &= \frac{1}{4}((x-x_1)^2-h^2)^2. \end{aligned}$$

Ekkor $p(x_0) = p(x_2) = 0$, így parciális integrálással

$$\int_{x_0}^{x_2} \frac{f'''(\xi(x))}{6}(x-x_0)(x-x_1)(x-x_2)\,dx = -\int_{x_0}^{x_2} \frac{d}{dx}\frac{f'''(\xi(x))}{6}p(x)\,dx.$$

p nemnegatív függvény, ezért a 2.6. és a 6.8. tételeket alkalmazva kapjuk, hogy

$$\int_{x_0}^{x_2} \frac{f'''(\xi(x))}{6}(x-x_0)(x-x_1)(x-x_2)\,dx = -\frac{f^{(4)}(\eta)}{24}\int_{x_0}^{x_2} p(x)\,dx = -\frac{h^5}{90}f^{(4)}(\eta).$$

Beláttuk tehát az

$$\int_{x_0}^{x_2} f(x)\,dx = \frac{h}{3}(f(x_0)+4f(x_1)+f(x_2)) - \frac{h^5}{90}f^{(4)}(\eta), \qquad \eta \in (x_0,x_2) \tag{7.32}$$

képletet, az ún. *elemi Simpson-formulát.*

A hibatag képlete mutatja, hogy a Simpson-formula meglepő módon harmadrendű polinomokra is az integrál pontos értékét adja vissza, mivel ekkor $f^{(4)}$ azonosan nulla. Másrészt a várt negyedrendű hiba helyett a képlet eggyel jobb, ötödrendű hibával rendelkezik. Ez a jobb hibarend megmutatható minden páros n-re felírt Newton–Cotes-képletnél.

Az összetett trapézformulához hasonlóan vezethető le az *összetett Simpson-formula*: Páros sok egyenlő részre, $2n$ részre osztjuk az $[a,b]$ intervallumot, azaz $h = (b-a)/2n$. Ekkor

$$\begin{aligned} \int_a^b f(x)\,dx &= \frac{h}{3}\left(f(x_0) + 4\sum_{i=1}^{n} f(x_{2i-1}) + 2\sum_{i=1}^{n-1} f(x_{2i}) + f(x_{2n})\right) \\ &\quad - \frac{(b-a)h^4}{180}f^{(4)}(\xi), \qquad \xi \in (a,b). \end{aligned} \tag{7.33}$$

7.8. példa. Számítsuk ki az $\int_0^1 x^2e^x\,dx$ integrál közelítő értékét a Simpson-formulával $h = 0.5$ (elemi Simpson-formula), $h = 0.25$ és $h = 0.125$ lépésközt használva! Az első esetben

$$\int_0^1 x^2e^x\,dx \approx \frac{0.5}{3}(0 + 4\cdot 0.5^2e^{0.5} + e) = 0.7278339.$$

A hiba ekkor 0.0095520. Ha $h = 0.25$-re alkalmazzuk az összetett Simpson-formulát, akkor

$$\int_0^1 x^2e^x\,dx \approx \frac{0.25}{3}(0 + 4\cdot 0.25^2e^{0.25} + 2\cdot 0.5^2e^{0.5} + 4\cdot 0.75^2e^{0.75} + e) = 0.7189082.$$

Ennek hibája 0.0006264. Végül $h = 0.125$-re

$$\int_0^1 x^2 e^x\,dx \approx \frac{0.125}{3}\Big(0 + 4\cdot 0.125^2 e^{0.125} + 2\cdot 0.25^2 e^{0.25} + 4\cdot 0.375^2 e^{0.375} + 2\cdot 0.5^2 e^{0.5} + 4\cdot 0.625^2 e^{0.625} + 2\cdot 0.75^2 e^{0.75} + 4\cdot 0.875^2 e^{0.875} + e\Big) = 0.7183215,$$

aminek a hibája 0.0000396. □

Most bizonyítás nélkül felsorolunk néhány egyéb zárt elemi Newton–Cotes-formulát: Simpson $\frac{3}{8}$-ados formula:

$$\int_{x_0}^{x_3} f(x)\,dx = \frac{3h}{8}\Big(f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\Big) - \frac{3h^5}{80} f^{(4)}(\xi) \tag{7.34}$$

$n = 4$:

$$\int_{x_0}^{x_4} f(x)\,dx = \frac{2h}{45}\Big(7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)\Big) - \frac{8h^7}{945} f^{(6)}(\xi) \tag{7.35}$$

Végül levezetés és bizonyítás nélkül felsoroljuk az első néhány nyílt Newton–Cotes-formulát:

$$\int_{x_{-1}}^{x_1} f(x)\,dx = 2hf(x_0) + \frac{h^3}{3} f''(\xi), \tag{7.36}$$

$$\int_{x_{-1}}^{x_2} f(x)\,dx = \frac{3h}{2}\Big(f(x_0) + f(x_1)\Big) + \frac{3h^3}{4} f''(\xi), \tag{7.37}$$

$$\int_{x_{-1}}^{x_3} f(x)\,dx = \frac{4h}{3}\Big(2f(x_0) - f(x_1) + 2f(x_2)\Big) + \frac{14h^5}{45} f^{(4)}(\xi), \tag{7.38}$$

$$\int_{x_{-1}}^{x_4} f(x)\,dx = \frac{5h}{24}\Big(11f(x_0) + f(x_1) + f(x_2) + 11f(x_3)\Big) + \frac{95h^5}{144} f^{(4)}(\xi). \tag{7.39}$$

Zárjuk ezt a szakaszt a numerikus integrálás stabilitásának vizsgálatával.

7.9. tétel. *Legyen $\sum_{i=1}^n c_i f(x_i)$ egy olyan kvadratúra formula, amely pontos a konstans függvényekre és minden c_i együttható pozitív. Legyen y_i közelítése a pontos $f(x_i)$ függvényértékeknek, és tegyük fel, hogy $|y_i - f(x_i)| \le \varepsilon$. Ekkor*

$$\left|\sum_{i=1}^n c_i f(x_i) - \sum_{i=1}^n c_i y_i\right| \le \varepsilon(b-a).$$

Bizonyítás. A feltétel szerint $(b-a) = \int_a^b 1\,dx = \sum_{i=1}^n c_i$, ezért

$$\left|\sum_{i=1}^n c_i f(x_i) - \sum_{i=1}^n c_i y_i\right| \le \sum_{i=1}^n c_i |f(x_i) - y_i| \le \varepsilon \sum_{i=1}^n c_i = \varepsilon(b-a).$$

□

Megjegyezzük, hogy az összes ebben a fejezetben ismertetendő kvadratúra képlet pontos a konstans függvényekre, és a legtöbb pozitív súlyokat használ. Ezek a módszerek tehát numerikusan stabilak a függvény kerekítési hibájára nézve.

Feladatok

1. Számítsa ki a következő integrálok közelítő értékét a trapézformula segítségével $h = 0.5, 0.25, 0.125$ lépésközt használva:
   (a) $\int_0^1 \sin^3 x\,dx$,
   (b) $\int_1^2 \ln(x+1)\,dx$,
   (c) $\int_1^2 e^{1/x}\,dx$.
2. Ismételje meg az 1. feladatot a Simpson-formulát használva!
3. Ismételje meg az 1. feladatot a (7.34)-(7.35) formulákat használva!
4. Ismételje meg az 1. feladatot a (7.36)-(7.39) formulákat használva!
5. Mutassa meg, hogy a (7.26) érintőformula az $[x_i, x_{i+1}]$ intervallumok felezőpontjához húzott érintő alatti területek összegét adja vissza!
6. Mutassa meg, hogy az érintőformula a Newton–Cotes-formulák egyik speciális esete, és vezesse le az érintőformula hibatagját!
7. Vezesse le a (7.34)-(7.35) formulákat (a hibatag alakja nélkül)!
8. Vezesse le a (7.36)-(7.39) formulákat (a hibatag alakja nélkül)!
9. Vezesse le a Simpson-formula képletét a trapézformulából Richardson-extrapolációval!

7.4. Gauss-féle kvadratúra formulák

Az előző szakaszban láttuk, hogy a Newton–Cotes-formulák a pontos integrált adják vissza bizonyos fokszámú polinomok esetén. Ebben a szakaszban olyan kvadratúra képletek levezetésével foglalkozunk, amelyek hasonló tulajdonságúak. Tekintsük az

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} c_i f(x_i)$$

általános kvadratúra képletet. Teljesül a következő állítás:

7.10. tétel. *Egy*

$$Q(f) \equiv \sum_{i=1}^{n} c_i f(x_i) \tag{7.40}$$

kvadratúra formula akkor és csak akkor pontos egy tetszőleges $p(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0$ legfeljebb m-edfokú polinomra, ha pontos az x^i hatványfüggvényekre minden $i = 0, 1, \ldots, m$-re.

Bizonyítás. Abból, hogy Q pontos minden legfeljebb m-edfokú polinomra, természetesen következik, hogy pontos az x^i hatványfüggvényekre minden $i = 0, 1, \ldots, m$-re.

Most tegyük fel, hogy Q pontos az x^i hatványfüggvényekre minden $i = 0, 1, \ldots, m$-re. Ekkor az integrál és a Q kvadratúra formula linearitásából következik

$$\begin{aligned}
&\int_a^b a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0\,dx \\
&\qquad = a_m \int_a^b x^m\,dx + a_{m-1}\int_a^b x^{m-1}\,dx + \cdots + a_0 \int_a^b 1\,dx \\
&\qquad = a_m Q(x^m) + a_{m-1}Q(x^{m-1}) + \cdots + a_0 Q(1) \\
&\qquad = Q(a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0).
\end{aligned}$$

□

A (7.40) képlettel definiált Q kvadratúra formulában $2n$ darab paraméter szerepel, a c_i, x_i számok ($i = 1, 2, \ldots, n$). Azt várhatjuk tehát az előző tétel alapján, hogy egy ilyen kvadratúra képlet legfeljebb $(2n-1)$-edfokú polinomokra adjon vissza pontos értéket, hiszen azokban is $2n$ együttható van. A 7.10. tétel szerint ekkor a Q kvadratúra formula akkor és csak akkor pontos a legfeljebb $(2n-1)$-edfokú polinomokra, ha teljesül a következő $2n$ egyenlet:

$$\begin{aligned}
\int_a^b dx &= \sum_{i=1}^n c_i \\
\int_a^b x\,dx &= \sum_{i=1}^n c_i x_i \\
\int_a^b x^2\,dx &= \sum_{i=1}^n c_i x_i^2 \\
&\vdots \qquad\qquad \vdots \\
\int_a^b x^{2n-1}\,dx &= \sum_{i=1}^n c_i x_i^{2n-1}
\end{aligned} \tag{7.41}$$

Azt a (7.40) alakú kvadratúra formulát, amelyet a (7.41) egyenletrendszer megoldása segítségével írunk fel, n pontra felírt *Gauss-féle kvadratúra formulának* nevezzük.

Most tekintsünk egy speciális esetet, legyen $[a, b] = [-1, 1]$ és $n = 2$. Ekkor a (7.41) egyenletekből kapjuk az integrálokat kiszámolva

$$\begin{aligned}
2 &= c_1 + c_2 \\
0 &= c_1 x_1 + c_2 x_2 \\
\frac{2}{3} &= c_1 x_1^2 + c_2 x_2^2 \\
0 &= c_1 x_1^3 + c_2 x_2^3.
\end{aligned}$$

Könnyen ellenőrizhető, hogy az egyenletrendszernek egyértelmű megoldása van (a sorrendtől eltekintve): $c_1 = c_2 = 1$ és $x_1 = -\frac{\sqrt{3}}{3}$, $x_2 = \frac{\sqrt{3}}{3}$. A másodrendű Gauss-féle kvadratúra formula képlete tehát:

$$\int_{-1}^1 f(x)\,dx \approx f\left(-\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right). \tag{7.42}$$

7.11. példa. Számítsuk ki az $f(x) = e^x$ függvény integráljának egy közelítését a $[-1, 1]$ intervallumon! A (7.42) Gauss-formula alapján

$$\int_{-1}^1 e^x\,dx \approx e^{-\frac{\sqrt{3}}{3}} + e^{\frac{\sqrt{3}}{3}} = 2.3426961.$$

Ezt az $e - 1/e = 2.350424$ pontos értékkel összehasonlítva kapjuk, hogy a közelítés hibája 0.0077062, ami a képlet egyszerűségéhez viszonyítva nagyon kicsi. □

Szükségünk lesz az ortogonális függvények fogalmára. Az f és g függvényeket egymásra *ortogonálisnak* nevezzük az $[a, b]$ intervallumon, ha

$$\int_a^b f(x)g(x)\,dx = 0.$$

Megmutatjuk, hogy létezik polinomoknak egy olyan $(P_i)_{i=0,1,\dots}$ sorozata, amelyek páronként ortogonálisak a $[-1,1]$ intervallumon, és P_i i-edfokú polinom. Legyen $P_0(x) \equiv 1$ és $P_1(x) \equiv x$. Ekkor P_0 és P_1 ortogonális egymásra a $[-1,1]$ intervallumon. Keressük P_2-t a $P_2(x) = x^2 + a_{2,1}P_1(x) + a_{2,0}P_0(x)$ alakban. Ekkor a kívánt ortogonalitás alapján

$$\begin{aligned} 0 &= \int_{-1}^{1} P_2(x)P_0(x)\,dx \\ &= \int_{-1}^{1} x^2 P_0(x)\,dx + a_{2,1}\int_{-1}^{1} P_1(x)P_0(x)\,dx + a_{2,0}\int_{-1}^{1} P_0^2(x)\,dx \\ &= \int_{-1}^{1} x^2 P_0(x)\,dx + a_{2,0}\int_{-1}^{1} P_0^2(x)\,dx, \end{aligned}$$

amit megoldva

$$a_{2,0} = -\frac{\int_{-1}^{1} x^2 P_0(x)\,dx}{\int_{-1}^{1} P_0^2(x)\,dx}.$$

Ehhez hasonlóan

$$\begin{aligned} 0 &= \int_{-1}^{1} P_2(x)P_1(x)\,dx \\ &= \int_{-1}^{1} x^2 P_1(x)\,dx + a_{2,1}\int_{-1}^{1} P_1^2(x)\,dx + a_{2,0}\int_{-1}^{1} P_0(x)P_1(x)\,dx \\ &= \int_{-1}^{1} x^2 P_1(x)\,dx + a_{2,1}\int_{-1}^{1} P_1^2(x)\,dx, \end{aligned}$$

amiből

$$a_{2,1} = -\frac{\int_{-1}^{1} x^2 P_1(x)\,dx}{\int_{-1}^{1} P_1^2(x)\,dx}.$$

P_2-t tehát egyértelműen felírhatjuk a keresett alakban. Ezt az eljárást folytatva ha $P_0,\dots,$ P_i már definiált, P_{i+1}-et a

$$P_{i+1}(x) = x^{i+1} + a_{i+1,i}P_i(x) + \cdots + a_{i+1,0}P_0(x) \tag{7.43}$$

alakban keressük. Ekkor az előbbi számoláshoz hasonlóan kapjuk, hogy

$$a_{i+1,j} = -\frac{\int_{-1}^{1} x^{i+1}P_j(x)\,dx}{\int_{-1}^{1} P_j^2(x)\,dx}, \quad j = 0,1,\dots,i, \tag{7.44}$$

tehát P_{i+1} egyértelműen definiálható. Ezt az eljárást *Gram–Schmidt-féle ortogonalizálásnak* nevezzük, a kapott P_i polinomokat pedig i-edfokú *Legendre-polinomnak* hívjuk. Az első néhány Legendre-polinom képlete:

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_2(x) &= x^2 - \frac{1}{3}, \\ P_3(x) &= x^3 - \frac{3}{5}x, \\ P_4(x) &= x^4 - \frac{6}{7}x^2 + \frac{3}{35} \end{aligned}$$

Megmutatható hogy a Legendre-polinomok teljesítik a

$$P_{n+1}(x) = xP_n(x) - \frac{n^2}{4n^2-1}P_{n-1}(x) \tag{7.45}$$

rekurzív képletet. A Legendre-polinomok fontosabb tulajdonságait foglalja össze a következő tétel:

7.12. tétel. *Legyen P_i az i-edik Legendre-polinom. Ekkor*

1. *P_i ortogonális egy tetszőleges legfeljebb $(i-1)$-edfokú polinomra.*
2. *P_i páros függvény ha i páros, és páratlan függvény, ha i páratlan.*
3. *P_i-nek i darab különböző valós gyöke van a $(-1,1)$ intervallumban, amelyek szimmetrikusak az origóra nézve.*
4. *Ha $(p_i)_{i=0,1,\dots}$ (pontosan) i-edfokú, páronként ortogonális polinomok egy sorozata, akkor minden i-re $p_i(x) = c_iP_i(x)$ valamely $c_i \neq 0$ konstansra.*

Az alábbi tétel szerint az n pontra felírt Gauss-féle kvadratúra képlet alappontjai a P_n Legendre-polinom gyökeivel egyeznek meg.

7.13. tétel. *Tegyük fel, hogy az $x_1, x_2, \dots, x_n$ számok az n-edfokú Legendre-polinom gyökei, és legyen*

$$c_i = \int_{-1}^{1} \frac{(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}\,dx. \tag{7.46}$$

Ekkor egy tetszőleges legfeljebb $(2n-1)$-edfokú p polinomra

$$\int_{-1}^{1} p(x)\,dx = \sum_{i=1}^{n} c_i p(x_i).$$

A következő tétel a Gauss-féle kvadratúra formula képlethibáját adja meg.

7.14. tétel. *Legyen $f \in C^{2n}(a,b)$. Ekkor létezik olyan $\xi \in (a,b)$, hogy az n pontra felírt Gauss-féle kvadratúra formulára*

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} c_k f(x_k) + \frac{f^{(2n)}(\xi)}{(2n)!}\int_{-1}^{1} P_n^2(x)\,dx.$$

A 7.14. tételből belátható, hogy a Gauss-féle kvadratúra formula maradéktagja közelítőleg

$$\frac{\pi f^{(2n)}(\xi)}{4^n(2n)!}$$

alakú, azaz ha például $f^{(2n)}$ korlátos n-től független korláttal, akkor a Gauss-féle kvadratúra formula exponenciális sebességgel tart 0-hoz, ha $n \to \infty$. Emlékezzünk, hogy a Newton–Cotes-formulák csak polinomiális sebességgel tartanak 0-hoz, ha $n \to \infty$.

7.6. táblázat. A Gauss-féle kvadratúra formula paraméterei

n	x_i	c_i
2	0.5773502692	1.0000000000
	-0.5773502692	1.0000000000
3	0.7745966692	0.5555555556
	0.0000000000	0.8888888889
	-0.7745966692	0.5555555556
4	0.8611363116	0.3478548451
	0.3399810436	0.6521451549
	-0.3399810436	0.6521451549
	-0.8611363116	0.3478548451
5	0.9061798459	0.2369268850
	0.5384693101	0.4786286705
	0.0000000000	0.5688888889
	-0.5384693101	0.4786286705
	-0.9061798459	0.2369268850

A 7.6. táblázatban felsoroltuk az első néhány Legendre-polinom gyökeit, és az előző tételből kapott hozzá tartozó c_i együtthatók értékét.

A Gauss-féle kvadratúra képletek a $[-1,1]$ intervallumra vonatkoznak. Egy tetszőleges $[a,b]$ intervallumon vett integrált az $x=((b-a)t+a+b)/2$ változó helyettesítéssel tudunk a $[-1,1]$ intervallumra visszavezetni:

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^1 f\left(\frac{(b-a)t+a+b}{2}\right)\,dt.$$

7.15. példa. Közelítsük az $\int_0^1 x^2e^x\,dx$ integrált másodrendű Gauss-féle kvadratúra képlettel:

$$\begin{aligned}\int_0^1 x^2e^x\,dx &= \frac{1}{2}\int_{-1}^1 \left(\frac{t+1}{2}\right)^2 e^{(t+1)/2}\,dt\\ &\approx \frac{1}{2}\left(\left(\frac{-\sqrt{3}/3+1}{2}\right)^2 e^{(-\sqrt{3}/3+1)/2} + \left(\frac{\sqrt{3}/3+1}{2}\right)^2 e^{(\sqrt{3}/3+1)/2}\right)\\ &= 0.7119418.\end{aligned}$$

amelynek hibája 0.0063400. □

Feladatok

1. Alkalmazza a kétpontos Gauss-féle kvadratúra képletet az előző szakasz 1. feladatában felsorolt integrálokra!
2. Alkalmazza a 3, 4 és 5 pontra felírt Gauss-féle kvadratúra képleteket az előző szakasz 1. feladatában felsorolt integrálokra!

8. fejezet

Szélsőértékszámítás

Egy- és többváltozós függvények lokális szélsőértékei keresésével foglalkozunk ebben a fejezetben. Elegendő csak minimumkeresési módszereket vizsgálni, mivel egy $f(x)$ függvény ott veszi fel a maximumát, ahol a $-f(x)$ függvény a minimumát, így a maximum keresés mindig visszavezethető minimum keresésre.

Algoritmusoknak három nagy csoportjával foglalkozunk: deriváltat nem használó, csak első deriváltat használó és első és második deriváltat is igénylő módszerekkel. Az első csoportba tartozik az aranymetszés szerinti keresés, a szimplex és a Nelder–Mead-módszer, a másodikba a gradiens módszer, a harmadikba pedig a Newton-módszer. Ez utóbbi csoportba sorolhatók talán a kvázi-Newton módszerek is, melyek nem a pontos derivált és második derivált értéket használják, hanem annak valamilyen közelítését.

8.1. Analízis előismeretek

8.1. tétel. *Legyen $f\colon \mathbb{R}^n \to \mathbb{R}$ parciálisan differenciálható minden változója szerint. Ekkor ha f-nek létezik lokális szélsőértéke az $\mathbf{a}$ pontban, akkor $\frac{\partial f(\mathbf{a})}{\partial x_i} = 0$ teljesül minden $i = 1, \ldots, n$-re.*

Ha $f \in C^2$, és valamely $\mathbf{a}$ pontban $f'(\mathbf{a}) = \mathbf{0}$, továbbá az $f''(\mathbf{a})$ Hesse-mátrix pozitív (negatív) definit, akkor f-nek lokális minimuma (maximuma) van $\mathbf{a}$-ban.

Kétváltozós függvényekre az előbbi tétel speciális esetekén kapjuk:

8.2. tétel. *Legyen $f\colon \mathbb{R}^2 \to \mathbb{R}$, $f \in C^2$. Ekkor ha f-nek létezik lokális szélsőértéke az (a, b) pontban, akkor*

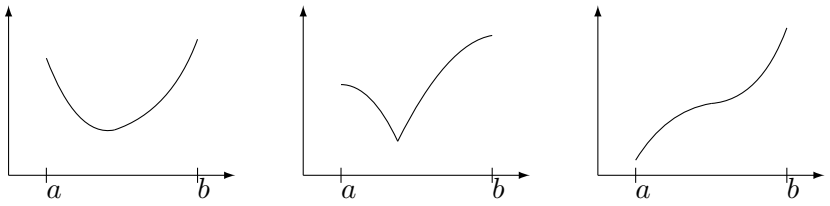
$$\frac{\partial f}{\partial x}(a, b) = 0, \qquad \frac{\partial f}{\partial y}(a, b) = 0 \tag{8.1}$$

teljesül.

Fordítva, ha valamely (a, b)-re (8.1) teljesül, továbbá

$$D(a, b) := \frac{\partial^2 f}{\partial x^2}(a, b) \cdot \frac{\partial^2 f}{\partial y^2}(a, b) - \left(\frac{\partial^2 f}{\partial x\, \partial y}(a, b)\right)^2 > 0$$

akkor f-nek létezik lokális szélsőértéke (a, b)-ben, mégpedig lokális maximuma, ha $\frac{\partial^2 f}{\partial x^2}(a, b) < 0$ ill. lokális minimuma, ha $\frac{\partial^2 f}{\partial x^2}(a, b) > 0$. Ha $D(a, b) < 0$, akkor f-nek nincs szélsőértéke (a, b)-ben.

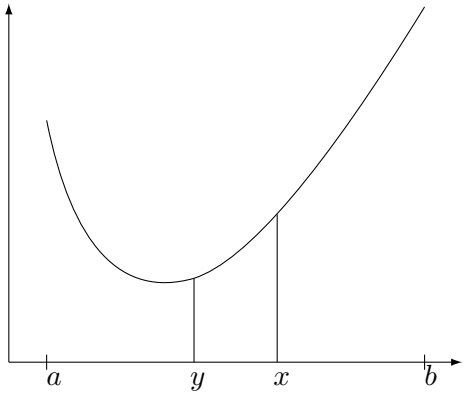


8.1. ábra. Unimodális függvények

8.2. Aranymetszés szerinti keresés módszere

Legyen $f\colon [a,b] \to \mathbb{R}$ folytonos, és feltesszük, hogy f *unimodális*, azaz f-nek egyértelmű lokális minimuma van $[a,b]$-ben. Ez teljesül pl. ha a függvény konvex az $[a,b]$ intervallumon, de a konvexitás nem szükséges ahhoz, hogy egy függvény unimodális legyen (lásd például a 8.1. ábrán szereplő második és harmadik függvényt). Jelölje p az f függvény minimumhelyét.

Az *aranymetszés szerinti keresés módszernél*, az intervallumfelezés módszeréhez hasonlóan, egyre szűkebb és szűkebb intervallumokra határoljuk be a függvény minimumhelyét: Legyen $a < y < x < b$. Ha $f(x) > f(y)$, akkor $p \in [a,x]$, ellenkező esetben $p \in [y,b]$ teljesül. (Lásd a 8.2. ábrát!) Ezután az $[a,x]$ illetve $[y,b]$ intervallummal folytatjuk az eljárást.



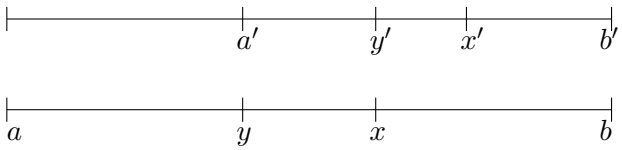
8.2. ábra.

Az x és y pontokat úgy választjuk, hogy az $[a,x]$ és $[y,b]$ intervallum hossza azonos legyen: $x - a = b - y = r(b-a)$ valamely $0 < r <$ 1-re. Ekkor

$$x = a + r(b-a), \qquad y = a + (1-r)(b-a) \tag{8.2}$$

alakú. Az $x > y$ feltételből kapjuk, hogy $0.5 < r < 1$ kell legyen. Jelölje $[a',b']$ a következő intervallumot. Válasszuk az új osztópontokat, x'-t és y'-t a (8.2) szabály szerint, és $f(x')$ és $f(y')$ összehasonlításával határozzuk meg a következő intervallumot. Még nem definiáltuk r-t. Az aranymetszés szerinti keresés módszere úgy választja meg r-t, hogy az új x', y' osztópontok közül az egyik egyezzen meg egy előző osztóponttal, azért hogy minden lépésben csak egy új függvényértéket kelljen kiértékelni.

A 8.3. ábrán azt az esetet tüntettük fel, ahol a jobb oldali, $[y,b]$ intervallumba esik a minimumhely. Ekkor azt követeljük meg az osztópontok választásától, hogy $y' = x$ legyen.



8.3. ábra.

Ekkor teljesülnek a következő összefüggések:

$$\begin{aligned} a + r(b-a) &= y' \\ &= a' + (1-r)(b'-a') \\ &= y + (1-r)(b-y) \\ &= a + (1-r)(b-a) + (1-r)(b-a-(1-r)(b-a)), \end{aligned}$$

és így

$$r = 1 - r + (1-r)(1-(1-r)),$$

amiből

$$r^2 + r - 1 = 0 \tag{8.3}$$

következik. Ennek pozitív megoldása $r = (\sqrt{5}-1)/2 \approx 0.61834$. Ez az aranymetszés arányossági tényezője: r teljesíti az

$$\frac{r}{1-r} = \frac{1}{r}$$

egyenlet.

Abban az esetben, amikor az $[a, x]$ intervallumban van a minimumhely, akkor úgy választjuk x', y'-t, hogy $x' = y$ legyen. Megmutatható (3. feladat), hogy ez a követelmény is a (8.3) egyenlethez vezet.

8.3. algoritmus. Aranymetszés szerinti keresés módszere

INPUT: $f(x)$,
$[a, b]$,
ε, - tolerancia
OUTPUT: p - a minimumhely közelítése

$r \leftarrow (\sqrt{5}-1)/2$
$x \leftarrow a + r(b-a)$
$y \leftarrow a + (1-r)(b-a)$
$fx \leftarrow f(x)$
$fy \leftarrow f(y)$
while $(b-a) > \varepsilon$ **do**
    if $fx > fy$ **do**
        $b \leftarrow x$
        $x \leftarrow y$
        $fx \leftarrow fy$
        $y \leftarrow a + (1-r)(b-a)$
        $fy \leftarrow f(y)$
    else do
        $a \leftarrow y$

$y \leftarrow x$
$fy \leftarrow fx$
$x \leftarrow a + r(b-a)$
$fx \leftarrow f(x)$
end do
end do
output$((a+b)/2)$

Könnyen igazolható a következő tétel:

8.4. tétel. *Legyen $f \in C(a,b)$ unimodális függvény. Ekkor az aranymetszés szerinti keresés módszere konvergál az f függvény minimumhelyéhez.*

Könnyű ellenőrizni, hogy az aranymetszés szerinti keresés módszere n lépése után az intervallum hossza $(b-a)r^n$ lesz. Így a 8.3. algoritmus az ε tolerancia értéket

$$n \geq \frac{\log \frac{\varepsilon}{b-a}}{\log r} \tag{8.4}$$

lépésben éri el.

8.5. példa. Keressük meg az $f(x) = x^2 - 0.8x + 1$ függvény minimumhelyét! Könnyű kiszámolni, hogy a függvény a minimumát a $p = 0.4$ pontban veszi fel. A 8.3. algoritmust alkalmaztuk az adott függvényre a $[-1, 2]$ kezdeti intervallumot és az $\varepsilon = 0.005$ tolerancia értéket használva, amelynek eredménye a 8.1. táblázatban látható. A (8.4) formula szerint $n \geq 13.29337586$ lépés kell az előírt tolerancia eléréséhez. A minimumhely az utolsó lépésben kapott $[0.3977741449, 0.4013328688]$ intervallumban helyezkedik el, a 8.3. algoritmus az intervallum felezőpontját, 0.3995535068-t adja meg, mint közelítő értéket. □

8.1. táblázat. Aranymetszés szerinti keresés módszere, $f(x) = x^2 - 0.8x + 1$

k	$[a_k, b_k]$	y_k	x_k
0	[-1.0000000000, 2.0000000000]	0.1458980338	0.8541019662
1	[-1.0000000000, 0.8541019662]	-0.2917960675	0.1458980338
2	[-0.2917960675, 0.8541019662]	0.1458980338	0.4164078650
3	[0.1458980338, 0.8541019662]	0.4164078650	0.5835921350
4	[0.1458980338, 0.5835921350]	0.3130823038	0.4164078650
5	[0.3130823038, 0.5835921350]	0.4164078650	0.4802665738
6	[0.3130823038, 0.4802665738]	0.3769410125	0.4164078650
7	[0.3769410125, 0.4802665738]	0.4164078650	0.4407997213
8	[0.3769410125, 0.4407997213]	0.4013328688	0.4164078650
9	[0.3769410125, 0.4164078650]	0.3920160087	0.4013328688
10	[0.3920160087, 0.4164078650]	0.4013328688	0.4070910050
11	[0.3920160087, 0.4070910050]	0.3977741449	0.4013328688
12	[0.3977741449, 0.4070910050]	0.4013328688	0.4035322811
13	[0.3977741449, 0.4035322811]	0.3999735572	0.4013328688
14	[0.3977741449, 0.4013328688]	0.3991334565	0.3999735572

Feladatok

1. Az aranymetszés szerinti keresés módszerét alkalmazva keresse meg az alábbi függvények minimumhelyét az adott intervallumon:

   (a) $f(x) = x^3 - 3x + 1$, $x \in [-1, 2]$, (b) $f(x) = |\cos x|$, $x \in [0, 2]$,

   (c) $f(x) = 1 - 10xe^{-x}$, $x \in [0, 2]$, (d) $f(x) = \cos(x^2 - x)$, $x \in [1, 3]$.

2. Alkalmazza az aranymetszés szerinti keresés módszerét az $f(x) = -1/x^2$ függvényre a $[-1,1]$ intervallumon! Mit tapasztal?
3. Igazolja, hogy az $[a',b'] = [a,x]$ választáskor az $x' = y$ egyenlet akkor teljesül, ha r megoldása a (8.3) egyenletnek!
4. Bizonyítsa be a 8.4. tételt!
5. Ellenőrizze a (8.4) formulát!

8.3. Szimplex módszer

Egy n-dimenziós *szimplexen* olyan $n+1$ darab n-dimenziós vektor konvex burkát, azaz az

$$\{\alpha_0\mathbf{x}^{(0)} + \cdots + \alpha_n\mathbf{x}^{(n)}\colon\ 0 \le \alpha_i \le 1, \quad \alpha_0 + \cdots + \alpha_n \le 1\}$$

halmazt értjük, ahol az $\mathbf{x}_1 - \mathbf{x}_0, \mathbf{x}_2 - \mathbf{x}_0, \ldots, \mathbf{x}_n - \mathbf{x}_0$ vektorok lineárisan függetlenek. Ekkor az $\mathbf{x}^{(0)}$,...,$\mathbf{x}^{(n)}$ vektorokat a szimplex csúcspontjainak hívjuk. Egydimenziós szimplexek a szakaszok, kétdimenziós szimplexek a háromszögek, háromdimenziós szimplexek pedig a tetraéderek.

A *szimplex módszert* n-változós függvények minimumhely keresésére használjuk. Vegyünk fel kiindulásként egy n-dimenziós szimplexet. Keressük meg, hogy melyik a „legrosszabb" csúcspont, azaz melyik csúcspontban veszi fel az f függvény a legnagyobb értéket. Legyen ez például az $\mathbf{x}^{(j)}$ pont. Ekkor a szimplex legrosszabb pontját tükrözzük az $\mathbf{x}^{(j)}$ ponttal szemben fekvő oldal középpontjára, azaz a többi csúcspont

$$\mathbf{x}_c := \frac{1}{n}\sum_{\substack{i=0\\ i\neq j}}^{n} \mathbf{x}^{(i)}$$

súlypontjára. A tükrözött pont koordinátáit az

$$\mathbf{x}_r = 2\mathbf{x}_c - \mathbf{x}^{(j)}$$

képlettel számíthatjuk ki. Ha $f(\mathbf{x}_r)$ nem kisebb, mint az előző lépésbeli legnagyobb függvényérték, $f(\mathbf{x}^{(j)})$, akkor a tükrözést nem fogadjuk el, hanem ahelyett a legjobb csúcspontból fele akkorára zsugorítjuk a szimplexet: legyen $\mathbf{x}^{(k)}$ a legjobb csúcspontja a szimplexnek, azaz ebben a legkisebb a függvényérték. Ekkor a többi csúcspontot az

$$\mathbf{x}^{(i)} \leftarrow \mathbf{x}^{(k)} + \frac{1}{2}(\mathbf{x}^{(i)} - \mathbf{x}^{(k)}), \quad i = 0,1,\ldots,k-1,k+1,\ldots,n$$

képlettel számoljuk újra. Ezután a kapott (tükrözött vagy zsugorított) szimplexszel megismételjük az eljárást.

Az előbbi iterációs módszerhez többféle megállási feltételt, illetve feltétel kombinációt adhatunk meg. Például megkövetelhetjük, hogy az eljárás akkor érjen véget, ha a szimplex egy előre megadott méretnél kisebb lesz. A szimplex méretét definiálhatjuk például a leghosszabb éle hosszaként, azaz a $\max\{\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|\colon\ i,j = 0,\ldots,n\}$ képlettel. Egy másik lehetőség lehet az, hogy a szimplexek súlypontjaiban felvett f_k függvényérték sorozatára alkalmazzuk az $|f_{k+1} - f_k| < \varepsilon$ feltételt. Egy harmadik megállási feltétel lehet a következő: Legyen $\bar{f}$ a csúcspontokban felvett függvényértékek átlaga, σ pedig a szórása, azaz

$$\bar{f} = \frac{1}{n+1}\sum_{i=0}^{n} f(\mathbf{x}^{(i)}), \qquad \sigma = \sqrt{\frac{1}{n+1}\sum_{i=0}^{n}(f(\mathbf{x}^{(i)}) - \bar{f})^2}.$$

Ekkor addig folytatjuk az iterációt, amíg σ kisebb nem lesz mint egy előre megadott tolerancia érték. A függvény minimumhelyét az algoritmus utolsó lépésében kapott szimplex súlypontjával szokás közelíteni.

8.6. példa. Keressük meg az $f(x,y) = (x^2-2y)^2 + 2(x-1)^2$ függvény minimumhelyét! Könnyen látható, hogy a függvénynek az $(1, 0.5)$ pontban van (globális) minimuma. A szimplex módszert alkalmaztuk a feladat megoldására a $(-2,4)$, $(-1,4)$ és $(-1.5,5)$ kezdeti háromszögből kiindulva. Az első 25 lépésben kapott háromszögeket és a csúcspontokhoz tartozó függvényértékeket a 8.2. táblázatban soroltuk fel. A 8.4. ábrán láthatók f szintvonalai és az egyes lépésekben kapott háromszögek. A 25. háromszög középpontja, $(0.9063, 0.3542)$, jó közelítése a pontos minimumhelynek. Az ebben a pontban felvett függvényérték 0.0303, ami közel van a pontos minimum értékhez, 0-hoz. □

8.2. táblázat. Szimplex módszer, $f(x,y) = (x^2-2y)^2 + 2(x-1)^2$

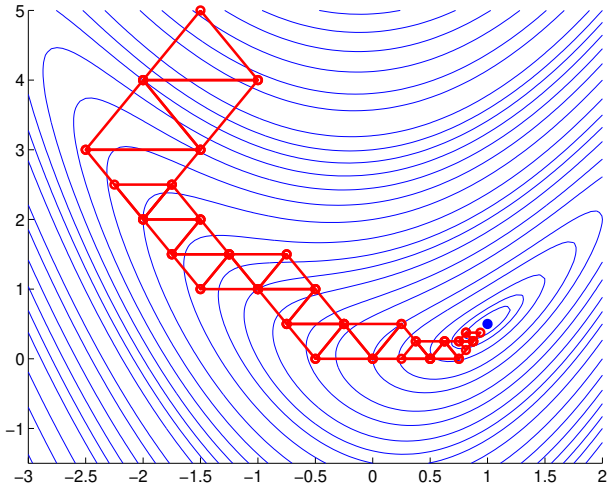
k	$\mathbf{x}^{(k,1)}$	$\mathbf{x}^{(k,2)}$	$\mathbf{x}^{(k,3)}$	$f(\mathbf{x}^{(k,1)})$	$f(\mathbf{x}^{(k,2)})$	$f(\mathbf{x}^{(k,3)})$
0	(-1.000, 4.000)	(-2.000, 4.000)	(-1.500, 5.000)	57.000	34.000	72.563
1	(-2.000, 4.000)	(-1.000, 4.000)	(-1.500, 3.000)	34.000	57.000	26.563
2	(-1.500, 3.000)	(-2.000, 4.000)	(-2.500, 3.000)	26.563	34.000	24.563
3	(-2.500, 3.000)	(-1.500, 3.000)	(-2.000, 2.000)	24.563	26.563	18.000
4	(-2.000, 2.000)	(-2.250, 2.500)	(-1.750, 2.500)	18.000	21.129	18.879
5	(-2.000, 2.000)	(-1.750, 2.500)	(-1.500, 2.000)	18.000	18.879	15.563
6	(-1.500, 2.000)	(-2.000, 2.000)	(-1.750, 1.500)	15.563	18.000	15.129
7	(-1.750, 1.500)	(-1.500, 2.000)	(-1.250, 1.500)	15.129	15.563	12.191
8	(-1.250, 1.500)	(-1.750, 1.500)	(-1.500, 1.000)	12.191	15.129	12.563
9	(-1.250, 1.500)	(-1.500, 1.000)	(-1.000, 1.000)	12.191	12.563	9.000
10	(-1.000, 1.000)	(-1.250, 1.500)	(-0.750, 1.500)	9.000	12.191	12.066
11	(-1.000, 1.000)	(-0.750, 1.500)	(-0.500, 1.000)	9.000	12.066	7.563
12	(-0.500, 1.000)	(-1.000, 1.000)	(-0.750, 0.500)	7.563	9.000	6.316
13	(-0.750, 0.500)	(-0.500, 1.000)	(-0.250, 0.500)	6.316	7.563	4.004
14	(-0.250, 0.500)	(-0.750, 0.500)	(-0.500, 0.000)	4.004	6.316	4.563
15	(-0.250, 0.500)	(-0.500, 0.000)	(0.000, 0.000)	4.004	4.563	2.000
16	(0.000, 0.000)	(-0.250, 0.500)	(0.250, 0.500)	2.000	4.004	2.004
17	(0.000, 0.000)	(0.250, 0.500)	(0.500, 0.000)	2.000	2.004	0.563
18	(0.500, 0.000)	(0.250, 0.000)	(0.375, 0.250)	0.563	1.129	0.910
19	(0.500, 0.000)	(0.375, 0.250)	(0.625, 0.250)	0.563	0.910	0.293
20	(0.625, 0.250)	(0.500, 0.000)	(0.750, 0.000)	0.293	0.563	0.441
21	(0.625, 0.250)	(0.750, 0.000)	(0.875, 0.250)	0.293	0.441	0.102
22	(0.875, 0.250)	(0.750, 0.250)	(0.813, 0.125)	0.102	0.129	0.239
23	(0.875, 0.250)	(0.750, 0.250)	(0.813, 0.375)	0.102	0.129	0.078
24	(0.813, 0.375)	(0.875, 0.250)	(0.938, 0.375)	0.078	0.102	0.024
25	(0.938, 0.375)	(0.875, 0.375)	(0.906, 0.313)	0.024	0.031	0.056

A szimplex módszernek egy módosított változata a *Nelder–Mead-módszer*. Ennél a módszernél a szimplexet tükrözzük, illetve megnyújtjuk vagy zsugorítjuk aszerint, hogy milyen értékeket vesz fel a függvény a csúcspontokban. Feltesszük, hogy minden egyes lépésben a csúcspontokat úgy indexezzük, hogy a függvényértékek növekvő sorrendben legyenek, azaz $f(\mathbf{x}^{(0)}) \le f(\mathbf{x}^{(1)}) \le \cdots \le f(\mathbf{x}^{(n)})$. Ekkor $\mathbf{x}^{(n)}$ a legrosszabb csúcspont, ezt tükrözzük a szemben fekvő oldal

$$\mathbf{x}_c = \frac{1}{n}\sum_{i=0}^{n-1} \mathbf{x}^{(i)}$$

súlypontjára. Legyen a tükrözött pont $\mathbf{x}_r = 2\mathbf{x}_c - \mathbf{x}^{(n)}$. Vizsgáljuk meg, hogy milyen értéket vesz fel az f függvény $\mathbf{x}_r$-ben. Három esetet különböztetünk meg: 1. $f(\mathbf{x}^{(0)}) < f(\mathbf{x}_r) < f(\mathbf{x}^{(n-1)})$, 2. $f(\mathbf{x}_r) \le f(\mathbf{x}^{(0)})$, azaz $\mathbf{x}_r$ lenne az új legjobb pont, és 3. $f(\mathbf{x}_r) \ge f(\mathbf{x}^{(n-1)})$, azaz $\mathbf{x}_r$ lenne az új legrosszabb pont.

Az 1. esetben $\mathbf{x}^{(n)}$-t $\mathbf{x}_r$-re kicseréljük (elfogadtuk a tükrözést), és folytatjuk az iterációt.



8.4. ábra. Szimplex módszer.

A 2. esetben először megpróbáljuk az $\mathbf{x}_r$ irányban megnyújtani egy kicsit a szimplexet, hátha még jobb pontot kapunk. Legyen

$$\mathbf{x}_e := \mathbf{x}_c + \alpha(\mathbf{x}_r - \mathbf{x}_c),$$

ahol $\alpha > 1$ egy rögzített szám (egy paraméter a módszerben). Ha ekkor $f(\mathbf{x}_e) < f(\mathbf{x}^{(0)})$ teljesül, akkor a megnyújtást sikeresnek ítéljük, és $\mathbf{x}^{(n)}$-t $\mathbf{x}_e$-re cseréljük ki. Ellenkező esetben viszont $\mathbf{x}^{(n)}$-t $\mathbf{x}_r$-re cseréljük ki, azaz tükrözünk, de nem nyújtjuk meg a szimplexet.

A 3. esetben azt gondoljuk, hogy túl messze tükröztük $\mathbf{x}^{(n)}$-t, így megpróbáljuk zsugorítani a szimplexet. Legyen

$$\mathbf{x}_z := \begin{cases} \mathbf{x}_c - \beta(\mathbf{x}_r - \mathbf{x}_c), & \text{ha } f(\mathbf{x}^{(n)}) < f(\mathbf{x}_r), \\ \mathbf{x}_c + \beta(\mathbf{x}_r - \mathbf{x}_c), & \text{ha } f(\mathbf{x}^{(n)}) \geq f(\mathbf{x}_r), \end{cases}$$

ahol $0 < \beta < 1$ egy újabb paraméter. Ha $f(\mathbf{x}_z) < \min\{f(\mathbf{x}^{(n)}),\ f(\mathbf{x}_r)\}$, akkor $\mathbf{x}^{(n)}$-t $\mathbf{x}_z$-vel cseréljük fel. Ellenkező esetben viszont a szimplexet a legjobb pontjából, $\mathbf{x}^{(0)}$-ból a felére zsugorítjuk össze:

$$\mathbf{x}^{(i)} \leftarrow \mathbf{x}^{(0)} + \frac{1}{2}(\mathbf{x}^{(i)} - \mathbf{x}^{(0)}), \quad i = 1, \ldots, n.$$

8.7. példa. A Nelder–Mead-módszert alkalmaztuk az $\alpha = 1.4$ és $\beta = 0.7$ paraméter értékekkel a 8.6. feladatban már vizsgált $f(x, y) = (x^2 - 2y)^2 + 2(x - 1)^2$ függvény minimumhelyének keresésére. Most is a $(-2, 4)$, $(-1, 4)$ és $(-1.5, 5)$ háromszögből indultunk ki. A kapott sorozat első 17 tagja látható a 8.3. táblázatban, illetve a 8.5. ábrán. A 17. háromszög középpontja $(1.0071, 0.5929)$, a hozzá tartozó függvényérték pedig 0.0295. Látható, hogy ez a módszer sokkal gyorsabban konvergál a minimumhelyhez, mint a szimplex módszer. □

Feladatok

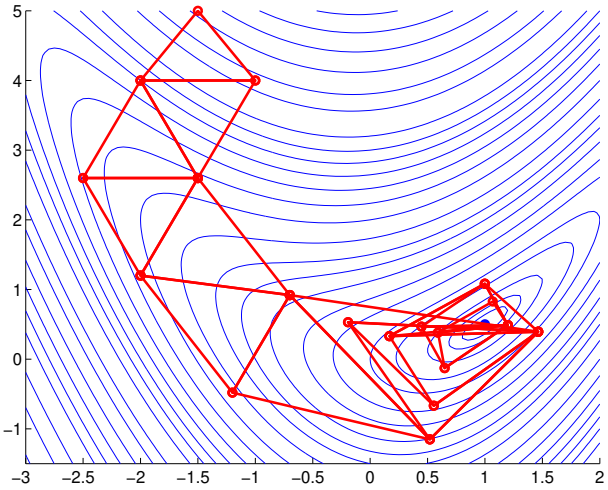
1. Keresse meg a következő függvények minimumhelyét Nelder–Mead-módszerrel!

   (a) $f(x, y) = x^2 + 5y^2$, (b) $f(x, y) = x^2 + (x + y - 2)^2$,

   (c) $f(x, y) = 3x^2 + e^{(x-y)^2}$, (d) $f(x, y) = x^2 + \cos^2(x - y)$

8.3. táblázat. Nelder–Mead-módszer, $f(x,y) = (x^2 - 2y)^2 + 2(x-1)^2$

k	$\mathbf{x}^{(k,1)}$	$\mathbf{x}^{(k,2)}$	$\mathbf{x}^{(k,3)}$	$f(\mathbf{x}^{(k,1)})$	$f(\mathbf{x}^{(k,2)})$	$f(\mathbf{x}^{(k,3)})$
0	(-1.000, 4.000)	(-2.000, 4.000)	(-1.500, 5.000)	57.000	34.000	72.563
1	(-2.000, 4.000)	(-1.000, 4.000)	(-1.500, 2.600)	34.000	57.000	21.203
2	(-1.500, 2.600)	(-2.000, 4.000)	(-2.500, 2.600)	21.203	34.000	25.603
3	(-1.500, 2.600)	(-2.500, 2.600)	(-2.000, 1.200)	21.203	25.603	20.560
4	(-2.000, 1.200)	(-1.500, 2.600)	(-0.700, 0.920)	20.560	21.203	7.602
5	(-0.700, 0.920)	(-2.000, 1.200)	(-1.200,-0.480)	7.602	20.560	15.440
6	(-0.700, 0.920)	(-1.200,-0.480)	(0.520,-1.152)	7.602	15.440	7.088
7	(0.520,-1.152)	(-0.700, 0.920)	(1.464, 0.394)	7.088	7.602	2.270
8	(1.464, 0.394)	(0.520,-1.152)	(-0.192, 0.530)	2.270	7.088	3.891
9	(1.464, 0.394)	(-0.192, 0.530)	(0.555,-0.668)	2.270	3.891	3.097
10	(1.464, 0.394)	(0.555,-0.668)	(0.168, 0.330)	2.270	3.097	1.783
11	(0.168, 0.330)	(1.464, 0.394)	(0.999, 1.083)	1.783	2.270	1.362
12	(0.999, 1.083)	(0.168, 0.330)	(1.200, 0.487)	1.362	1.783	0.296
13	(1.200, 0.487)	(0.999, 1.083)	(0.448, 0.467)	0.296	1.362	1.147
14	(1.200, 0.487)	(0.448, 0.467)	(0.648,-0.129)	0.296	1.147	0.707
15	(1.200, 0.487)	(0.648,-0.129)	(0.591, 0.380)	0.296	0.707	0.505
16	(1.200, 0.487)	(0.591, 0.380)	(1.068, 0.828)	0.296	0.505	0.274
17	(1.068, 0.828)	(1.200, 0.487)	(0.754, 0.464)	0.274	0.296	0.251



8.5. ábra. Nelder–Mead-módszer.

Több különböző α, β paraméter értékekkel próbálja ki a módszert!

2. Alkalmazza a Nelder–Mead-módszert tetszőleges $\alpha > 1$ és $0 < \beta < 1$ paraméter értékeket használva az $f(x) = x^2 - y^2$ függvényre és a $[0,1]$, $[0,-1]$, $[1,0]$ kezdeti pontokra! Mit tapasztal? Mit tapasztal, ha ugyanerre a függvényre és pontokra a szimplex módszert alkalmazza?

3. Fogalmazza meg a szimplex módszert egyváltozós függvények minimumhelyének meghatározására, és alkalmazza a 8.2. szakasz 1. feladatában szereplő függvényekre!

4. Tekintsük a következő, deriváltat nem használó módszert kétváltozós függvények minimalizálására: legyen f egy kétváltozós függvény, $(p_1^{(0)}, p_2^{(0)})$ egy adott kezdeti pont. Minimalizáljuk a $t \mapsto f(p_1^{(0)} + t, p_2^{(0)})$ egyváltozós függvényt (például szimplex módszerrel, lásd az előző példát). Legyen t_1 a minimumhely. Ekkor jelölje $(p_1^{(1)}, p_2^{(1)}) := (p_1^{(0)} + t_1, p_2^{(0)})$. Ezután minimalizáljuk a $t \mapsto f(p_1^{(1)}, p_2^{(1)} + t)$ egyváltozós függvényt. A kapott t_2 minimumhelyhez tartozó $(p_1^{(2)}, p_2^{(2)}) = (p_1^{(1)}, p_2^{(1)} + t_2)$ pontból megismételjük az eljárást. Így felváltva az x- illetve y-tengely irányában egydimenziós minimumkeresési feladatokat megoldva kapjuk a sorozat következő pontját. Alkalmazza ezt a

módszert az 1. feladatban felsorolt függvényekre! Hasonlítsa össze a kapott sorozat konvergenciájának gyorsaságát a Nelder–Mead-módszer gyorsaságával!

8.4. Gradiens módszer

Tekintsünk egy $f : \mathbb{R}^n \to \mathbb{R}$ függvényt. Analízisből ismert tétel szerint egy $\mathbf{p}$ pontban az f függvény a $-f'(\mathbf{p})$ irányban csökken a leggyorsabban:

8.8. tétel. *Legyen $f \in C^1$. Ekkor a*

$$\lim_{t\to 0+} \frac{f(\mathbf{p}+t\mathbf{u}) - f(\mathbf{p})}{t}, \qquad \|\mathbf{u}\|_2 = 1$$

iránymenti deriváltak minimuma az $\mathbf{u} = -f'(\mathbf{p})/\|f'(\mathbf{p})\|_2$ irányban van.

Egy $\mathbf{u}$ irányt az f függvény $\mathbf{p}$ pontbeli *lejtőjének* nevezzük, ha létezik olyan $\delta > 0$, hogy $f(\mathbf{p}+t\mathbf{u}) < f(\mathbf{p})$ minden $0 < t < \delta$-ra, azaz a függvény csökken a $\mathbf{p}$ pontból az $\mathbf{u}$ irány mentén indulva. A 8.8. tételt úgy is megfogalmazhatjuk, hogy az f függvénynek a $\mathbf{p}$ pontban a $-f'(\mathbf{p})$ irányban legmeredekebb a lejtője.

A *gradiens módszer* szerint egy $\mathbf{p}^{(0)}$ kezdeti pontból a negatív gradiensvektor irányában kell elmozdulni. Szokás az előbbiek miatt ezt a *legmeredekebb lejtő módszerének* is nevezni. A módszer általános képlete ezért:

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} - \alpha_k f'(\mathbf{p}^{(k)}), \tag{8.5}$$

ahol α_k a lépésközt meghatározó szorzótényező. A (8.5) gradiens módszernek több változata van. A legegyszerűbb esetben a lépésköz állandó. Legyen h rögzített, és használjuk az $\alpha_k = h/\|f'(\mathbf{p}^{(k)})\|_2$ számot. Ekkor az egyes pontok közötti távolság konstans h lesz. Természetesen ekkor általában nem várható el, hogy h-nál pontosabban megközelítsük a minimumhelyet.

Egy másik változatban úgy választjuk meg a lépésközt, hogy

$$\phi_k(\alpha_k) = \min_{t\in\mathbb{R}} \phi_k(t)$$

legyen, ahol

$$\phi_k(t) := f\Big(\mathbf{p}^{(k)} - tf'(\mathbf{p}^{(k)})\Big). \tag{8.6}$$

Ekkor minden egyes lépésben a gradiensvektor által meghatározott egyenes mentén egy egyváltozós függvényt kell minimalizálni. Ez utóbbi módon választott lépésközt használó gradiens módszert *optimális gradiens módszernek* hívjuk.

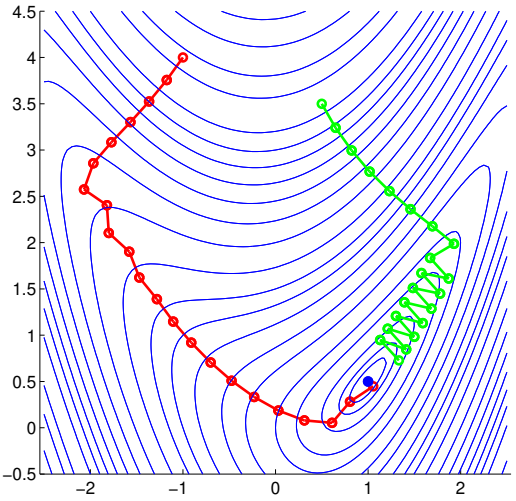
Az optimális gradiens módszernél a gradiensvektorral párhuzamos egyenes mentén egy olyan pontig lépünk, ahol az egyenes érint egy szintvonalat. Abból a pontból pedig a pontbeli gradiensvektorral párhuzamosan lépünk tovább. Ebből következik, hogy az optimális gradiens módszernél az egymás utáni lépések irányai merőlegesek egymásra. (Lásd a 3. feladatot!)

Megmutatható, hogy az optimális gradiens módszer lokálisan lineárisan konvergens. A sorozat aszimptotikus hibakonstansa néha közel van 1-hez, azaz a konvergencia lassú is lehet.

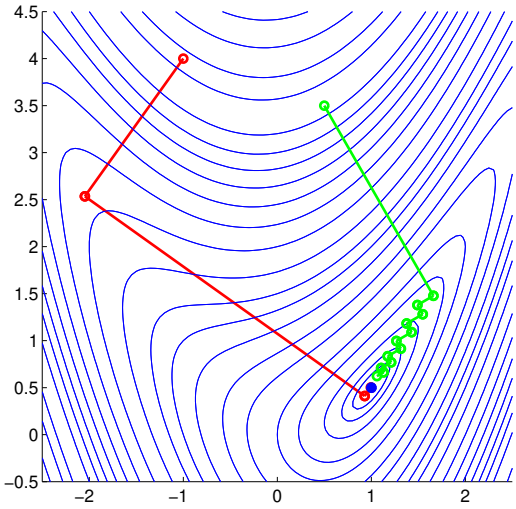
8.9. példa. Tekintsük újra a 8.6. és 8.7. példákban vizsgált $f(x,y) = (x^2-2y)^2 + 2(x-1)^2$ függvényt. Először az $\alpha_k = 0.3/\|f'(\mathbf{p}^{(k)})\|_2$ lépésközzel futtajuk a gradiens módszert, két kezdeti pontból indítva a módszert: a $(-1,4)$ kezdeti értékből (piros karikák) és a $(0.5, 3.5)$ kezdeti értékből (zöld karikák). A kapott sorozatok első 25 tagja a 8.6. ábrán látható. A sorozatok lassan közelítik meg az $(1, 0.5)$

minimumhelyet (kék pont), és annak közelében oszcillálnak. Vegyük észre, hogy ahogy az az analízisből ismert, a gradiensvektor merőleges a ponthoz tartozó szintvonalra, így a gradiens módszer sorozata mindig a szintvonalra merőleges irányban mozdul el.

Ezután az optimális gradiens módszert alkalmaztuk a $(-1, 4)$ és az $(0.5, 3.5)$ kezdőpontból indulva. A két sorozat első 3 illetve 12 tagját a 8.7. ábrán láthatjuk. Az első sorozat gyorsan a minimumhely közelébe került. A második is gyorsan a minimumhelyet tartalmazó hosszúkás „völgybe” került, de ezután ott csak lassan, cikcakkban haladt a minimumhely felé. □



8.6. ábra. Gradiens módszer konstans lépésközt használva.



8.7. ábra. Optimális gradiens módszer.

Ha f gradiensvektorát nem tudjuk vagy nem akarjuk kiszámolni (túl sok műveletet igényel), használhatjuk (8.5) következő változatát:

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} - \alpha_k \mathbf{v}^{(k)}, \tag{8.7}$$

ahol a $\mathbf{v}^{(k)}$ vektor i-edik komponensét a

$$v_i^{(k)} = \frac{1}{h}\Big(f(\mathbf{p}^{(k)} + h\mathbf{e}^{(i)}) - f(\mathbf{p}^{(k)})\Big), \qquad i = 1, \ldots, n$$

képlettel számoljuk ($\mathbf{e}^{(i)}$ az i-edik egységvektor).

Feladatok

1. Alkalmazza a gradiens módszert a 8.3. szakasz 1. feladatában felsorolt függvényekre! Válasszon tetszőleges kezdőpontot, és használja az $\alpha_k = h/\|f'(\mathbf{p}^{(k)})\|_2$ lépésközt valamely $h > 0$-ra, illetve az optimális gradiens módszert!
2. Ismételje meg az előző feladatot az $\alpha_k = h$ lépésközt használva!
3. Számítsa ki a (8.6) képlettel definiált ϕ_k függvény deriváltját! A derivált $t = \alpha_k$ pontbeli értékéből vezesse le, hogy a $\mathbf{p}^{(k+2)} - \mathbf{p}^{(k+1)}$ és $\mathbf{p}^{(k+1)} - \mathbf{p}^{(k)}$ vektorok merőlegesek egymásra! Magyarázza meg, hogy a numerikus módszerrel generált 8.7. ábrán a jobb oldali sorozat első és második lépése miért nem merőleges egymásra!

8.5. Lineáris egyenletrendszerek megoldása gradiens módszerrel

Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ szimmetrikus mátrix, $\mathbf{b} \in \mathbb{R}^n$, $c \in \mathbb{R}$, és tekintsük a

$$g\colon \mathbb{R}^n \to \mathbb{R}, \qquad g(\mathbf{x}) := \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} + c \tag{8.8}$$

alakú kvadratikus függvényt. Az $\mathbf{A} = (a_{ij})$, $\mathbf{x} = (x_1, \dots, x_n)^T$, $\mathbf{b} = (b_1, \dots, b_n)^T$ jelöléseket használva felírhatjuk g-t a következő alakban:

$$g(x_1, \dots, x_n) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j - \sum_{i=1}^{n} b_ix_i + c.$$

Számítsuk ki a $\frac{\partial g}{\partial x_i}$ parciális deriváltat. A feltevés szerint $a_{ij} = a_{ji}$, ezért

$$\frac{\partial g}{\partial x_i}(x_1, \dots, x_n) = \frac{1}{2}\sum_{j=1}^{n}(a_{ij}x_j + a_{ji}x_j) - b_i = \sum_{j=1}^{n} a_{ij}x_j - b_i,$$

azaz vektoriális alakban

$$g'(\mathbf{x}) = \left(\frac{\partial g}{\partial x_1}(\mathbf{x}), \dots, \frac{\partial g}{\partial x_n}(\mathbf{x})\right)^T = \mathbf{A}\mathbf{x} - \mathbf{b}. \tag{8.9}$$

Így ha $\mathbf{A}$ invertálható, akkor g-nek pontosan egy kritikus pontja van, amely az $\mathbf{A}\mathbf{x} = \mathbf{b}$ egyenlet megoldása. Legyen $\bar{\mathbf{x}}$ a g függvény kritikus pontja és $\mathbf{x} = \bar{\mathbf{x}} + \Delta\mathbf{x}$.

$$\begin{aligned} g(\bar{\mathbf{x}} + \Delta\mathbf{x}) &= \frac{1}{2}(\bar{\mathbf{x}} + \Delta\mathbf{x})^T\mathbf{A}(\bar{\mathbf{x}} + \Delta\mathbf{x}) - \mathbf{b}^T(\bar{\mathbf{x}} + \Delta\mathbf{x}) + c \\ &= \frac{1}{2}\bar{\mathbf{x}}^T\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\bar{\mathbf{x}}^T\mathbf{A}\Delta\mathbf{x} + \frac{1}{2}(\Delta\mathbf{x})^T\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}(\Delta\mathbf{x})^T\mathbf{A}\Delta\mathbf{x} \\ &\quad - \mathbf{b}^T\bar{\mathbf{x}} - \mathbf{b}^T\Delta\mathbf{x} + c. \end{aligned}$$

Ebből kapjuk az $\mathbf{A} = \mathbf{A}^T$, $\bar{\mathbf{x}}^T\mathbf{A}\Delta\mathbf{x} = (\Delta\mathbf{x})^T\mathbf{A}\bar{\mathbf{x}}$, $\mathbf{b}^T\Delta\mathbf{x} = (\Delta\mathbf{x})^T\mathbf{b}$ és az $\mathbf{A}\bar{\mathbf{x}} = \mathbf{b}$ összefüggéseket felhasználva, hogy

$$\begin{aligned} g(\bar{\mathbf{x}} + \Delta\mathbf{x}) &= \frac{1}{2}\bar{\mathbf{x}}^T\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}^T\bar{\mathbf{x}} + (\Delta\mathbf{x})^T(\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}) + \frac{1}{2}(\Delta\mathbf{x})^T\mathbf{A}\Delta\mathbf{x} + c \\ &= g(\bar{\mathbf{x}}) + \frac{1}{2}(\Delta\mathbf{x})^T\mathbf{A}\Delta\mathbf{x}. \end{aligned}$$

Ezért

$$g(\bar{\mathbf{x}} + \Delta\mathbf{x}) - g(\bar{\mathbf{x}}) = \frac{1}{2}(\Delta\mathbf{x})^T\mathbf{A}\Delta\mathbf{x}. \tag{8.10}$$

Ha $\mathbf{A}$ pozitív definit mátrix, akkor $g(\bar{\mathbf{x}} + \Delta\mathbf{x}) - g(\bar{\mathbf{x}}) > 0$ minden $\Delta\mathbf{x} \neq \mathbf{0}$ vektorra, azaz $\bar{x}$ minimalizálja a g függvényt. Ehhez hasonlóan, ha $\mathbf{A}$ negatív definit, akkor a (8.10) egyenletből következik, hogy g-nek maximuma van $\bar{x}$-ben. Pozitív ill. negatív definit mátrixok a 3.9. tétel szerint invertálhatók. Ezzel beláttuk tehát a következő tételt:

8.10. tétel. *Legyen $\mathbf{A}$ szimmetrikus. Ekkor a $g(\mathbf{x}) := \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} + c$ kvadratikus függvény gradiensvektora $g'(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$. Ha $\mathbf{A}$ pozitív (negatív) definit, akkor g-nek létezik globális minimuma (maximuma), amelyet a függvény az $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ pontban vesz fel.*

Az előző tétel bizonyításából könnyen belátható:

8.11. következmény. *Ha egy kvadratikus függvénynek egy pontban lokális minimuma (maximuma) van, akkor ott a függvénynek globális minimuma (maximuma) is van.*

Ha $\mathbf{A}$ egy szimmetrikus pozitív definit mátrix, akkor a 8.10. tétel szerint az $\mathbf{Ax} = \mathbf{b}$ lineáris egyenletrendszert megoldhatjuk úgy, hogy definiáljuk a g kvadratikus függvényt a (8.8) képlettel, és optimális gradiens módszerrel minimalizáljuk azt. Definiáljuk tehát a

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} - \alpha_k \mathbf{v}^{(k)}$$

sorozatot, ahol

$$\mathbf{v}^{(k)} = g'(\mathbf{p}^{(k)}) = \mathbf{A}\mathbf{p}^{(k)} - \mathbf{b}.$$

α_k-t az optimális gradiens módszer definíciójának megfelelően a $\phi_k(t) = g(\mathbf{p}^{(k)} - t\mathbf{v}^{(k)})$ egyváltozós függvény minimumhelyének választjuk. Az ϕ_k függvény egy másodfokú polinom, hiszen

$$\begin{aligned}\phi_k(t) &= \frac{1}{2}\left(\mathbf{p}^{(k)} - t\mathbf{v}^{(k)}\right)^T \mathbf{A}\left(\mathbf{p}^{(k)} - t\mathbf{v}^{(k)}\right) - \mathbf{b}^T\left(\mathbf{p}^{(k)} - t\mathbf{v}^{(k)}\right) + c \\ &= t^2\frac{1}{2}\left(\mathbf{v}^{(k)}\right)^T \mathbf{A}\mathbf{v}^{(k)} - t\left(\mathbf{v}^{(k)}\right)^T(\mathbf{A}\mathbf{p}^{(k)} - \mathbf{b}) + c - \mathbf{b}^T\mathbf{p}^{(k)}.\end{aligned}$$

Ezért ϕ_k minimumhelyét explicit módon meg tudjuk adni:

$$\alpha_k = \frac{\left(\mathbf{v}^{(k)}\right)^T(\mathbf{A}\mathbf{p}^{(k)} - \mathbf{b})}{\left(\mathbf{v}^{(k)}\right)^T\mathbf{A}\mathbf{v}^{(k)}}.$$

Ha bevezetjük az $\mathbf{r}^{(k)} := \mathbf{b} - \mathbf{A}\mathbf{p}^{(k)}$ reziduális vektort, akkor az előbbi képleteket összefoglalhatjuk a következő alakban:

$$\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{p}^{(k)} \tag{8.11}$$

$$\alpha_k = \frac{\left(\mathbf{r}^{(k)}\right)^T\mathbf{r}^{(k)}}{\left(\mathbf{r}^{(k)}\right)^T\mathbf{A}\mathbf{r}^{(k)}} \tag{8.12}$$

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} + \alpha_k\mathbf{r}^{(k)}. \tag{8.13}$$

8.12. példa. A

$$\begin{array}{rcrcrcl} 4x_1 & + & 2x_2 & - & x_3 & = & 0 \\ 2x_1 & + & 5x_2 & & & = & 8 \\ -x_1 & & & + & 3x_3 & = & 1. \end{array}$$

lineáris egyenletrendszerre alkalmaztuk a gradiens módszert a (8.11)-(8.13) rekurzív képletekkel a $\mathbf{p}^{(0)} = (3,3,3)^T$ kezdőértékből kiindulva. Megjegyezzük, hogy a módszer alkalmazható, hiszen a lineáris rendszer együtthatómátrixa szimmetrikus és pozitív definit. A kapott $\mathbf{p}^{(k)}$ sorozat első 13 tagját a 8.4. táblázatban soroltuk fel a közelítés hibájával együtt. Megjegyezzük, hogy a pontos megoldás $(-1, 2, 0)$. □

Feladatok

1. Mutassa meg, hogy tetszőleges

$$g(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{a}_{ij}x_ix_j + \sum_{i=1}^{n}\tilde{b}_ix_i + c$$

kvadratikus függvény felírható (8.8) alakban! Hogy írhatjuk fel $g'(\mathbf{x})$-et és $g''(\mathbf{x})$-et mátrix jelölést használva?

8.4. táblázat. Lineáris egyenletrendszer megoldása gradiens módszerrel

k	$\mathbf{p}^{(k)}$	$\|\mathbf{p}^{(k)}-\mathbf{p}\|_2$
0	(3.00000000, 3.00000000, 3.00000000)	5.09901951
1	(0.43469388, 0.77673469, 2.14489796)	2.85575065
2	(0.03799038, 1.89938726, 0.41611180)	1.12280719
3	(-0.59954375, 1.61568290, 0.37817223)	0.67162421
4	(-0.75093609, 1.98854968, 0.13393796)	0.28302529
5	(-0.90321440, 1.90857051, 0.10622765)	0.17032651
6	(-0.93575911, 1.99605148, 0.03257991)	0.07213829
7	(-0.97504377, 1.97631917, 0.02650106)	0.04342696
8	(-0.98365956, 1.99904876, 0.00839916)	0.01839730
9	(-0.99365117, 1.99398134, 0.00679190)	0.01107528
10	(-0.99583018, 1.99975420, 0.00213698)	0.00469196
11	(-0.99837993, 1.99846385, 0.00173029)	0.00282459
12	(-0.99893668, 1.99993749, 0.00054530)	0.00119662
13	(-0.99958687, 1.99960829, 0.00044139)	0.00072037

2. Igazolja a 8.11. következményt!
3. Ellenőrizze a (8.11)-(8.13) képletek levezetését!
4. Alkalmazza a gradiens módszert a következő kvadratikus függvények minimumhelyének meghatározására:

$$\text{(a)}\quad f(x,y)=2x^2-12x+3y^2+30y, \qquad \text{(b)}\quad f(x,y)=2x^2-4xy+3y^2-2y$$

5. Oldja meg a következő lineáris egyenletrendszereket gradiens módszerrel:

$$\text{(a)}\quad \begin{aligned} 4x_1 &- 3x_2 = 4\\ -3x_1 &+ 3x_2 = 3 \end{aligned} \qquad\qquad \text{(b)}\quad \begin{aligned} 6x_1 &+ 3x_2 - 2x_3 = 6\\ 3x_1 &+ 5x_2 - x_3 = -4\\ -2x_1 &- x_2 + 3x_3 = -2 \end{aligned}$$

6. Legyen $f(x,y)=\frac{1}{2}x^2+\frac{9}{2}y^2$. Igazolja, hogy a gradiens módszert alkalmazva a $\mathbf{p}^{(0)}=(9,1)^T$ pontból indulva a

$$\mathbf{p}^{(k)}=\begin{pmatrix} 9 \\ (-1)^k \end{pmatrix} 0.8^k$$

pontokat kapjuk! Mi a sorozat aszimptotikus hibakonstansa? Adjon meg egy olyan függvényt, ahol a gradiens módszer sorozatának aszimptotikus hibakonstansa egy előre megadott $0<\alpha<1$ szám!

8.6. Newton-módszer

Most tekintsünk egy $f\colon \mathbb{R}^n\to\mathbb{R}$ függvényt. Rögzítsünk egy $\mathbf{p}^{(0)}$ vektort. Ha $f\in C^3$, akkor $\mathbf{p}^{(0)}$ egy környezetében f közelíthető a

$$g(\mathbf{x}):=f(\mathbf{p}^{(0)})+f'(\mathbf{p}^{(0)})^T(\mathbf{x}-\mathbf{p}^{(0)})+\frac{1}{2}(\mathbf{x}-\mathbf{p}^{(0)})^T f''(\mathbf{p}^{(0)})(\mathbf{x}-\mathbf{p}^{(0)}) \tag{8.14}$$

másodfokú Taylor-polinomjával, ahol $f'(\mathbf{p}^{(0)})$ f gradiensvektora, $f''(\mathbf{p}^{(0)})$ pedig f Hesse-mátrixa $\mathbf{p}^{(0)}$-ban. Tegyük fel, hogy $f''(\mathbf{p}^{(0)})$ pozitív definit. Ekkor a 8.10. tétel szerint g-nek globális minimuma létezik, amelyet a

$$\mathbf{p}^{(1)}=\mathbf{p}^{(0)}-\left(f''(\mathbf{p}^{(0)})\right)^{-1}f'(\mathbf{p}^{(0)})$$

pontban vesz fel. Ekkor $\mathbf{p}^{(1)}$-et tekinthetjük f minimumhelye közelítésének. Ezután megismételjük az eljárást a $\mathbf{p}^{(1)}$ pontbeli Taylor-közelítést használva. Így definiálhatjuk a következő iterációs módszert:

$$\mathbf{p}^{(k+1)}=\mathbf{p}^{(k)}-\left(f''(\mathbf{p}^{(k)})\right)^{-1}f'(\mathbf{p}^{(k)}) \tag{8.15}$$

A (8.15) iterációs módszert *Newton-féle minimumkeresési módszernek* hívjuk. Könnyen látható, hogy ez azonos az $f'(\mathbf{x}) = \mathbf{0}$ egyenletrendszer megoldására felírt Newton-iterációval. Ebből kapjuk rögtön a következő tételt.

8.13. tétel. *Legyen $f\colon \mathbb{R}^n \to \mathbb{R}$, $f \in C^3$, $f'(\mathbf{p}) = \mathbf{0}$ és $\mathbf{f}''(\mathbf{p})$ pozitív definit. Ekkor f-nek $\mathbf{p}$-ben lokális minimuma van, és a (8.15) Newton-iteráció lokálisan kvadratikusan konvergál $\mathbf{p}$-hez.*

Bizonyítás. A 8.1. tételt alkalmazva kapjuk, hogy $\mathbf{p}$-ben f-nek lokális minimuma van. Mivel a (8.15) iteráció ekvivalens az $f'(\mathbf{x}) = \mathbf{0}$ egyenlet $\mathbf{p}$ gyökének keresésére felírt Newton-módszerrel, ezért a 2.56. tételből kapjuk, hogy a (8.15) iteráció lokálisan kvadratikusan konvergál $\mathbf{p}$-hez. □

8.14. példa. Alkalmazzuk a Newton-módszert a 8.6., 8.7. és 8.9. példákban vizsgált $f(x,y) = (x^2 - 2y)^2 + 2(x-1)^2$ függvényre. A $(-1,4)^T$ pontból indított (8.15) iteráció első 5 tagját a 8.5. táblázatban tüntettük fel. A sorozat igen gyorsan megközelítette a pontos $(1,0.5)^T$ minimumhelyet. Megjegyezzük, hogy az $(1,3)^T$ pontból indított Newton-sorozat egy lépésben már a pontos minimumhelyet adja vissza. □

8.5. táblázat. Newton-módszer, $f(x,y) = (x^2 - 2y)^2 + 2(x-1)^2$

k	$\mathbf{p}^{(k)}$	$f(\mathbf{p}^{(k)})$	$\|\mathbf{p}^{(k)} - \mathbf{p}\|_2$	$\frac{\|\mathbf{p}^{(k)}-\mathbf{p}\|_2}{\|\mathbf{p}^{(k-1)}-\mathbf{p}\|_2^2}$
0	(-1.00000000, 4.00000000)	57.00000000	4.03112887	
1	(-1.33333333, 0.83333333)	10.90123457	2.35702260	0.14504754
2	(0.76666667,-1.91111111)	19.55698889	2.42237512	0.43602752
3	(0.80979667, 0.32695523)	0.07235807	0.25714159	0.04382173
4	(0.99964684, 0.48162536)	0.00129935	0.01837803	0.27794212
5	(0.99998771, 0.49998766)	0.00000000	0.00001742	0.05156519

8.15. példa. Tekintsük az $f(x,y) = 0.1(x^2 - 2y)^4 + (x-1)^2$ függvényt. Könnyű látni, hogy ennek a függvénynek is $(1,0.5)^T$ a minimumhelye. Ellenőrizhető, hogy a minimumpontban a függvény Hasse mátrixa $f''(1,0.5) = \mathbf{0}$, ami nem pozitív definit. Ennek ellenére a Newton-módszer a $(-1,4)^T$ kezdőértékből indítva konvergens lesz (lásd a 8.6. táblázatot), csak a konvergencia sebessége lineáris lesz. □

Feladatok

1. Alkalmazza a Newton-féle minimumkeresési módszert a 8.3. szakasz 1. feladatában felsorolt függvényekre!
2. Mutassa meg, hogy olyan kvadratikus függvényekre, melyeknek Hesse-mátrixa pozitív definit, a Newton-módszer egy lépésben a pontos minimumhelyet adja vissza!
3. Igazolja, hogy ha a 8.13. tétel feltételei teljesülnek, és ha $\mathbf{p}^{(0)}$ elegendően közel van $\mathbf{p}$-hez, akkor a (8.15) sorozat minden k-ra definiálható, azaz $f''(\mathbf{p}^{(k)})$ invertálható!

8.6. táblázat. Newton-módszer, $f(x,y) = 0.1(x^2-2y)^4 + (x-1)^2$

k	$\mathbf{p}^{(k)}$	$f(\mathbf{p}^{(k)})$	$\|\mathbf{p}^{(k)}-\mathbf{p}\|_2$	$\frac{\|\mathbf{p}^{(k)}-\mathbf{p}\|_2}{\|\mathbf{p}^{(k-1)}-\mathbf{p}\|_2}$
0	(-1.00000000, 4.00000000)	244.10000000	4.03112887	
1	(-1.01468429, 2.84801762)	51.47734819	3.09388745	0.76749902
2	(-1.06550085, 2.12183854)	13.60182932	2.62614813	0.84881825
3	(-1.25304590, 1.80360379)	6.79822461	2.60299802	0.99118476
4	(-2.19917836, 2.64963726)	10.23933318	3.85430701	1.48071838
5	(1.13216300,-4.75372475)	1355.09401353	5.25538684	1.36351018
6	(1.13190045,-2.95581491)	267.68684927	3.45833116	0.65805454
7	(1.13102026,-1.75800646)	52.89017856	2.26180447	0.65401616
8	(1.12811546,-0.96208855)	10.46057564	1.46769088	0.64890263
9	(1.11900871,-0.43955842)	2.07752857	0.94706552	0.64527588
10	(1.09458417,-0.11167347)	0.41720946	0.61894313	0.65353781
11	(1.05056809, 0.07705747)	0.08386326	0.42595483	0.68819704
12	(1.01290080, 0.19574848)	0.01637137	0.30452490	0.71492300
13	(1.00119582, 0.28963767)	0.00320655	0.21036572	0.69079974
14	(1.00003517, 0.35899525)	0.00063312	0.14100475	0.67028386
15	(1.00000031, 0.40597370)	0.00012506	0.09402630	0.66683071
16	(1.00000000, 0.43731559)	0.00002470	0.06268441	0.66666888
17	(1.00000000, 0.45821040)	0.00000488	0.04178960	0.66666668
18	(1.00000000, 0.47214026)	0.00000096	0.02785974	0.66666667
19	(1.00000000, 0.48142684)	0.00000019	0.01857316	0.66666667
20	(1.00000000, 0.48761789)	0.00000004	0.01238211	0.66666667

8.7. Kvázi-Newton módszerek

Az előző szakaszhoz hasonlóan közelítsük az f függvényt egy $\mathbf{p}^{(k)}$ pontja környezetében a

$$g(\mathbf{x}) := f(\mathbf{p}^{(k)}) + \left(\mathbf{v}^{(k)}\right)^T (\mathbf{x}-\mathbf{p}^{(k)}) + \frac{1}{2}(\mathbf{x}-\mathbf{p}^{(k)})^T \mathbf{A}^{(k)} (\mathbf{x}-\mathbf{p}^{(k)}) \tag{8.16}$$

kvadratikus függvénnyel. Ha $\mathbf{v}^{(k)} \approx f'(\mathbf{p}^{(k)})$ és $\mathbf{A}^{(k)} \approx f''(\mathbf{p}^{(k)})$, akkor (8.16) közelíti f másodfokú $\mathbf{p}^{(k)}$-körüli Taylor-polinomját, így valóban f közelítésének tekinthető $\mathbf{p}^{(k)}$ egy kis környezetében. Azt várjuk, hogy g minimumhelye közelíteni fogja f minimumhelyét. Ha $\mathbf{A}^{(k)}$ pozitív definit, akkor a 8.10. tétel szerint g minimumhelye a

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} - \left(\mathbf{A}^{(k)}\right)^{-1} \mathbf{v}^{(k)}. \tag{8.17}$$

pontban van. Ezeket az iterációs eljárásokat *kvázi-Newton minimumkeresési módszereknek* hívjuk.

Választhatjuk $\mathbf{A}^{(k)}$-t és $\mathbf{v}^{(k)}$-t az $f''(\mathbf{p}^{(k)})$ Hesse-mátrix és az $f'(\mathbf{p}^{(k)})$ gradiensvektor numerikus közelítésének: $\mathbf{A}^{(k)} = (a_{ij}^{(k)})$ és $\mathbf{v}^{(k)} = (v_1^{(k)}, \ldots, v_n^{(k)})^T$, ahol

$$a_{ij}^{(k)} = \frac{1}{h^2}\left(f(\mathbf{p}^{(k)} + h\mathbf{e}^{(i)} + h\mathbf{e}^{(j)}) - f(\mathbf{p}^{(k)} + h\mathbf{e}^{(i)}) - f(\mathbf{p}^{(k)} + h\mathbf{e}^{(j)}) + f(\mathbf{p}^{(k)})\right) \tag{8.18}$$

és

$$v_i^{(k)} = \frac{1}{h}\Big(f(\mathbf{p}^{(k)} + h\mathbf{e}^{(i)}) - f(\mathbf{p}^{(k)})\Big),$$

$i, j = 1, \ldots, n$ ($\mathbf{e}^{(i)}$ az i-edik egységvektor, $h > 0$ rögzített kis lépésköz). Itt elsőrendű jobb oldali differencia képlettel közelítettük f elsőrendű parciális deriváltjait, illetve a (7.18)–(7.19) képletekkel a másodrendű parciális deriváltakat. Ezzel a módosítással nincs szükség a pontos Jacobi- és Hesse-mátrix ismeretére, viszont minden iterációs lépésben n^2 nagyságrendű függvény kiértékelést kell elvégezni, arról nem is beszélve, hogy nem tudjuk, mi a h lépésköz ideális választása.

Most tekintsük azt az esetet, amikor a (8.17) képletben $\mathbf{v}^{(k)} = f'(\mathbf{p}^{(k)})$, azaz vizsgáljuk a

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} - \left(\mathbf{A}^{(k)}\right)^{-1} f'(\mathbf{p}^{(k)}) \tag{8.19}$$

alakú kvázi-Newton módszereket. Feltesszük tehát, hogy a függvény gradiensvektorát ki tudjuk számítani, és a kérdés az, hogyan közelítsük a függvény Hesse-mátrixát. Erre egy lehetőség a 2.13. szakaszban vizsgált Broyden-módszer alkalmazása az $f'(\mathbf{x}) = \mathbf{0}$ egyenletrendszer gyökének meghatározására:

$$\mathbf{A}^{(k)}\mathbf{s}^{(k)} = -f'(\mathbf{p}^{(k)}), \tag{8.20}$$

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} + \mathbf{s}^{(k)}, \tag{8.21}$$

$$\mathbf{y}^{(k)} = f'(\mathbf{p}^{(k+1)}) - f'(\mathbf{p}^{(k)}), \tag{8.22}$$

$$\mathbf{A}^{(k+1)} = \mathbf{A}^{(k)} + \frac{(\mathbf{y}^{(k)} - \mathbf{A}^{(k)}\mathbf{s}^{(k)})(\mathbf{s}^{(k)})^T}{\|\mathbf{s}^{(k)}\|_2^2}. \tag{8.23}$$

8.16. példa. Alkalmazzuk a (8.20)–(8.23) képletekkel definiált Broyden-módszert az $f(x,y) = (x^2 - 2y)^2 + 2(x-1)^2$ függvényre. A $(2,2)^T$ pontból indítottuk a sorozatot, az $\mathbf{A}^{(0)}$ mátrix pedig az $f''(2,2)$ Hesse-mátrix $h = 0.05$ lépésközű (8.18) másodrendű differencia képlettel számított közelítése volt. A kapott sorozat első 10 tagját a 8.7. táblázatban láthatjuk. □

8.7. táblázat. Broyden-módszer, $f(x,y) = (x^2 - 2y)^2 + 2(x-1)^2$

k	$\mathbf{p}^{(k)}$	$f(\mathbf{p}^{(k)})$	$\|\mathbf{p}^{(k)} - \mathbf{p}\|_2$	$\frac{\|\mathbf{p}^{(k)}-\mathbf{p}\|_2}{\|\mathbf{p}^{(k-1)}-\mathbf{p}\|_2}$
0	(2.00000000, 2.00000000)	2.00000e+00	1.80277564	
1	(1.28952043, 0.56127886)	4.59574e-01	0.29593441	0.16415488
2	(1.35039835, 0.89916410)	2.46195e-01	0.53114121	1.79479368
3	(1.24875073, 0.73204681)	1.32833e-01	0.34018032	0.64047058
4	(1.12570322, 0.59780553)	3.67287e-02	0.15927091	0.46819553
5	(1.05911935, 0.54518730)	7.97359e-03	0.07441095	0.46719737
6	(0.99939685, 0.49649610)	3.43894e-05	0.00355544	0.04778109
7	(1.01133354, 0.50962433)	2.69479e-04	0.01486866	4.18194987
8	(1.00464762, 0.50384065)	4.58758e-05	0.00602918	0.40549562
9	(1.00047293, 0.50036811)	4.91375e-07	0.00059931	0.09940111
10	(1.00008014, 0.50006497)	1.37638e-08	0.00010316	0.17213595

A (8.23) iterációs módszerrel az a probléma, hogy mivel $\mathbf{A}^{(k)}$ az $f''(\mathbf{p})$ Hesse-mátrix közelítése, így természetes megkövetelni, hogy $\mathbf{A}^{(k)}$ pozitív definit legyen minden k-ra. Ez ahhoz is kell, hogy a (8.16) kvadratikus függvénynek legyen minimuma minden k-ra. A numerikus tapasztalat is azt támasztja alá, hogy azok a (8.19) alakú kvázi-Newton módszerek a leghatékonyabbak, ahol $\mathbf{A}^{(k)}$ pozitív definit közelítése a Hesse-mátrixnak. A Broyden-módszerrel generált $\mathbf{A}^{(k)}$ mátrixsorozat viszont pozitív definit mátrixból kiindulva még csak nem is szimmetrikus mátrixokat generál.

Az 5.6. tétel szerint ha egy $\mathbf{A}$ mátrix pozitív definit, akkor az $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ Cholesky-felbontása létezik, ahol $\mathbf{L}$ nemszinguláris. Fordítva, ha $\mathbf{A} = \mathbf{M}\mathbf{M}^T$ alakú, ahol $\mathbf{M}$ nemszinguláris, akkor $\mathbf{A}$ pozitív definit, hiszen $\mathbf{x}^T\mathbf{M}\mathbf{M}^T\mathbf{x} = \|\mathbf{M}^T\mathbf{x}\|_2^2 \geq 0$, és egyenlőség csak akkor van, ha $\mathbf{M}^T\mathbf{x} = \mathbf{0}$, és ezért $\mathbf{x} = \mathbf{0}$.

Legyen $\mathbf{A}^{(k)} = \mathbf{M}^{(k)}(\mathbf{M}^{(k)})^T$ alakú, ahol $\mathbf{M}^{(k)}$ invertálható (de nem feltétlenül alulról háromszögmátrix). A következő Hesse-mátrix közelítést, $\mathbf{A}^{(k+1)}$-et $\mathbf{A}^{(k+1)} = \mathbf{M}^{(k+1)}(\mathbf{M}^{(k+1)})^T$ alakban keressük, ahol $\mathbf{A}^{(k+1)}$-től megköveteljük, hogy teljesítse az $\mathbf{A}^{(k+1)}\mathbf{s}^{(k)} = \mathbf{y}^{(k)}$ szelő egyenleteket. A szelő egyenletből következik, hogy $(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)} = (\mathbf{s}^{(k)})^T\mathbf{A}^{(k+1)}\mathbf{s}^{(k)}$, ezért ha $\mathbf{A}^{(k+1)}$ pozitív definit, akkor az

$$(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)} > 0 \tag{8.24}$$

egyenlőtlenség teljesül. Megmutatjuk, hogy (8.24) teljesülése esetén a szelő egyenletnek van pozitív definit megoldása.

Vezessük be a $\mathbf{v}^{(k)} := (\mathbf{M}^{(k+1)})^T\mathbf{s}^{(k)}$ jelölést. Ekkor a szelő egyenlet felírható a következőképpen:

$$(\mathbf{M}^{(k+1)})^T\mathbf{s}^{(k)} = \mathbf{v}^{(k)}, \tag{8.25}$$

$$\mathbf{M}^{(k+1)}\mathbf{v}^{(k)} = \mathbf{y}^{(k)}. \tag{8.26}$$

Az $\mathbf{M}^{(k+1)}$ mátrixot az $\mathbf{M}^{(k)}$ mátrixot módosítva szeretnénk előállítani, ezért a Broyden-módszer levezetését követve (8.26) alapján természetes $\mathbf{M}^{(k+1)}$-et az

$$\mathbf{M}^{(k+1)} = \mathbf{M}^{(k)} + \frac{(\mathbf{y}^{(k)} - \mathbf{M}^{(k)}\mathbf{v}^{(k)})(\mathbf{v}^{(k)})^T}{\|\mathbf{v}^{(k)}\|_2^2} \tag{8.27}$$

alakban keresni. Ekkor $\mathbf{M}^{(k+1)}$ teljesíti a (8.26) egyenletet, és a legkevésbé tér el $\mathbf{M}^{(k)}$-tól abban az értelemben, hogy minden $\mathbf{z} \perp \mathbf{v}^{(k)}$-ra $\mathbf{M}^{(k+1)}\mathbf{z} = \mathbf{M}^{(k)}\mathbf{z}$. $\mathbf{M}^{(k+1)}$-et visszahelyettesítve a (8.25) egyenletbe kapjuk, hogy

$$\begin{aligned}
\mathbf{v}^{(k)} &= (\mathbf{M}^{(k)})^T\mathbf{s}^{(k)} + \frac{\left((\mathbf{y}^{(k)} - \mathbf{M}^{(k)}\mathbf{v}^{(k)})(\mathbf{v}^{(k)})^T\right)^T}{\|\mathbf{v}^{(k)}\|_2^2}\mathbf{s}^{(k)} \\
&= (\mathbf{M}^{(k)})^T\mathbf{s}^{(k)} + \frac{\mathbf{v}^{(k)}(\mathbf{y}^{(k)} - \mathbf{M}^{(k)}\mathbf{v}^{(k)})^T}{\|\mathbf{v}^{(k)}\|_2^2}\mathbf{s}^{(k)} \\
&= (\mathbf{M}^{(k)})^T\mathbf{s}^{(k)} + \frac{(\mathbf{y}^{(k)} - \mathbf{M}^{(k)}\mathbf{v}^{(k)})^T\mathbf{s}^{(k)}}{\|\mathbf{v}^{(k)}\|_2^2}\mathbf{v}^{(k)}.
\end{aligned}$$

Ebből következik, hogy $(\mathbf{M}^{(k)})^T\mathbf{s}^{(k)} = \alpha\mathbf{v}^{(k)}$ alakú, ahol

$$\begin{aligned}
\alpha &= 1 - \frac{(\mathbf{y}^{(k)} - \mathbf{M}^{(k)}\mathbf{v}^{(k)})^T\mathbf{s}^{(k)}}{\|\mathbf{v}^{(k)}\|_2^2} \\
&= 1 - \frac{(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)}}{\|\mathbf{v}^{(k)}\|_2^2} + \frac{(\mathbf{v}^{(k)})^T(\mathbf{M}^{(k)})^T\mathbf{s}^{(k)}}{\|\mathbf{v}^{(k)}\|_2^2} \\
&= 1 - \alpha^2\frac{(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)}}{(\mathbf{s}^{(k)})^T\mathbf{M}^{(k)}(\mathbf{M}^{(k)})^T\mathbf{s}^{(k)}} + \alpha,
\end{aligned}$$

és így

$$\begin{aligned}
\alpha^2 &= \frac{(\mathbf{s}^{(k)})^T\mathbf{M}^{(k)}(\mathbf{M}^{(k)})^T\mathbf{s}^{(k)}}{(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)}} \\
&= \frac{(\mathbf{s}^{(k)})^T\mathbf{A}^{(k)}\mathbf{s}^{(k)}}{(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)}}.
\end{aligned} \tag{8.28}$$

Mivel a számláló pozitív, hiszen feltettük, hogy $\mathbf{A}^{(k)}$ pozitív definit, ezért α kifejezhető a (8.28) egyenletből, és

$$\mathbf{v}^{(k)} = \frac{1}{\alpha}(\mathbf{M}^{(k)})^T\mathbf{s}^{(k)} = \left(\frac{(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)}}{(\mathbf{s}^{(k)})^T\mathbf{A}^{(k)}\mathbf{s}^{(k)}}\right)^{1/2}(\mathbf{M}^{(k)})^T\mathbf{s}^{(k)}.$$

Ezt visszahelyettesítve a (8.27) egyenletbe

$$\begin{aligned}\mathbf{M}^{(k+1)} &= \mathbf{M}^{(k)} + \frac{(\mathbf{y}^{(k)} - \frac{1}{\alpha}\mathbf{M}^{(k)}(\mathbf{M}^{(k)})^T\mathbf{s}^{(k)})\frac{1}{\alpha}(\mathbf{s}^{(k)})^T\mathbf{M}^{(k)}}{\frac{1}{\alpha^2}\|(\mathbf{M}^{(k)})^T\mathbf{s}^{(k)}\|_2^2} \\ &= \mathbf{M}^{(k)} + \alpha\frac{\mathbf{y}^{(k)}(\mathbf{s}^{(k)})^T\mathbf{M}^{(k)}}{(\mathbf{s}^{(k)})^T\mathbf{A}^{(k)}\mathbf{s}^{(k)}} - \frac{\mathbf{A}^{(k)}\mathbf{s}^{(k)}(\mathbf{s}^{(k)})^T\mathbf{M}^{(k)}}{(\mathbf{s}^{(k)})^T\mathbf{A}^{(k)}\mathbf{s}^{(k)}}.\end{aligned}$$

Kis számolással ebből levezethető (2. feladat), hogy

$$\mathbf{A}^{(k+1)} = \mathbf{A}^{(k)} + \frac{\mathbf{y}^{(k)}(\mathbf{y}^{(k)})^T}{(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)}} - \frac{\mathbf{A}^{(k)}\mathbf{s}^{(k)}(\mathbf{s}^{(k)})^T\mathbf{A}^{(k)}}{(\mathbf{s}^{(k)})^T\mathbf{A}^{(k)}\mathbf{s}^{(k)}}. \tag{8.29}$$

Hátra van még azt megmutatni, hogy az iteráció pozitív definit mátrixot generál. Mivel $\mathbf{A}^{(k+1)} = \mathbf{M}^{(k+1)}(\mathbf{M}^{(k+1)})^T$, ezért elegendő azt belátni, hogy $\mathbf{M}^{(k+1)}$ invertálható. A feltevés szerint $\mathbf{M}^{(k)}$ pozitív definit, és ezért invertálható. Ha feltesszük, hogy (8.24) teljesül, akkor $\mathbf{M}^{(k+1)}$ invertálhatóságát könnyen kapjuk a (8.27) képletből a 2.58. tételt alkalmazva. A részletek kidolgozását az olvasóra hagyjuk (3. feladat).

A (8.29) formulát Broyden, Flecher, Goldfarb és Shanno vezették be 1970-ben, ezért *BFGS-iterációnak* nevezzük. Ez a jelenleg ismert legjobb iterációs formula a Hesse-mátrix közelítésére. Az iteráció kezdeti mátrixának vagy $f''(\mathbf{p}^{(0)})$-t vagy ennek egy (8.18) másodrendű differencia közelítését célszerű használni. Ha $\mathbf{p}^{(0)}$ elegendően közel van $\mathbf{p}$-hez, és $f''(\mathbf{p})$ pozitív definit, akkor $f''(\mathbf{p}^{(0)})$, és ezért $\mathbf{A}^{(0)}$ is az lesz.

Végül vizsgáljuk meg, hogy a (8.24) feltétel milyen megszorítást jelent. A Lagrange-féle középértéktételt (2.40. tétel) és a (8.21), (8.22) egyenleteket alkalmazva kapjuk, hogy

$$\begin{aligned}(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)} &= \left(f'(\mathbf{p}^{(k+1)}) - f'(\mathbf{p}^{(k)})\right)^T(\mathbf{p}^{(k+1)} - \mathbf{p}^{(k)}) \\ &= \sum_{i=1}^{n}\left(\frac{\partial f_i(\mathbf{p}^{(k+1)})}{\partial x_i} - \frac{\partial f_i(\mathbf{p}^{(k)})}{\partial x_i}\right)(p_i^{(k+1)} - p_i^{(k)}) \\ &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{\partial^2 f_i(\xi^{(k,i)})}{\partial x_i\,\partial x_j}(p_j^{(k+1)} - p_j^{(k)})\right)(p_i^{(k+1)} - p_i^{(k)}).\end{aligned}$$

Ha a $\mathbf{p}^{(k)}$ iteráltak elegendően közel maradnak $\mathbf{p}$-hez az iteráció közben, akkor $\xi^{(k,i)}$ is $\mathbf{p}$ közelében marad, és ezért f'' folytonossága miatt

$$\begin{aligned}(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)} &\approx \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{\partial^2 f_i(\mathbf{p})}{\partial x_i\,\partial x_j}(p_j^{(k+1)} - p_j^{(k)})\right)(p_i^{(k+1)} - p_i^{(k)}) \\ &= (\mathbf{p}^{(k+1)} - \mathbf{p}^{(k)})^T f''(\mathbf{p})(\mathbf{p}^{(k+1)} - \mathbf{p}^{(k)}),\end{aligned}$$

ami pozitív, hiszen $f''(\mathbf{p})$ pozitív definit. Ez a feltétel tehát, ha a sorozat $\mathbf{p}$-hez tart, $\mathbf{p}$ közelében teljesülni fog. Természetesen ha (8.24) nem teljesül, akkor is definiálható a (8.29) iteráció, csak ekkor $\mathbf{A}^{(k+1)}$ pozitív szemidefinit lesz, nem pozitív definit.

Belátható a következő tétel.

8.17. tétel. *Legyen $f \in C^3$, $f'(\mathbf{p}) = 0$, $f''(\mathbf{p})$ pozitív definit. Ekkor létezik olyan $\varepsilon, \delta > 0$, hogy a (8.20)–(8.22), (8.29) iteráció definiált minden k-ra, és szuperlineárisan konvergál $\mathbf{p}$-hez, ha $\|\mathbf{p}^{(0)} - \mathbf{p}\|_2 < \varepsilon$ és $\|\mathbf{A}^{(0)} - f''(\mathbf{p})\|_2 < \delta$.*

8.8. táblázat. A (8.19) kvázi-Newton módszer BFGS-iterációval

k	$\mathbf{p}^{(k)}$	$f(\mathbf{p}^{(k)})$	$\|\mathbf{p}^{(k)}-\mathbf{p}\|_2$	$\frac{\|\mathbf{p}^{(k)}-\mathbf{p}\|_2}{\|\mathbf{p}^{(k-1)}-\mathbf{p}\|_2}$
0	(2.00000000, 2.00000000)	2.00000e+00	1.80277564	
1	(1.28952043, 0.56127886)	4.59574e-01	0.29593441	0.16415488
2	(1.25102079, 0.70409379)	1.50630e-01	0.32352080	1.09321792
3	(1.19910219, 0.73444653)	8.02473e-02	0.30758228	0.95073416
4	(1.14966546, 0.69907469)	5.06393e-02	0.24905919	0.80973192
5	(1.00399514, 0.50473229)	3.40491e-05	0.00619320	0.02486638
6	(0.99975498, 0.49938607)	6.64526e-07	0.00066102	0.10673251
7	(1.00003118, 0.49997474)	1.46839e-08	0.00004012	0.06070113
8	(1.00001593, 0.50000889)	7.05953e-10	0.00001824	0.45466117
9	(1.00000627, 0.50000724)	8.24492e-11	0.00000958	0.52515860
10	(1.00000015, 0.50000024)	7.49020e-14	0.00000028	0.02901243

8.18. példa. A BFGS-iterációval kaptuk a 8.8. táblázatban szereplő sorozatot az $f(x,y)=(x^2-2y)^2+2(x-1)^2$ függvényre. Ugyanabból a kezdőértékekből indítottuk a módszert, mint a 8.16. példában. □

Teljes indukcióval ellenőrizhető, hogy a BFGS-módszerrel képzett $\mathbf{A}^{(k)}$ mátrixok $\mathbf{B}^{(k)} := (\mathbf{A}^{(k)})^{-1}$ inverzét a

$$\begin{aligned}\mathbf{B}^{(k+1)} &= \mathbf{B}^{(k)} + \left(1+\frac{(\mathbf{y}^{(k)})^T\mathbf{B}^{(k)}\mathbf{y}^{(k)}}{(\mathbf{s}^{(k)})^T\mathbf{y}^{(k)}}\right)\frac{\mathbf{s}^{(k)}(\mathbf{s}^{(k)})^T}{(\mathbf{s}^{(k)})^T\mathbf{y}^{(k)}} \\ &\quad - \frac{\mathbf{s}^{(k)}(\mathbf{y}^{(k)})^T\mathbf{B}^{(k)}+\mathbf{B}^{(k)}\mathbf{y}^{(k)}(\mathbf{s}^{(k)})^T}{(\mathbf{s}^{(k)})^T\mathbf{y}^{(k)}}\end{aligned} \tag{8.30}$$

rekurzív képlettel is kiszámíthatjuk. Ezt az összefüggést használva a (8.20) egyenlet helyettesíthető az

$$\mathbf{s}^{(k)} = -\mathbf{B}^{(k)}f'(\mathbf{p}^{(k)}) \tag{8.31}$$

egyenlettel, és így a módszer alkalmazásakor nincs szükség lineáris egyenletrendszer megoldására vagy mátrix invertálásra.

A BFGS-iteráció levezetéséhez hasonlóan kaphatjuk a DFP-iteráció képletét. Újra $\mathbf{A}^{(k+1)} = \mathbf{M}^{(k+1)}(\mathbf{M}^{(k+1)})^T$ alakban keressük a módosított Hesse-közelítést, de a (8.25)–(8.26) szelő egyenletek helyett most az azzal ekvivalens

$$\begin{aligned}(\mathbf{M}^{(k+1)})^{-1}\mathbf{y}^{(k)} &= \mathbf{v}^{(k)} \\ \left(\mathbf{M}^{(k+1)^T}\right)^{-1}\mathbf{v}^{(k)} &= \mathbf{s}^{(k)}\end{aligned}$$

egyenletekből indulunk ki. Ennek megoldását

$$\left(\mathbf{M}^{(k+1)}\right)^{-1} = \left(\mathbf{M}^{(k)}\right)^{-1} + \frac{\left(\mathbf{s}^{(k)}-(\mathbf{M}^{(k)})^{-1}\mathbf{v}^{(k)}\right)(\mathbf{v}^{(k)})^T}{\|\mathbf{v}^{(k)}\|_2^2}$$

alakban keresve kapjuk, hogy

$$\mathbf{v}^{(k)} = \left(\frac{(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)}}{(\mathbf{y}^{(k)})^T(\mathbf{A}^{(k)})^{-1}\mathbf{y}^{(k)}}\right)^{1/2}(\mathbf{M}^{(k)})^{-1}\mathbf{y}^{(k)},$$

feltéve, hogy a (8.24) teljesül. Ebből a 2.58. tétel alkalmazásával kiszámítható, hogy

$$\mathbf{A}^{(k+1)} = \mathbf{A}^{(k)} + \frac{(\mathbf{y}^{(k)} - \mathbf{A}^{(k)}\mathbf{s}^{(k)})(\mathbf{y}^{(k)})^T + \mathbf{y}^{(k)}(\mathbf{y}^{(k)} - \mathbf{A}^{(k)}\mathbf{s}^{(k)})^T}{(\mathbf{y}^{(k)})^T\mathbf{s}^{(k)}} - \frac{(\mathbf{y}^{(k)} - \mathbf{A}^{(k)}\mathbf{s}^{(k)})^T\mathbf{s}^{(k)}}{((\mathbf{y}^{(k)})^T\mathbf{s}^{(k)})^2}\mathbf{y}^{(k)}(\mathbf{y}^{(k)})^T. \tag{8.32}$$

Ezt a formulát *DFP-iterációnak* nevezzük felfedezői után: Davidon (1959) és Flecher, Powell (1963). Erre az iterációra is teljesül 8.17 tétellel analóg konvergencia eredmény.

Ellenőrizhető, hogy a DFP-iterációval generált $\mathbf{A}^{(k)}$ mátrix inverze kiszámítható a következő rekurzív módon:

$$(\mathbf{A}^{(k+1)})^{-1} = (\mathbf{A}^{(k)})^{-1} + \frac{\mathbf{s}^{(k)}(\mathbf{s}^{(k)})^T}{(\mathbf{s}^{(k)})^T\mathbf{y}^{(k)}} - \frac{(\mathbf{A}^{(k)})^{-1}\mathbf{y}^{(k)}(\mathbf{y}^{(k)})^T(\mathbf{A}^{(k)})^{-1}}{(\mathbf{y}^{(k)})^T(\mathbf{A}^{(k)})^{-1}\mathbf{y}^{(k)}}. \tag{8.33}$$

8.19. példa. A DFP-iterációt vizsgáltuk a 8.16. és 8.18. példák feladatára. Ez a módszer is a BFGS-iterációhoz hasonlóan gyorsan konvergál. A sorozat a 8.9. táblázatban látható. □

8.9. táblázat. A (8.19) kvázi-Newton módszer DFP-iterációval

k	$\mathbf{p}^{(k)}$	$f(\mathbf{p}^{(k)})$	$\|\mathbf{p}^{(k)} - \mathbf{p}\|_2$	$\frac{\|\mathbf{p}^{(k)}-\mathbf{p}\|_2}{\|\mathbf{p}^{(k-1)}-\mathbf{p}\|_2}$
0	(2.00000000, 2.00000000)	2.00000e+00	1.80277564	
1	(1.28952043, 0.56127886)	4.59574e-01	0.29593441	0.16415488
2	(1.25682024, 0.70394625)	1.61396e-01	0.32794924	1.10818219
3	(1.09891338, 0.59229507)	2.00977e-02	0.13528576	0.41252041
4	(1.01148073, 0.50204318)	6.24877e-04	0.01166112	0.08619621
5	(1.00103666, 0.50022718)	4.77384e-06	0.00106126	0.09100838
6	(1.00001771, 0.50001111)	8.01068e-10	0.00002090	0.01969409
7	(0.99999976, 0.49999958)	2.45621e-13	0.00000049	0.02332123
8	(1.00000001, 0.50000002)	4.22000e-16	0.00000002	0.03601757

Feladatok

1. Alkalmazza az ebben a szakaszban bevezetett kvázi-Newton módszereket a 8.3. szakasz 1. feladatában felsorolt függvényekre!
2. Ellenőrizze a (8.29) formula levezetését!
3. Igazolja, hogy $\mathbf{M}^{(k+1)}$ invertálható, ha (8.24) teljesül!
4. Igazolja a (8.30) rekurzív összefüggést!
5. Dolgozza ki a DFP-iteráció levezetésének részleteit!
6. Igazolja a (8.33) rekurzív összefüggést!

9. fejezet

Legkisebb négyzetek módszere

Tegyük fel, hogy egy fizikai folyamatot egy g függvénnyel írhatunk le, amelynek ismerjük vagy feltételezzük az általános képletét, de bizonyos paraméterek a képletben ismeretlenek. A paramétereket egy $\mathbf{a}$ vektorban tárolva a $g(x;\mathbf{a})$ jelöléssel hangsúlyozhatjuk, hogy g az $\mathbf{a}$ paraméterektől függ. Feltesszük, hogy vannak y_i $(i = 0, 1, \ldots, n)$ mérési adataink a g függvényről az x_i alappontokban. Tegyük fel a példa kedvéért, hogy tudjuk vagy sejtjük, hogy g egy másodfokú polinom. Ekkor g-t 3 paraméter, az együtthatói határozzák meg. Ha 3-nál több mérési értékünk van, akkor általában már nem tudunk egy parabolát rajzolni a pontokon keresztül (a mérési hibák miatt az adataink valószínűleg nem a parabola grafikonján helyezkednek el). Ezért a célunk az, hogy keressük meg azokat a paraméter értékeket, amelyhez tartozó g függvény a „legkevésbé" tér el a mérési adatoktól. Ezt a feladatot hívjuk *görbeillesztésnek*. Nem nyilvánvaló, hogy mit értsünk azon, hogy a függvény „legkevésbé" tér el az adatoktól. Attól függően, hogyan definiáljuk az illesztés hibáját, más és más matematikai feladatként fogalmazhatjuk meg a görbeillesztés feladatát. Lehetséges az illesztés hibáját mérni az

$$F_1(\mathbf{a}) := \max\{|g(x_i;\mathbf{a}) - y_i| \colon i = 0, 1, \ldots, n\}$$

vagy az

$$F_2(\mathbf{a}) := \sum_{i=0}^{n} |g(x_i;\mathbf{a}) - y_i|$$

képletekkel. Mindkettőt természetes választásnak érezhetjük, hiszen ha a képlet értéke kis szám, akkor a $g(x_i)$ függvényérték és az y_i mérési érték eltérése is kicsi lesz minden pontban. A probléma az, hogy ha $F_1(\mathbf{a})$-t ill. $F_2(\mathbf{a})$-t szeretnénk minimalizálni $\mathbf{a}$ szerint, akkor ez matematikailag nehéz feladat amiatt, hogy egyik függvény sem differenciálható $\mathbf{a}$ szerint. Ezt a technikai problémát kiküszöbölhetjük azzal, ha az

$$F(\mathbf{a}) := \sum_{i=0}^{n} (g(x_i;\mathbf{a}) - y_i)^2,$$

ún. négyzetes hibával mérjük a függvény és a mérési adatok eltérését. A matematikai feladat tehát az, hogy minimalizáljuk az $F(\mathbf{a})$ függvényt, és a minimumhelyhez tartozó paraméter értékekkel definiált $g(x;\mathbf{a})$ függvényt tekintjük a pontokra legjobban illeszkedő adott típusú függvénynek. Ezt a módszert hívjuk a *legkisebb négyzetek módszerének*.

A négyzetes hiba segítségével történő görbeillesztést tanulmányozzuk ebben a fejezetben. Először lineáris függvény, majd tetszőleges polinom, és végül néhány speciális nemlineáris függvény és trigonometrikus polinom illesztésével foglalkozunk. A fejezet végén rosszul definiált lineáris egyenletrendszerek legkisebb négyzetes megoldását vizsgáljuk.

9.1. Egyenes illesztése

Adottak (x_i, y_i), $i = 0, 1, \ldots, n$ pontok, ahol x_i-k páronként különböznek. Keresünk egy olyan $g(x) = ax + b$ lineáris függvényt, amelynek az adatoktól számított négyzetes eltérése, azaz

$$F(a,b) := \sum_{i=0}^{n} (ax_i + b - y_i)^2 \tag{9.1}$$

minimális. Az így definiált F függvény folytonosan parciálisan differenciálható a és b szerint, és

$$\begin{aligned} \frac{\partial F}{\partial a}(a,b) &= 2\sum_{i=0}^{n} (ax_i + b - y_i)x_i, \\ \frac{\partial F}{\partial b}(a,b) &= 2\sum_{i=0}^{n} (ax_i + b - y_i). \end{aligned} \tag{9.2}$$

A (9.2) parciális deriváltakat 0-val egyenlővé téve, átrendezés után kapjuk az ún. *Gauss-féle normálegyenleteket.*

$$\begin{aligned} a\sum_{i=0}^{n} x_i^2 + b\sum_{i=0}^{n} x_i &= \sum_{i=0}^{n} x_i y_i, \\ a\sum_{i=0}^{n} x_i + b(n+1) &= \sum_{i=0}^{n} y_i. \end{aligned} \tag{9.3}$$

Érdemes hangsúlyozni, hogy a második egyenletben b együtthatója, $n+1$ az adott mérési adatok számát adja vissza. Ez egy lineáris egyenletrendszer a-ra és b-re. Az egyenletrendszer akkor és csak akkor oldható meg, ha az együtthatómátrix determinánsa,

$$d := \det \begin{pmatrix} \sum_{i=0}^{n} x_i^2 & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & n+1 \end{pmatrix} = (n+1)\sum_{i=0}^{n} x_i^2 - \left(\sum_{i=0}^{n} x_i\right)^2$$

nem nulla. A Cauchy-Bunyakovszkij-Schwarz egyenlőtlenség (2.42. tétel) szerint

$$\left(\sum_{i=0}^{n} x_i\right)^2 = \left(\sum_{i=0}^{n} 1 \cdot x_i\right)^2 \le \sum_{i=0}^{n} 1 \sum_{i=0}^{n} x_i^2 = (n+1)\sum_{i=0}^{n} x_i^2.$$

Ebből következik, hogy $d \ge 0$. Ha feltesszük, hogy legalább két x_i különbözik, akkor a 2.42. tétel szerint egyenlőtlenség nem állhat fenn, azaz $d > 0$. Ezért a (9.3) egyenletrendszernek pontosan egy megoldása van, amely a következő alakban adható meg:

$$\begin{aligned} \bar{a} &= \frac{(n+1)\left(\sum_{i=0}^{n} x_i y_i\right) - \left(\sum_{i=0}^{n} x_i\right)\left(\sum_{i=0}^{n} y_i\right)}{(n+1)\left(\sum_{i=0}^{n} x_i^2\right) - \left(\sum_{i=0}^{n} x_i\right)^2}, \\ \bar{b} &= \frac{\left(\sum_{i=0}^{n} x_i^2\right)\left(\sum_{i=0}^{n} y_i\right) - \left(\sum_{i=0}^{n} x_i y_i\right)\left(\sum_{i=0}^{n} x_i\right)}{(n+1)\left(\sum_{i=0}^{n} x_i^2\right) - \left(\sum_{i=0}^{n} x_i\right)^2}. \end{aligned}$$

A 8.2. tétel szerint F-nek az $(\bar{a}, \bar{b})$ pontban lokális szélsőértéke van, ha

$$D(\bar{a}, \bar{b}) = \frac{\partial^2 F}{\partial a^2}(\bar{a}, \bar{b}) \cdot \frac{\partial^2 F}{\partial b^2}(\bar{a}, \bar{b}) - \left(\frac{\partial^2 F}{\partial a\, \partial b}(\bar{a}, \bar{b})\right)^2 > 0.$$

Könnyen kiszámítható, hogy

$$\frac{\partial^2 F}{\partial a^2}(\bar{a},\bar{b}) = 2\sum_{i=0}^{n} x_i^2, \quad \frac{\partial^2 F}{\partial b^2}(\bar{a},\bar{b}) = 2(n+1), \quad \frac{\partial^2 F}{\partial a\,\partial b}(\bar{a},\bar{b}) = 2\sum_{i=0}^{n} x_i.$$

Ezért

$$D(\bar{a},\bar{b}) = 4(n+1)\sum_{i=0}^{n} x_i^2 - 4\left(\sum_{i=0}^{n} x_i\right)^2 = 4d,$$

amiről már megmutattuk, hogy pozitív. Mivel $\frac{\partial^2 F}{\partial a^2}(\bar{a},\bar{b}) > 0$, ezért a 8.2. tételből következik, hogy az F függvénynek lokális minimuma van az $(\bar{a},\bar{b})$ pontban, ami a 8.11. következmény szerint egyben globális minimum is. Ezzel beláttuk a következő tételt:

9.1. tétel. *Adottak az (x_i, y_i) $(i = 0, 1, \ldots, n)$ pontok, ahol van olyan i és j, hogy $x_i \neq x_j$. Ekkor a*

$$\min_{(a,b)\in\mathbb{R}^2} \sum_{i=0}^{n} (ax_i + b - y_i)^2$$

szélsőérték feladatnak létezik egyértelmű megoldása, amely teljesíti a (9.1) normálegyenleteket.

9.2. példa. Tekintsük a következő adatokat:

x_i	-1.0	1.0	2.5	3.0	4.0	4.5	6.0
y_i	0.0	1.2	1.9	2.5	3.1	3.2	4.5

Keressük meg az adatokra legjobban illeszkedő egyenest! Kézi számoláskor írjuk le az adatokat a 9.1. táblázatban látható módon! Külön oszlopban kiszámoljuk az x_i^2 és $x_i y_i$ számokat, ill. az utolsó sorban az oszlopban szereplő számok összegét. Ezen összegeket használjuk a (9.3) normálegyenletek felírásához:

$$\begin{array}{rcrcl} 67.25a & + & 20.0b & = & 67.25 \\ 20.0a & + & 7b & = & 16.4. \end{array}$$

amelynek megoldása $a = 0.630243$ és $b = 0.542163$. A megadott pontok és az $y = 0.630243x + 0.542163$ egyenes grafikonja a 9.1. ábrán látható. Az illesztés hibája:

$$\sum_{i=0}^{6} (0.630243x_i + 0.542163 - y_i)^2 = 0.124691.$$

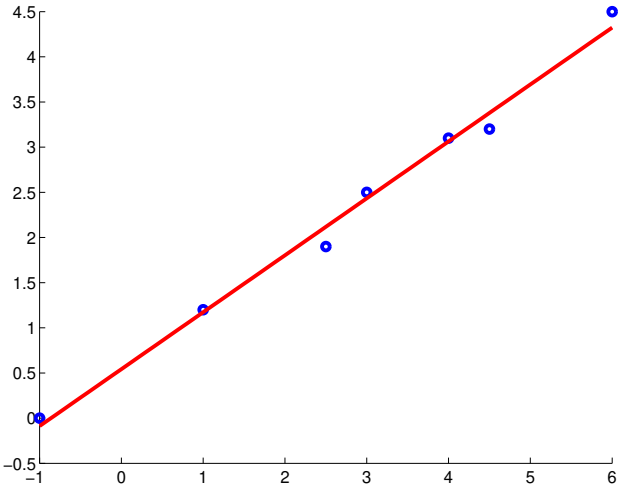
□

9.1. táblázat. Egyenes illesztése

x_i	y_i	x_i^2	$x_i y_i$
-1.0	0.0	1.00	0.00
1.0	1.2	1.00	1.20
2.5	1.9	6.25	4.75
3.0	2.5	9.00	7.50
4.0	3.1	16.00	12.40
4.5	3.2	20.25	14.40
6.0	4.5	36.00	27.00
20.0	16.4	89.50	67.25

Feladatok

1. Illesszen egyenest a megadott adatokra és számítsa ki az illesztés hibáját:



9.1. ábra. Egyenes illesztése: $y = 0.630243x + 0.542163$

(a)

x_i	0.0	1.0	1.5	2.0	3.0
y_i	-1.8	1.3	2.5	3.9	8.3

(b)

x_i	-1.0	1.0	2.0	3.0	4.0	5.0	6.0
y_i	4.2	2.1	1.3	2.1	2.8	-2.1	-3.0

(c)

x_i	-1.0	1.0	3.0	5.0	9.0	10.0	13.0
y_i	-0.1	3.4	7.3	15.1	29.1	35.6	56.3

9.2. Polinom illesztése

Ebben a szakaszban m-edfokú polinom illesztését vizsgáljuk megadott (x_i, y_i) $(i = 0, 1, \ldots, n)$ pontokra, azaz keresünk olyan a_m, a_{m-1}, …, a_0 számokat, amelyek minimalizálják az

$$F(a_m, a_{m-1}, \ldots, a_1, a_0) := \sum_{i=0}^{n} (a_m x_i^m + a_{m-1} x_i^{m-1} + \cdots + a_1 x_i + a_0 - y_i)^2$$

$m+1$-változós függvényt. Ha $n \leq m$, akkor a megadott pontokon keresztül rajzolható m-edfokú polinom (F minimális értéke 0). Ebben az esetben interpolációval meghatározhatók az együtthatók. Így az $m < n$ esetre érdekes vizsgálnunk a feladatot, hiszen ekkor F nem veszi fel a 0 értéket.

A 8.2. tétel alapján az F függvénynek ott lehet csak szélsőértéke, ahol a parciális deriváltjai nullák:

$$\begin{aligned}
\frac{\partial F}{\partial a_m}(a_m, a_{m-1}, \ldots, a_0) &= 2\sum_{i=0}^{m} (a_m x_i^m + a_{m-1} x_i^{m-1} + \cdots + a_0 - y_i) x_i^m, \\
\frac{\partial F}{\partial a_{m-1}}(a_m, a_{m-1}, \ldots, a_0) &= 2\sum_{i=0}^{m} (a_m x_i^m + a_{m-1} x_i^{m-1} + \cdots + a_0 - y_i) x_i^{m-1}, \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
\frac{\partial F}{\partial a_0}(a_m, a_{m-1}, \ldots, a_0) &= 2\sum_{i=0}^{m} (a_m x_i^m + a_{m-1} x_i^{m-1} + \cdots + a_0 - y_i).
\end{aligned}$$

Ezeket nullával egyenlővé téve és átrendezve a kapott egyenleteket

$$\begin{array}{llll}
a_m\sum_{i=0}^n x_i^{2m} & + a_{m-1}\sum_{i=0}^n x_i^{2m-1} + \cdots + a_1\sum_{i=0}^n x_i^{m+1} + a_0\sum_{i=0}^n x_i^m & = \sum_{i=0}^n x_i^m y_i \\
a_m\sum_{i=0}^n x_i^{2m-1} & + a_{m-1}\sum_{i=0}^n x_i^{2m-2} + \cdots + a_1\sum_{i=0}^n x_i^m + a_0\sum_{i=0}^n x_i^{m-1} & = \sum_{i=0}^n x_i^{m-1} y_i \\
\vdots & & \vdots \\
a_m\sum_{i=0}^n x_i^{m+1} & + a_{m-1}\sum_{i=0}^n x_i^m + \cdots + a_1\sum_{i=0}^n x_i^2 + a_0\sum_{i=0}^n x_i & = \sum_{i=0}^n x_i y_i \\
a_m\sum_{i=0}^n x_i^m & + a_{m-1}\sum_{i=0}^n x_i^{m-1} + \cdots + a_1\sum_{i=0}^n x_i + a_0(n+1) & = \sum_{i=0}^n y_i
\end{array}$$

Most belátjuk, hogy a (9.4) lineáris egyenletrendszernek létezik egyértelmű megoldása, azaz az

$$\mathbf{A} = \begin{pmatrix}
\sum_{i=0}^n x_i^{2m} & \sum_{i=0}^n x_i^{2m-1} & \cdots & \sum_{i=0}^n x_i^{m+1} & \sum_{i=0}^n x_i^m \\
\sum_{i=0}^n x_i^{2m-1} & \sum_{i=0}^n x_i^{2m-2} & \cdots & \sum_{i=0}^n x_i^m & \sum_{i=0}^n x_i^{m-1} \\
\vdots & \vdots & & \vdots & \vdots \\
\sum_{i=0}^n x_i^m & \sum_{i=0}^n x_i^{m-1} & \cdots & \sum_{i=0}^n x_i & \sum_{i=0}^n 1
\end{pmatrix}$$

együtthatómátrix invertálható. Ehhez a 3.9. tétel szerint elegendő megmutatni, hogy $\mathbf{A}$ pozitív definit. Az $\mathbf{A}$ mátrix jk-adik elemét a $\sum_{i=0}^n x_i^{2m+2-j-k}$ képlettel adhatjuk meg, ahol $j,k = 1,2,\ldots,m+1$. Legyen $\mathbf{z} = (z_1, z_2, \ldots, z_{m+1}) \in \mathbb{R}^{m+1}$. Egyszerű átalakításokkal adódik

$$\begin{aligned}
\mathbf{z}^T\mathbf{A}\mathbf{z} &= \sum_{j=1}^{m+1}\sum_{k=1}^{m+1}\sum_{i=0}^{n} x_i^{2m+2-j-k} z_j z_k \\
&= \sum_{i=0}^{n}\sum_{j=1}^{m+1}\sum_{k=1}^{m+1} x_i^{m+1-j} z_j x_i^{m+1-k} z_k \\
&= \sum_{i=0}^{n}\left(\sum_{j=1}^{m+1} x_i^{m+1-j} z_j\right)^2 .
\end{aligned}$$

Tegyük fel, hogy $\mathbf{z}^T\mathbf{A}\mathbf{z} = 0$. Ekkor az előbbi számolásból következik, hogy $\sum_{j=1}^{m+1} x_i^{m+1-j} z_j = 0$ minden $i = 0, 1, \ldots, n$-re. Eszerint ha az x_i alappontok páronként különböznek, akkor a $p(x) := \sum_{j=1}^{m+1} z_j x^{m+1-j}$ m-edfokú polinomnak $n+1$ különböző gyöke van. Ha feltesszük, hogy $m \le n$, akkor az algebra alaptétele szerint ebből következik, hogy p azonosan nulla, azaz $z_j = 0$ minden $j = 1, 2, \ldots, m+1$-re. Ezzel beláttuk, hogy $\mathbf{A}$ pozitív definit, és így a (9.4) egyenletrendszernek létezik egyértelmű megoldása, amit $\bar{\mathbf{a}}$-val jelölünk. Mivel

$$\frac{\partial^2 F}{\partial a_j\,\partial a_k}(\bar{\mathbf{a}}) = 2\sum_{i=0}^n x_i^{j+k},$$

ezért $F''(\bar{\mathbf{a}}) = 2\mathbf{A}$. Ebből következik a 8.1. tétel alapján, hogy F-nek $\bar{\mathbf{a}}$-ban lokális minimuma van, és mivel F kvadratikus függvény, ezért ez globális minimum is. Az eredményeinket a következő tételben összegezhetjük:

9.3. tétel. *Adottak az (x_i, y_i) $(i = 0, 1, \ldots, n)$ pontok, ahol az x_i alappontok páronként különböznek. Legyen $m \leq n$. Ekkor a*

$$\min_{(a_m,\ldots,a_0)\in\mathbb{R}^{m+1}} \sum_{i=0}^{n} (a_m x_i^m + a_{m-1} x_i^{m-1} + \cdots + a_1 x_i + a_0 - y_i)^2$$

szélsőérték feladatnak létezik egyértelmű megoldása, amely teljesíti a (9.4) normálegyenleteket.

9.4. példa. Illesszünk parabolát az

x_i	-1.0	-0.5	0.0	1.0	2.0	3.0	3.5
y_i	1.6	1.7	1.9	1.5	0.6	-0.1	-1.0

adatokra! Kézi számoláskor a 9.2. táblázatban látható módon helyezzük el az adatokat. Az utolsó sorban szereplő összegeket használjuk a (9.4) egyenletrendszerhez:

$$\begin{aligned} 249.1250a &+ 77.750b &+ 27.50c &= -7.225 \\ 77.750a &+ 27.50b &+ 8.0c &= -3.55 \\ 27.50a &+ 8.0b &+ 7c &= 6.2. \end{aligned}$$

amelyet megoldva kapjuk, hogy $a = -0.196021$, $b = -0.084748$ és $c = 1.752653$. A megadott pontokat és a számított parabola grafikonját a 9.2. ábrán láthatjuk. Az illesztés hibája

$$\sum_{i=0}^{6} (-0.196021x_i^2 - 0.084748x_i + 1.752653 - y_i)^2 = 0.0964456.$$

□

9.2. táblázat. Parabola illesztése

x_i	y_i	x_i^4	x_i^3	x_i^2	$x_i^2 y_i$	$x_i y_i$
-1.0	1.4	1.0000	-1.000	1.00	1.400	-1.40
0.0	1.9	0.0000	0.000	0.00	0.000	0.00
0.5	1.6	0.0625	0.125	0.25	0.400	0.80
1.0	1.7	1.0000	1.000	1.00	1.700	1.70
2.0	0.2	16.0000	8.000	4.00	0.800	0.40
2.5	-0.1	39.0625	15.625	6.25	-0.625	-0.25
3.0	-2.0	81.0000	27.000	9.00	-18.000	-6.00
8.0	4.7	138.1250	50.750	21.50	-14.325	-4.75

Feladatok

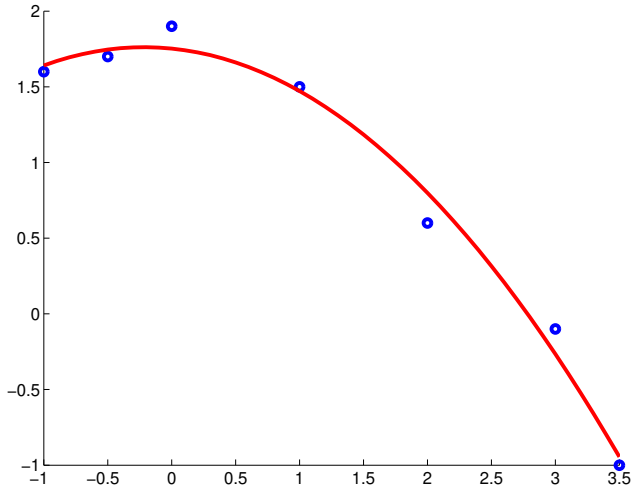
1. Illesszen parabolát a megadott adatokra és számítsa ki az illesztés hibáját:

(a)

x_i	-2.0	-1.0	1.0	2.0	3.0
y_i	-2.1	1.4	0.5	-2.5	-7.2

(b)

x_i	1.0	2.0	3.0	4.0	5.0	6.0
y_i	2.5	1.2	-2.0	3.9	6.2	8.3



9.2. ábra. Parabola illesztése: $y = -0.196021x^2 - 0.084748x + 1.752653$

9.3. Nemlineáris függvény illesztése

Az előző szakaszokban vizsgált módszert alkalmazhatjuk olyan nemlineáris függvény illesztésre is, ahol az ismeretlen paraméterek lineárisan szerepelnek, mert ekkor a kapott normálegyenletek lineáris egyenletek lesznek. Az általános esetben viszont a normálegyenletek is lehetnek nemlineárisak. Nézzünk egy példát. Tegyük fel, hogy egy be^{ax} alakú exponenciális függvényt szeretnénk illeszteni az (x_i, y_i) $(i = 0, 1, \ldots, n)$ pontokra. A négyzetes hibát felírva az

$$F(a, b) = \sum_{i=0}^{n} (be^{ax_i} - y_i)^2$$

függvényt kapjuk, amelynek kritikus pontjait a

$$\begin{aligned} 2\sum_{i=0}^{n} (be^{ax_i} - y_i)be^{ax_i}x_i &= 0 \\ 2\sum_{i=0}^{n} (be^{ax_i} - y_i)e^{ax_i} &= 0 \end{aligned}$$

egyenletrendszer megoldásai adják. Ezt már analitikusan nem tudjuk megoldani, és azt sem könnyű látni, hogy hány megoldás van, és ha több van, melyik megoldás fogja minimalizálni F-et. Természetesen meg tudjuk oldani az egyenletrendszert numerikusan, ill. a 8. fejezetben ismertetett numerikus módszerek segítségével tudjuk közelíteni F minimumát.

Az előbb vázolt számolás helyett alkalmazható a következő, ún. *linearizációs módszer*: Vegyük észre, hogy ha az $y = be^{ax}$ egyenlet mindkét oldalának vesszük a logaritmusát, akkor az $\ln y = \ln b + ax$ összefüggést kapjuk, ahol $\ln y$ lineárisan függ x-től. Vezessünk be új változókat: $X := x$, $Y := \ln y$, $A := a$ és $B := \ln b$. Illesszünk tehát $Y = AX + B$ alakú egyenest az adott $(x_i, \ln y_i)$ adatokra. Legyenek $\bar{A}$ és $\bar{B}$ az egyenes illesztésekor kapott konstansok. Ekkor a $\bar{b}e^{\bar{a}x}$ függvényt tekintjük az (x_i, y_i) pontokra legjobban illeszkedő exponenciális függvénynek, ahol $\bar{a} = \bar{A}$, $\bar{b} = e^{\bar{B}}$. Megjegyezzük, hogy a linearizációs módszerrel illesztett függvény természetesen nem megoldása az eredeti nemlineáris illesztési feladatnak, viszont könnyű kiszámolni, így a gyakorlatban ezt az illesztést célszerű alkalmazni.

9.5. példa. Illesszünk be^{ax} alakú függvényt az

x_i	0.0	1.0	1.5	2.0	3.0	4.0
y_i	0.3	0.7	0.9	1.2	1.8	2.7

pontokra! A linearizált adatok a 9.3. táblázatban láthatók. Az egyenes illesztésekor kapott

$$\begin{aligned} 32.25A &+ 11.5B &= 5.586294 \\ 11.5A &+ 6B &= 0.097352, \end{aligned}$$

normálegyenletek megoldása $A = 0.528951$ és $B = -0.997597$, azaz a linearizálás módszerével illesztett függvény képlete $0.368765^{0.528951x}$. Ennek a függvénynek és az adatoknak a grafikonja a 9.3. ábrán látható. A linearizált illesztés hibája

$$\sum_{i=0}^{5}(0.528951x_i - 0.997597 - \ln y_i)^2 = 0.095396,$$

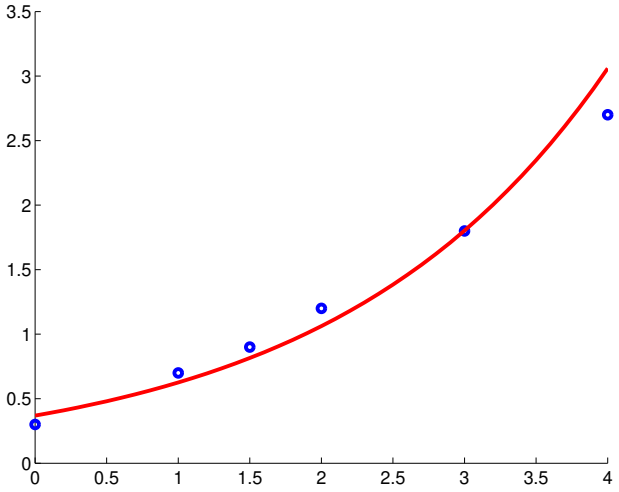
az eredeti hiba a kapott függvényre pedig

$$\sum_{i=0}^{5}(0.368765^{0.528951x_i} - y_i)^2 = 0.165543.$$

□

9.3. táblázat. be^{ax} alakú függvény illesztése

x_i	y_i	$\ln y_i$	x_i^2	$x_i \ln y_i$
0.0	0.3	-1.203973	0.00	0.000000
1.0	0.7	-0.356675	1.00	-0.356675
1.5	0.9	-0.105361	2.25	-0.158041
2.0	1.2	0.182322	4.00	0.364643
3.0	1.8	0.587787	9.00	1.763360
4.0	2.7	0.993252	16.00	3.973007
11.5		0.097352	32.25	5.586294



9.3. ábra. be^{ax} alakú függvény illesztése: be^{ax}: $y = 0.368765^{0.528951x}$

9.6. példa. Illesszünk egy bx^a alakú hatványfüggvényt a következő pontokra:

x_i	0.5	1.0	1.5	2.5	3.0
y_i	0.7	1.1	1.6	2.1	2.3

Ebben az esetben is alkalmazható a linearizálás módszere: tekintsük az $\ln y = a \ln x + \ln b$ összefüggést. Ekkor $\ln y$ lineárisan függ $\ln x$-től. Illesszünk tehát egy egyenest az $(\ln x_i, \ln y_i)$ pontokra. A számolást a 9.4. táblázatban láthatjuk, a kapott normálegyenletek:

$$\begin{array}{rcrcl} 2.691393A & + & 1.727221B & = & 2.032673 \\ 1.727221A & + & 5B & = & 1.783485. \end{array}$$

Ennek megoldása $A = 0.676257$, $B = 0.123088$. Ebből az eredeti paraméterek: $a = A = 0.676257$ és $b = e^B = e^{0.123088} = 1.130984$. A linearizált illesztés hibája

$$\sum_{i=0}^{4} (0.676257 \ln x_i + 0.123088 - \ln y_i)^2 = 0.007279,$$

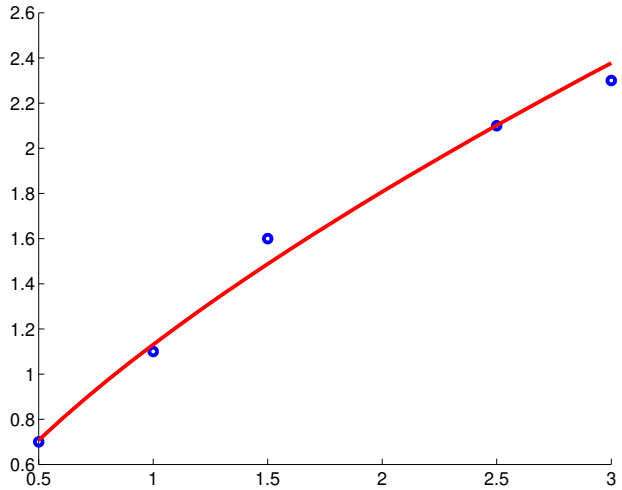
az eredeti négyzetes hiba pedig

$$\sum_{i=0}^{4} (1.130984 x_i^{0.676257} - y_i)^2 = 0.019616.$$

□

9.4. táblázat. bx^a alakú függvény illesztése

x_i	y_i	$\ln x_i$	$\ln y_i$	$(\ln x_i)^2$	$\ln x_i \ln y_i$
0.5	0.7	-0.693147	-0.356675	0.480453	0.247228
1.0	1.1	0.000000	0.095310	0.000000	0.000000
1.5	1.6	0.405465	0.470004	0.164402	0.190570
2.5	2.1	0.916291	0.741937	0.839589	0.679830
3.0	2.3	1.098612	0.832909	1.206949	0.915044
		1.727221	1.783485	2.691393	2.032673



9.4. ábra. bx^a alakú függvény illesztése: $y = 1.130984x^{0.676257}$

Feladatok

1. Illesszen be^{ax} alakú függvényt a megadott adatokra és számítsa ki az illesztés hibáját:

(a)

x_i	-2.0	-1.0	1.0	2.0	3.0
y_i	0.6	0.9	1.6	2.3	2.9

(b)

x_i	1.0	1.5	2.0	2.5	3.0	3.5
y_i	1.3	1.6	1.9	2.2	3.0	4.1

2. Illesszen bx^a alakú függvényt a megadott adatokra és számítsa ki az illesztés hibáját:

(a)

x_i	1.0	3.0	4.0	5.0	6.0	9.0
y_i	1.6	1.9	2.2	2.3	3.4	4.9

(b)

x_i	1.0	2.0	3.0	4.0	5.0
y_i	0.7	2.8	7.5	14.8	25.6

3. Oldja meg az előző két feladatot az eredeti nemlineáris négyzetes hibát minimalizálva Newton-módszerrel!

10. fejezet
Közönséges differenciálegyenletek

Ebben a fejezetben közönséges differenciálegyenletek numerikus megoldásait vizsgáljuk az Euler-, Taylor-, és Runge–Kutta módszerekkel.

10.1. Differenciálegyenletek előismeretek

Ebben a fejezetben az

$$y' = f(t, y), \qquad y(t_0) = y_0 \tag{10.1}$$

kezdeti érték probléma közelítő megoldását keressük egy véges $[t_0, T]$ intervallumon. Az egyszerűség kedvéért a közelítő módszerek tárgyalásakor azt az esetet vizsgáljuk, ahol $y = y(t)$ valós értékű függvény, azaz feltesszük, hogy

$$f\colon [t_0, T] \times \mathbb{R} \to \mathbb{R}, \qquad y_0 \in \mathbb{R}.$$

A kapott eredmények könnyen átvihetők differenciálegyenlet-rendszerekre: ekkor $\mathbf{y} = \mathbf{y}(t)$ az ismeretlen függvényekből képzett m-dimenziós vektort jelöl, és a vizsgált egyenletrendszert vektor jelöléssel az

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}^{(0)}, \tag{10.2}$$

alakban írjuk fel, ahol

$$\mathbf{f}\colon [t_0, T] \times \mathbb{R}^m \to \mathbb{R}^m, \qquad \mathbf{y}^{(0)} \in \mathbb{R}^m.$$

Vezessük be a következő definíciót: Az $f\colon [t_0, T] \times \mathbb{R} \to \mathbb{R}$ függvény a második változójában teljesíti a *Lipschitz-tulajdonságot* az L Lipschitz-konstanssal, ha

$$|f(t, y) - f(t, \tilde{y})| \le L|y - \tilde{y}| \qquad \text{minden } t \in [t_0, T] \text{ és } y, \tilde{y} \in \mathbb{R}\text{-re.} \tag{10.3}$$

Ezt a fogalmat könnyen általánosíthatjuk a vektor értékű esetre, ha abszolút érték helyett normát használunk az előző definícióban.

A differenciálegyenletek elméletéből tudjuk, hogy a (10.1) ill. (10.2) kezdeti érték problémák megoldhatóságához annyit kell csak feltenni, hogy az f ill. $\mathbf{f}$ függvények folytonosak legyenek, valamint a megoldások egyértelműségéhez még azt is fel kell tenni, hogy a második változójukban Lipschitz-tulajdonságúak legyenek. Érvényes tehát a következő állítás (a skalár esetre megfogalmazva):

10.1. tétel. *Tegyük fel, hogy az $f : [t_0, T] \times \mathbb{R} \to \mathbb{R}$ folytonos függvény a második változójában Lipschitz-tulajdonságú (valamely L Lipschitz-konstanssal). Ekkor a (10.1) kezdeti érték problémának minden $y_0 \in \mathbb{R}$ kezdeti értékhez létezik egyértelmű megoldása a $[0, T]$ intervallumon.*

Megjegyezzük, hogy a 10.1. tétel és a későbbiekben megfogalmazandó tételek feltételeiben szereplő Lipschitz-tulajdonság, azaz a (10.3) egyenlőtlenség teljesülésének megkövetelése minden $y, \bar{y} \in \mathbb{R}$-re elég erős megszorítás f-re nézve. Ehelyett szokás gyengébb, ún. lokális Lipschitz-tulajdonságot megkövetelni: minden $T > t_0$ és $[a, b]$ intervallumhoz, amelyre $y_0 \in (a, b)$, létezik

olyan $L > 0$ szám (amely T-től és $[a, b]$-től függ), hogy (10.3) teljesül minden $t \in [t_0, T]$, $y, \bar{y} \in [a, b]$-re. Ez a feltétel a gyakorlatban fellépő f függvények nagyrészére teljesül. Például elég azt feltenni, hogy a folytonos f függvény folytonosan differenciálható a második változója szerint, abból következik, hogy lokálisan Lipschitz-tulajdonságú a második változójában (3. feladat). A lokális Lipschitz-feltételből viszont nem garantálható, hogy a (10.1) feladat megoldása az egész $[t_0, T]$ intervallumon létezik, csak annyit mondhatunk, hogy létezik olyan $0 < \bar{T} \leq T$ szám, hogy a (10.1) feladatnak egyértelmű megoldása létezik a $[t_0, \bar{T}]$ intervallumon (lásd 4. feladat). Ennek a technikai problémának elkerülésére a későbbi bizonyításainkhoz feltesszük, hogy f globálisan, azaz (10.3) értelmében Lipschitz-tulajdonságú.

Ismert, hogy az

$$y^{(m)} = f(t, y, y', \ldots, y^{(m-1)}), \quad y(t_0) = y_0,\ y'(t_0) = y_1, \ldots,\ y^{(m-1)}(t_0) = y_{m-1}$$

m-edrendű kezdeti érték feladat ekvivalens egy (10.2) alakú elsőrendű differenciálegyenlet-rendszerrel, ahol

$$\mathbf{y} = (y, y', \ldots, y^{(m-1)})^T, \quad \text{és} \quad \mathbf{y}^{(0)} = (y_0, y_1, \ldots, y_{m-1})^T.$$

Mi az egyszerűség kedvéért csak a (10.1) alakú elsőrendű skaláris differenciálegyenletekkel foglalkozunk a továbbiakban, de a később ismertetett módszerek egyrésze könnyen átfogalmazható differenciálegyenlet-rendszerekre is.

Feladatok

1. Alakítsa át a következő magasabbrendű differenciálegyenletekhez tartozó kezdeti érték feladatokat (10.2) alakra:
   (a) $y'' + 5y' = e^{2t-1}, \qquad y(0) = 3, \quad y'(0) = -1,$
   (b) $y'' - t^2 y' + ty = 0, \qquad y(1) = 1, \quad y'(1) = 0,$
   (c) $y''' + 4y'' - 2y' + 5y = t^3, \qquad y(-1) = 2, \quad y'(-1) = -3.$
2. Bizonyítsa be, hogy az $y' = \sqrt{|y|}$, $y(0) = 0$ kezdeti érték feladatnak $y(t) = 0$ és $y(t) = t^2/4$ is megoldása. Mutassa meg, hogy az $f(y) = \sqrt{|y|}$ függvény nem Lipschitz-tulajdonságú y-ban.
3. Bizonyítsa be, hogy ha az $f : [t_0, T] \times \mathbb{R} \to \mathbb{R}$ folytonos függvény a második változója szerint folytonosan parciálisan differenciálható, akkor f lokális Lipschitz-tulajdonságú a második változójában.
4. Igazolja, hogy az $y' = y^2$, $y(0) = 1$ kezdeti érték feladatnak nem létezik megoldása a $[0, T]$ intervallumon, ha $T \geq 1$! Mutassa meg, hogy a $g(y) = y^2$ függvény nem globális Lipschitz-tulajdonságú y-ban, viszont lokális Lipschitz-tulajdonságú!

10.2. Euler-módszer

Tekintsük a (10.1) kezdeti érték problémát. Ebben a szakaszban a probléma legegyszerűbb numerikus megoldási módszerét, az ún. *Euler-módszert* vizsgáljuk. A célunk az, hogy egy $[t_0, T]$ véges intervallumon, előre megadott véges sok pontban közelítsük a megoldást. Jelöljük ezeket az alappontokat (a t_0 ponttal kezdve): $t_0 < t_1 < \cdots < t_n = T$-vel, és az alappontok távolságát h_i-vel, azaz $h_i = t_{i+1} - t_i$ $(i = 0, \ldots, n-1)$. Nem kell feltennünk a módszer definiálásakor, hogy az alappontok ekvidisztánsak (azaz $h_i = h$ állandó), de a gyakorlatban természetesen erre az esetre alkalmazzuk a leggyakrabban a módszert. Az alappontokbeli $y(t_i)$ megoldásértékek közelítésére definiáljuk a z_i, ún. *Euler-sorozatot* a

$$z_{i+1} = z_i + h_i f(t_i, z_i), \quad (i = 0, 1, 2, \ldots, n-1), \qquad z_0 = y_0 \tag{10.4}$$

rekurzív képlettel.

Most háromféleképpen is levezetjük az Euler-módszer képletét, majd utána vizsgáljuk a közelítés hibáját.

1. levezetés: Tegyük fel, hogy $y(t)$ megoldása a (10.1) kezdeti érték problémának. Mivel $y(t)$ teljesíti a kezdeti feltételt, tudjuk az t_0 alappontbeli értékét: $y(t_0) = y_0$, ezért z_0 a pontos értéke a megoldásnak a t_0 pontban. Hogyan becsülhetjük $y(t_1)$-et? Közelítsük az $y(t)$ függvényt a t_0 pontjához tartozó elsőrendű Taylor-polinomjával (azaz geometriailag a függvény grafikonját az adott ponthoz tartozó érintőjével közelítjük): $y(t) \approx y(t_0) + y'(t_0)(t - t_0)$. Ekkor a $t = t_1$ pontban kapjuk, hogy

$$y(t_1) \approx y(t_0) + y'(t_0)h_1. \tag{10.5}$$

Ebben a képletben szerepel még a megoldás deriváltja a t_0 pontban, ami a (10.1) egyenletet alapján $y'(t_0) = f(t_0, y(t_0))$. Mivel $y(t_0) = y_0 = z_0$, így $y'(t_0)$-t ki tudjuk számítani az egyenlet jobb oldala, t_0 és a már definiált z_0 érték segítségével: $y'(t_0) = f(t_0, z_0)$. Ezért a (10.5) összefüggésből kapjuk, hogy $y(t_1) \approx z_1 := z_0 + h_1 f(t_0, z_0)$. Használhatjuk tehát z_1-et, mint a megoldás t_1-beli közelítését. Hogyan közelítsük $y(t_2)$-t, ill általában $y(t_{i+1})$-et, ha már ismert az $y(t_i)$ megoldásérték z_i közelítése? Az előző ötletet követve $y(t_{i+1}) \approx y(t_i) + y'(t_i)h_i$, és mivel $y(t_i) \approx z_i$ és így $y'(t_i) = f(t_i, y(t_i)) \approx f(t_i, z_i)$, kapjuk, hogy $y(t_{i+1}) \approx z_{i+1}$, ahol z_{i+1}-et a (10.4) képlettel definiáltuk.

2. levezetés: A megoldás teljesíti az $y'(t_i) = f(t_i, y(t_i))$ összefüggést. Elsőrendű numerikus differenciálási képletet használva

$$y'(t_i) \approx \frac{y(t_{i+1}) - y(t_i)}{h_i},$$

azaz

$$\frac{y(t_{i+1}) - y(t_i)}{h_i} \approx f(t_i, y(t_i)).$$

Ezt átrendezve kapjuk, hogy $y(t_{i+1}) \approx y(t_i) + h_i f(t_i, y(t_i))$. Feltéve hogy $y(t_i) \approx z_i$, a (10.4) képlettel definiált z_{i+1} teljesíti az $y(t_{i+1}) \approx z_{i+1}$ összefüggést.

3. levezetés: Az $y'(t) = f(t, y(t))$ differenciálegyenlet mindkét oldalát integrálva t_i-től t_{i+1}-ig kapjuk, hogy

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} f(s, y(s))\, ds,$$

azaz

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(s, y(s))\, ds. \tag{10.6}$$

A probléma az, hogy nem ismerjük az $f(s, y(s))$ összetett függvényt, mivel nem ismerjük $y(s)$ képletét. Így az integrál pontos értékét nem tudjuk kiszámítani. Használjunk egy egyszerű integrál közelítő képletet:

$$\int_a^b g(s)\, ds \approx g(a)(b - a). \tag{10.7}$$

Ez a közelítő képlet alkalmazható ebben az esetben, mivel ehhez csak a függvény intervallum bal oldali végpontjához tartozó értéke szükséges, amit felteszünk, hogy már ismerünk. Ezt a közelítést alkalmazva $\int_{t_i}^{t_{i+1}} f(s, y(s))\, ds \approx h_i f(t_i, y(t_i))$, azaz

$$y(t_{i+1}) \approx y(t_i) + h_i f(t_i, y(t_i)),$$

amiből szintén megkapjuk a (10.4) formulát.

Az Euler-módszer 1. levezetése alapján a módszerhez a következő geometriai interpretációt rendelhetjük hozzá: az i-edik lépésben megkapott (t_i, z_i) pontból egy egyenes (a ponton átmenő megoldás érintője) mentén lépünk tovább egy „egységet", azaz az egyenesen levő, t_{i+1} első koordinátájú pontba.

10.2. példa. Tekintsük az

$$y' = 2y - 10t^2 + 2t, \qquad y(0) = 1. \tag{10.8}$$

kezdeti érték feladatot! Könnyen ellenőrizhetjük, hogy a feladat analitikus megoldása $y(t) = 5t^2 + 4t + 2 - e^{2t}$. Vegyünk egy h lépésközhöz tartozó $t_i = ih$ ekvidisztáns beosztást! Az Euler-sorozatot a

$$z_{i+1} = z_i + h\Big(2z_i - 10t_i^2 + 2t_i\Big), \qquad i = 0, 1, 2, \ldots, \qquad z_0 = 1.$$

rekurzív definícióval számoljuk ki. A 10.1. táblázat tartalmazza a közelítő sorozat $h = 0.2$, 0.1 és 0.05 lépésközökhöz tartozó első néhány tagját és a közelítés $e_i = |y(t_i) - z_i|$ hibáját. Láthatjuk, hogy a lépésközt csökkentve a közelítés hibája is csökken, sőt azt is észrevehetjük, hogy a hiba h-val lineárisan arányos: ha felezzük a lépésközt, a hiba is körülbelül fele akkora lesz. □

10.1. táblázat. Euler-módszer

		$h = 0.2$			$h = 0.1$			$h = 0.05$		
t_i	$y(t_i)$	i	z_i	e_i	i	z_i	e_i	i	z_i	e_i
0.0	1.0000	0	1.0000	0.0000	0	1.0000	0.0000	0	1.0000	0.0000
0.2	1.0652	1	1.1000	0.0348	2	1.0830	0.0178	4	1.0742	0.0090
0.4	1.0614	2	1.1340	0.0726	4	1.0986	0.0372	8	1.0802	0.0188
0.6	0.9899	3	1.1034	0.1135	6	1.0481	0.0583	12	1.0194	0.0295
0.8	0.8518	4	1.0097	0.1579	8	0.9329	0.0811	16	0.8930	0.0411
1.0	0.6487	5	0.8547	0.2060	10	0.7547	0.1060	20	0.7025	0.0538

Most rátérünk az Euler-módszer konvergenciájának vizsgálatára. Az egyszerűség kedvéért tegyük fel, hogy ekvidisztáns osztópontokra alkalmazzuk az Euler-módszert, azaz $h_i = h$ konstans. Szükségünk lesz a következő definícióra: Az Euler-módszer $(i+1)$-edik *lokális képlethibáján* a

$$\tau_{i+1} := \frac{y(t_{i+1}) - y(t_i)}{h} - f(t_i, y(t_i)), \qquad (i = 0, 1, \ldots, n-1) \tag{10.9}$$

számot értjük, ahol $y(t)$ a (10.1) feladat pontos megoldása.

Átrendezve a (10.9) egyenletet következik

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \tau_{i+1}h. \tag{10.10}$$

Innen látható, hogy $\tau_{i+1}h$ adja a numerikus módszer hibáját az $(i+1)$-edik lépés megtételekor, ha feltesszük, hogy az i-edik lépésben a pontos értékből indulunk ki.

Vegyük $y(t)$ elsőrendű Taylor-közelítését a t_i pont körül:

$$y(t) = y(t_i) + y'(t_i)(t - t_i) + \frac{1}{2}y''(\xi)(t - t_i)^2.$$

Ebből kapjuk, használva az $y'(t_i) = f(t_i, y(t_i))$ összefüggést és a (10.10) egyenletet, hogy az Euler-módszer lokális képlethibája

$$\tau_{i+1} = \frac{h}{2}y''(\xi) \tag{10.11}$$

alakú, ahol $\xi \in (t_i, t_{i+1})$.

Szükségünk lesz az alábbi állításra:

10.3. tétel. *Legyenek a, b pozitív valós számok, $x_0, x_1, x_2, \dots$ egy számsorozat, amelyre $x_0 \geq -b/a$, és*

$$x_{i+1} \leq (1+a)x_i + b, \qquad i \geq 0.$$

Ekkor

$$x_i \leq e^{ia}\left(\frac{b}{a} + x_0\right) - \frac{b}{a}$$

teljesül minden $i \geq 0$-ra.

Bizonyítás. Egymás után alkalmazva a feltételt és elemi átalakításokat kapjuk a következő összefüggéseket:

$$\begin{aligned} x_i &\leq (1+a)x_{i-1} + b \\ &\leq (1+a)((1+a)x_{i-2} + b) + b \\ &\vdots \\ &\leq (1+a)((1+a)(\cdots((1+a)x_0 + b)\cdots) + b) + b \\ &= (1+a)^i x_0 + (1 + (1+a) + (1+a)^2 + \cdots + (1+a)^{i-1})b \\ &= (1+a)^i x_0 + \frac{(1+a)^i - 1}{a} b \\ &= (1+a)^i \left(\frac{b}{a} + x_0\right) - \frac{b}{a}. \end{aligned} \tag{10.12}$$

Az $1+x \leq e^x$ elemi egyenlőtlenségből kapjuk, hogy $(1+x)^i \leq e^{ix}$, ami a (10.12) egyenlőtlenséggel együtt adja a tétel állítását. □

10.4. tétel. *Legyen az $f : [t_0, T] \times \mathbb{R} \to \mathbb{R}$ folytonos függvény a második változójában Lipschitz-tulajdonságú az L Lipschitz-konstanssal, jelölje $z_0, z_1, \dots, z_n$ az Euler-sorozatot, és $\tau = \max\{|\tau_{i+1}| \colon i = 0, 1, \dots, n-1\}$. Ekkor*

$$|y(t_i) - z_i| \leq \left(e^{L(T-t_0)} - 1\right)\frac{\tau}{L}, \qquad (i = 0, 1, \dots, n). \tag{10.13}$$

Bizonyítás. A (10.10) és (10.4) egyenleteket egymásból kivonva

$$y(t_{i+1}) - z_{i+1} = y(t_i) - z_i + h\Big(f(t_i, y(t_i)) - f(t_i, z_i)\Big) + \tau_{i+1}h$$

adódik. Ebből a háromszög-egyenlőtlenséget, f Lipschitz-tulajdonságát, τ definícióját és a $h = \max\{h_i \colon i = 0, 1, \dots, n-1\}$ jelölést használva:

$$\begin{aligned} |y(t_{i+1}) - z_{i+1}| &\leq |y(t_i) - z_i| + h\Big|f(t_i, y(t_i)) - f(t_i, z_i)\Big| + |\tau_{i+1}|h \\ &\leq |y(t_i) - z_i| + Lh|y(t_i) - z_i| + |\tau_{i+1}|h \\ &\leq (1 + Lh)|y(t_i) - z_i| + \tau h. \end{aligned}$$

Ez utóbbi egyenlőtlenségre alkalmazva a 10.3. tételt az $x_i = |y(t_i) - z_i|$, $a = Lh$, $b = \tau h$ választással, és használva az $x_0 = 0$ és $nh = t_n - t_0 = T - t_0$ relációkat adódik (10.13). □

A tételből következik, hogy a közelítés hibája

$$|y(t_i) - z_i| \leq K_1\tau, \qquad i = 0, 1, \dots, n \tag{10.14}$$

alakban becsülhető (ahol K_1 egy adott konstans), azaz az Euler-sorozat közelítési hibája kicsi, feltéve hogy minden egyes lépés lokális képlethibája kicsi. A (10.11) képlet szerint τ_{i+1} megbecsülhető a

$$|\tau_{i+1}| \le \frac{M_2}{2}h, \qquad i = 0, 1, \dots, n-1 \tag{10.15}$$

alakban, ahol $M_2 = \max\{|y''(t)| : t \in [t_0, T]\}$ (feltéve persze, hogy a megoldás kétszer differenciálható). Ebből adódik, hogy ha h kicsi, akkor a közelítés hibája is kicsi.

A megoldás (definíció szerint) mindig differenciálható függvény, és a deriváltja teljesíti az $y'(t) = f(t, y(t))$ egyenletet. Ha tehát feltesszük, hogy f folytonosan parciálisan differenciálható mindkét változója szerint, akkor a többváltozós függvényekre vonatkozó láncszabály szerint y kétszer differenciálható, és

$$y''(t) = \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t))y'(t).$$

Itt viszont használhatjuk újra az egyenletet $y'(t)$ helyettesítésére:

$$y''(t) = \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t))f(t, y(t)). \tag{10.16}$$

Ha például f és parciális deriváltjai korlátosak, akkor (10.16) segítségével rögtön kaphatunk egy explicit becslést M_2-re.

Összegezve az eddigieket, beláttuk a következő állítást:

10.5. tétel. *Legyen $f : [t_0, T] \times \mathbb{R} \to \mathbb{R}$ folytonos függvény a második változójában Lipschitz-tulajdonságú, és folytonosan parciálisan differenciálható mindkét változója szerint. Ekkor az Euler-sorozat elsőrendben konvergál a megoldáshoz, azaz létezik egy $K > 0$ konstans, hogy*

$$|y(t_i) - z_i| \le Kh, \qquad i = 0, 1, \dots, n.$$

Feladatok

1. Számítsa ki a megadott lépésközhöz tartozó Euler-sorozat első tíz tagját és a közelítés hibáját (használva a megadott analitikus megoldást) a következő feladatokra:
   (a) $ty' - y = 2t$, $y(1) = 1$, $h = 0.1$, the solution: $y(t) = 2t\ln t + t$,
   (b) $y' - 2y = 6$, $y(0) = 2$, $h = 0.1$, $y(t) = -3 + 5e^{2t}$,
   (c) $y' - \frac{2}{t}y = 1$, $y(1) = 1$, $h = 0.2$, $y(t) = 2t^2 - t$,
   (d) $y' = \frac{t}{1+y}$, $y(1) = 2$, $h = 0.1$, $y(t) = \sqrt{t^2+8} - 1$.
2. Fogalmazza meg az Euler-módszert differenciálegyenlet-rendszerekre!
3. Oldja meg a következő differenciálegyenlet-rendszereket Euler-módszerrel, és adja meg a közelítés hibáját (a megadott analitikus megoldás segítségével)!
   (a) $\left.\begin{array}{rcl} y_1' &=& 2y_1 - 3y_2, \\ y_2' &=& -y_1 + 4y_2, \end{array}\right\}$ $t \in [0, 2]$, $y_1(0) = 1$, $y_2(0) = -5$,
   $h = 0.1$, $y_1(t) = -3e^t + 4e^{5t}$, $y_2(t) = -4e^{5t} - e^t$.
   (b) $\left.\begin{array}{rcl} y_1' &=& 2y_1 - 3y_2, \\ y_2' &=& 3y_1 + 2y_2, \end{array}\right\}$ $t \in [0, 1]$, $y_1(0) = 1$, $y_2(0) = 0$,
   $h = 0.1$, $y_1(t) = e^{2t}\cos 3t$, $y_2(t) = e^{2t}\sin 3t$.
4. Fogalmazza át a problémát differenciálegyenlet-rendszerre, majd számítsa ki annak Euler-közelítését a megadott lépésközzel az adott intervallumon! Mi az eredeti feladat közelítő megoldása az osztópontokban? Adja meg a közelítés hibáját (a megadott analitikus megoldás ismeretében)!

(a) $y'' - 3y' + 2y = 2, \quad t \in [0,1] \qquad y(0) = 1, \; y'(0) = -1, \quad h = 0.1, \quad y(t) = 1 + e^t - e^{2t},$

(b) $y'' - 2y' + 5y = 0, \quad t \in [0,2], \qquad y(1) = 1, \quad y'(0) = 3, \quad h = 0.2, \quad y(t) = e^t \sin 2t + e^t \cos 2t.$

5. Legyen $t_i = t_0 + ih$ ekvidisztáns beosztása a $[t_0, T]$ intervallumnak, $\{z_i\}$ a hozzá tartozó Euler-sorozat, és $z(t;h)$ az a lineáris spline függvény, amely a z_i értékeket interpolálja az alappontokban: $z(t_i;h) = z_i$, $i = 0, 1, \ldots, n$. Bizonyítsa be, hogy

$$\sup_{t \in [t_0, T]} |y(t) - z(t;h)| \to 0, \qquad \text{ha } h \to 0.$$

10.3. A kerekítési hiba hatása az Euler-módszerre

A gyakorlatban az Euler-módszer (és bármely más numerikus módszer) alkalmazásakor számítanunk kell a kerekítési hibák fellépésére. Először is az y_0 pontos kezdőérték helyett annak gépi megfelelőjét tároljuk és használjuk kezdeti értékként, valamint minden egyes iterációs lépésben követünk el kerekítési hibát. Jelöljük z_i-vel az előző szakaszban definiált Euler-sorozat pontos értékét, és w_i-vel a ténylegesen számolt értékét. Legyen w_0 a kezdeti érték gépi megfelelője. Legyen $\delta_0 := y_0 - w_0$, és jelölje δ_i az egyes iterációs lépések közben elkövetett kerekítési hibát, azaz tegyük fel, hogy

$$w_{i+1} = w_i + hf(t_i, w_i) + \delta_{i+1}, \qquad i = 0, 1, 2, \ldots, n-1. \tag{10.17}$$

A (10.17) egyenletből kivonva a (10.4) egyenletet kapjuk

$$w_{i+1} - z_{i+1} = w_i - z_i + h(f(t_i, w_i) - f(t_i, z_i)) + \delta_{i+1}.$$

Tegyük fel, hogy f Lipschitz-tulajdonságú a második változójában az L Lipschitz-konstanssal. Jelölje $\delta := \max\{|\delta_1|, |\delta_2|, \ldots, |\delta_n|\}$. Ekkor a háromszög-egyenlőtlenséget használva:

$$\begin{aligned} |w_{i+1} - z_{i+1}| &\le |w_i - z_i| + h|f(t_i, w_i) - f(t_i, z_i)| + |\delta_{i+1}| \\ &\le |w_i - z_i| + hL|w_i - z_i| + \delta, \qquad i = 0, 1, 2, \ldots. \end{aligned}$$

Ebből az egyenlőtlenségből a 10.3. tétel segítségével belátható a következő állítás:

10.6. tétel. *Legyen $f\colon [t_0, T] \times \mathbb{R} \to \mathbb{R}$ folytonos függvény a második változójában Lipschitz-tulajdonságú az L Lipschitz-konstanssal, és folytonosan parciálisan differenciálható mindkét változója szerint. Ekkor*

$$|y(t_i) - w_i| \le \frac{e^{L(T-t_0)} - 1}{L} \left(\frac{hM_2}{2} + \frac{\delta}{h} \right) + |\delta_0| e^{L(T-t_0)}, \qquad i = 0, 1, \ldots, n,$$

ahol $M_2 := \max\{|y''(t)|\colon t \in [t_0, T]\}$ és $\delta := \max\{|\delta_1|, |\delta_2|, \ldots, |\delta_n|\}$.

A 10.6. tételben szereplő $\frac{hM_2}{2} + \frac{\delta}{h}$ tényező már nem lineáris h-ban, sőt

$$\lim_{h \to 0+} \left(\frac{hM_2}{2} + \frac{\delta}{h} \right) = \infty.$$

Ezért túlságosan kis lépésköz választása esetén jelentős lehet az Euler-módszer hibája. A gyakorlatban persze ha a lépésköz nagyságrendekkel nagyobb, mint a kerekítési hiba (ami általában teljesül), akkor a kerekítési hiba hatása kicsi.

Feladatok

1. Dolgozza ki a 10.6. tétel bizonyításának részleteit!
2. Rajzolja fel a 10.6. tételben szereplő $g(h) = \frac{hM_2}{2} + \frac{\delta}{h}$ függvény grafikonját! Határozza meg a függvény minimumát!
3. Az előző feladatban megkapott optimális, azaz a $g(h)$ függvényt minimalizáló lépésköz értékét számítsa ki a 10.2. példában vizsgált feladat esetén, feltéve, hogy $\delta = 0.00001$!

10.4. Taylor-módszer

A 10.2. szakaszban levezetett eredmények könnyen átvihetők általánosabb módszerekre is. Az Euler-módszer képletéből kiindulva definiáljuk a következő általános egylépéses módszert a (10.1) feladat megoldására:

$$z_{i+1} = z_i + hF(t_i, z_i; h), \qquad i = 0, 1, \ldots, n-1, \qquad z_0 = y_0, \tag{10.18}$$

ahol $F : [t_0, T] \times \mathbb{R} \times [0, H] \to \mathbb{R}$, valamely $H > 0$-ra. (Az Euler-módszernél $F(t, z; h) = f(t, z)$.) Megjegyezzük, hogy ebben a szakaszban ekvidisztáns osztópontokra fogalmazzuk meg a módszereket, de a levezetett képleteket alkalmazhatjuk az általános esetben is a $z_{i+1} = z_i + h_i F(t_i, z_i; h_i)$ rekurzív definíció szerint.

Az Euler-módszerhez hasonlóan, a (10.18) módszer $(i+1)$-edik *lokális képlethibáján* a

$$\tau_{i+1} := \frac{y(t_{i+1}) - y(t_i)}{h} - F(t_i, y(t_i); h), \qquad (i = 0, 1, \ldots, n-1) \tag{10.19}$$

számot értjük, ahol $y(t)$ a (10.1) feladat pontos megoldása.

Nyilvánvalóan a 10.4. tétel átvihető a (10.18) módszerre, ha F folytonos és Lipschitz-tulajdonságú a második változójában. A 10.4. tétel utáni levezetések is megismételhetők, és teljesül a (10.14) egyenlőtlenség is. Ha feltesszük, hogy (10.15) is teljesül (ez nem teljesül automatikusan), akkor ebből következik a 10.5. tétel megfelelő változata erre az általános módszerre. Sőt ennél többet is beláthatunk. Könnyen bizonyítható a következő állítás:

10.7. tétel. *Legyen $F : [t_0, T] \times \mathbb{R} \times [0, H] \to \mathbb{R}$ folytonos függvény a második változójában Lipschitz-tulajdonságú, és folytonosan parciálisan differenciálható az első és második változója szerint. Feltesszük, hogy a (10.18) módszer lokális képlethibája α rendű, azaz létezik egy olyan $K_2 > 0$ konstans, hogy*

$$|\tau_{i+1}| \le K_2 h^\alpha$$

minden $i = 0, 1, \ldots, n-1$-re. Ekkor a (10.18) közelítő megoldás is α rendben konvergál a (10.1) feladat megoldáshoz, azaz létezik egy $K > 0$ konstans, hogy

$$|y(t_i) - z_i| \le Kh^\alpha, \qquad i = 0, 1, \ldots, n.$$

Hogyan válasszuk meg F-et, hogy a 10.7. tétel feltételei teljesüljenek? Az Euler-módszer 1. levezetéséből és a (10.11) becslés bizonyításából kézenfekvően adódik az ötlet, hogy ne elsőrendű, hanem magasabbrendű Taylor-polinommal közelítsük a megoldást (feltéve, hogy a megoldás elég sokszor differenciálható):

$$\begin{aligned} y(t) &= y(t_i) + y'(t_i)(t - t_i) + \frac{1}{2}y''(t_i)(t - t_i)^2 + \ldots + \frac{1}{\alpha!}y^{(\alpha)}(t_i)(t - t_i)^\alpha \\ &\quad + \frac{1}{(\alpha+1)!}y^{(\alpha+1)}(\xi_i)(t - t_i)^{\alpha+1}, \end{aligned}$$

ahol $\xi_i \in \langle t, t_i\rangle$. Hogy számolhatók y magasabbrendű deriváltjai? Tudjuk, hogy $y'(t) = f(t, y(t))$. Ha mindkét oldalt deriváljuk, kapjuk a (10.16) egyenletet. Ha (10.16) jobb oldalát deriváljuk t szerint, és használjuk az $y'(t) = f(t, y(t))$ összefüggést, megkapjuk $y'''(t)$-t t, $y(t)$, f és f parciális deriváltjai segítségével. Vezessük be a következő jelölést:

$$f^{(i)}(t, y(t)) := \frac{d^i}{d\,t^i}\Big(f(t, y(t))\Big), \tag{10.20}$$

(azaz $f^{(i)}(t, y(t))$ az $f(t, y(t))$ összetett függvény t-szerinti i-edrendű deriváltja). $f^{(i)}(t, z)$ pedig jelöli azt a képletet, amit az előbb definiált $f^{(i)}(t, y(t))$ képletéből $y(t)$ z-re cserélésével kapunk. Ezt a jelölést használva $y^{(i)}(t) = f^{(i-1)}(t, y(t))$, és így

$$\begin{aligned} y(t_{i+1}) &= y(t_i) + f(t_i, y(t_i))h + \frac{1}{2}f^{(1)}(t_i, y(t_i))h^2 + \ldots + \frac{1}{\alpha!}f^{(\alpha-1)}(t_i, y(t_i))h^\alpha \\ &\quad + \frac{1}{(\alpha+1)!}f^{(\alpha)}(\xi_i, y(\xi_i))h^{\alpha+1}. \end{aligned}$$

Tegyük fel tehát hogy $f \in C^\alpha$, és definiáljuk F-et a következőképpen:

$$F(t, z; h) := f(t, z) + \frac{1}{2}f^{(1)}(t, z)h + \ldots + \frac{1}{\alpha!}f^{(\alpha-1)}(t, z)h^{\alpha-1} \tag{10.21}$$

Ekkor

$$\tau_{i+1} = \frac{1}{(\alpha+1)!}f^{(\alpha)}(\xi_i, y(\xi_i))h^\alpha,$$

azaz a lokális képlethiba h-ban α rendű. A (10.18) és (10.21) képlettel definiált módszert α rendű *Taylor-módszernek* nevezzük.

10.8. példa. Tekintsük újra a (10.8) feladatot, és alkalmazzuk rá először a másodrendű Taylor-módszert! Ehhez számítsuk ki $f^{(1)}$-et:

$$\begin{aligned} f^{(1)}(t, y(t)) &= \frac{d}{dt}\left(2y(t) - 10t^2 + 2t\right) = 2y'(t) - 20t + 2 \\ &= (4y(t) - 20t^2 + 4t) - 20t + 2 = 4y(t) - 20t^2 - 16t + 2. \end{aligned}$$

Ezért a közelítő sorozatunk definíciója:

$$z_{i+1} = z_i + h\Big(2z_i - 10t_i^2 + 2t_i\Big) + \frac{h^2}{2}\Big(4z_i - 20t_i^2 - 16t_i + 2\Big), \qquad i = 0, 1, 2, \ldots, \qquad z_0 = 1.$$

A 10.2. táblázatban felsoroltuk a sorozat $h = 0.2$ és 0.1 lépésközökhöz tartozó első néhány tagját. Látható, hogy a lépésközt felezve a hiba kb. negyedére csökken, ami mutatja a másodrendű konvergenciát. Összehasonlítva a kapott eredményt a 10.1. táblázattal, látható, hogy ezzel a képlettel jelentősen kisebb hibát kapunk, mint az Euler-módszerrel.

10.2. táblázat. Másodrendű Taylor-módszer

		$h = 0.2$			$h = 0.1$		
t_i	$y(t_i)$	i	z_i	$\lvert y(t_i) - z_i\rvert$	i	z_i	$\lvert y(t_i) - z_i\rvert$
0.0	1.00000	0	1.00000	0.0000e-01	0	1.00000	0.0000e-01
0.2	1.50818	1	1.52000	1.1825e-02	2	1.51160	3.4247e-03
0.4	2.17446	2	2.20960	3.5141e-02	4	2.18467	1.0206e-02
0.6	2.87988	3	2.95821	7.8325e-02	6	2.90270	2.2813e-02
0.8	3.44697	4	3.60215	1.5518e-01	8	3.49229	4.5325e-02
1.0	3.61094	5	3.89918	2.8823e-01	10	3.69537	8.4425e-02

Most alkalmazzuk a harmadrendű Taylor-módszert a feladatra. Egyszerű számolással kapjuk, hogy

$$f^{(2)}(t,y(t)) = \frac{d}{dt}\left(4y(t) - 20t^2 - 16t + 2\right) = 4y'(t) - 40t - 16 = 8y(t) - 40t^2 - 32t - 16.$$

Így a közelítő sorozat definíciója:

$$z_{i+1} = z_i + h\Big(2z_i - 10t_i^2 + 2t_i\Big) + \frac{h^2}{2}\Big(4z_i - 20t_i^2 - 16t_i + 2\Big) + \frac{h^3}{6}(8z_i - 40t_i^2 - 32t_i - 16),$$

$i = 0, 1, 2, \ldots$ és $z_0 = 1$. A numerikus eredményeket a 10.3. táblázatban közöljük. □

10.3. táblázat. Harmadrendű Taylor-módszer

		$h = 0.2$			$h = 0.1$		
t_i	$y(t_i)$	i	z_i	$\lvert y(t_i) - z_i\rvert$	i	z_i	$\lvert y(t_i) - z_i\rvert$
0.0	1.00000	0	1.00000	0.0000e-01	0	1.00000	0.0000e-01
0.2	1.50818	1	1.50933	1.1580e-03	2	1.50834	1.6959e-04
0.4	2.17446	2	2.17791	3.4538e-03	4	2.17497	5.0596e-04
0.6	2.87988	3	2.88761	7.7257e-03	6	2.88102	1.1321e-03
0.8	3.44697	4	3.46233	1.5361e-02	8	3.44922	2.2518e-03
1.0	3.61094	5	3.63958	2.8634e-02	10	3.61514	4.1989e-03

Feladatok

1. Ismételje meg a 10.2. szakasz 1. feladatát másod- és harmadrendű Taylor-módszert használva!
2. Fogalmazza meg és alkalmazza a negyed- és ötödrendű Taylor-módszereket a (10.8) feladatra!

10.5. Runge–Kutta-módszerek

A Taylor-módszer nehézsége az, hogy a módszer alkalmazásához ki kell számítani az $f^{(i)}$ deriváltakat, amikor könnyen kaphatunk olyan bonyolult képleteket, amelyek kiértékelése jelentős gépidőt igényelhet, és a sok aritmetikai művelet elvégzése közben a számolási hibák felhalmozódásától is tarthatunk. A *Runge–Kutta-módszerek* a Taylor-módszerek számolási igényét igyekeznek csökkenteni, megőrizve azok magasrendű konvergenciáját. Az alapötletet először másodrendű esetben mutatjuk meg.

Legyen $f \in C^2$, és tekintsük a másodrendű Taylor-módszert definiáló

$$F(t,z;h) = f(t,z) + \frac{h}{2}\left(\frac{\partial f}{\partial t}(t,z) + \frac{\partial f}{\partial y}(t,z)f(t,z)\right)$$

függvényt! (Itt is, mint eddig, $\frac{\partial f}{\partial y}$ jelöli az f függvény második változó szerinti parciális deriváltját.) Hasonlítsuk össze ezt a képletet a következő Taylor-formulával:

$$f(t+a, z+b) = f(t,z) + \frac{\partial f}{\partial t}(t,z)a + \frac{\partial f}{\partial y}(t,z)b + E(t,z,a,b),$$

ahol a hibatag másodrendű, azaz

$$E(t,z,a,b) = \frac{1}{2}\left(\frac{\partial^2 f}{\partial t^2}(\xi,\eta)a^2 + 2\frac{\partial^2 f}{\partial t\,\partial y}(\xi,\eta)ab + \frac{\partial^2 f}{\partial y^2}(\xi,\eta)b^2\right) \tag{10.22}$$

valamely $\xi \in \langle t, t+a\rangle$ és $\eta \in \langle z, z+b\rangle$-re. Ha az $a = h/2$ és $b = f(t,z)h/2$ paraméter választást használjuk, kapjuk hogy

$$f\left(t+\frac{h}{2}, z+\frac{h}{2}f(t,z)\right) = F(t,z;h) + E\left(t,z,\frac{h}{2},\frac{h}{2}f(t,z)\right),$$

azaz $f\left(t+\frac{h}{2}, z+\frac{h}{2}f(t,z)\right)$ „lényeges része" megegyezik $F(t,z;h)$-val. Jelentős különbség viszont, hogy $f\left(t+\frac{h}{2}, z+\frac{h}{2}f(t,z)\right)$-t sokkal egyszerűbb kiszámolni, mint $F(t,z;h)$-t. Ez adja az ötletet, hogy tekintsük a

$$z_{i+1} = z_i + hf\left(t_i+\frac{h}{2}, z_i+\frac{h}{2}f(t_i,z_i)\right), \quad i = 0,1,2,\ldots, \qquad z_0 = y_0 \tag{10.23}$$

közelítő módszert. Ezt a módszert *felezőpont-módszernek* nevezzük. Legyen τ_{i+1} a felezőpont-módszer, $\bar\tau_{i+1}$ pedig a másodrendű Taylor-módszer $(i+1)$-edik lokális képlethibája. Ekkor

$$\begin{aligned}
\tau_{i+1} &= \frac{y(t_{i+1})-y(t_i)}{h} - f\left(t_i+\frac{h}{2}, y(t_i)+\frac{h}{2}f(t_i,y(t_i))\right)\\
&= \frac{y(t_{i+1})-y(t_i)}{h} - F(t_i,y(t_i);h) - E\left(t_i,y(t_i),\frac{h}{2},\frac{h}{2}f(t_i,y(t_i))\right)\\
&= \bar\tau_{i+1} - E\left(t_i,y(t_i),\frac{h}{2},\frac{h}{2}f(t_i,y(t_i))\right).
\end{aligned}$$

Az előző szakaszból ismert, hogy $|\bar\tau_{i+1}| \le \bar K h^2$, és (10.22) valamint $f \in C^2$ biztosítja, hogy létezik olyan $\tilde K$, hogy $\left|E\left(t_i,y(t_i),\frac{h}{2},\frac{h}{2}f(t_i,y(t_i))\right)\right| \le \tilde K h^2$. Ebből viszont következik, hogy $|\tau_{i+1}| \le (\bar K + \tilde K)h^2$, és így a (10.23) módszer másodrendben konvergál, feltéve, hogy a 10.7. tételben a Lipschitz-feltétel is teljesül. Ez a feltétel nyilvánvalóan teljesül, ha feltesszük, hogy f Lipschitz-tulajdonságú a második változójában. (Lásd a 2. feladatot!)

Az előzőekkel analóg módon definiáljuk F-et a következő módon:

$$\begin{aligned}
F(t,z;h) &:= \sum_{j=1}^{p}\gamma_j G_j(t,z;h),\\
G_1(t,z;h) &:= f(t,z),\\
G_j(t,z;h) &:= f\left(t+\alpha_j h, z+h\sum_{k=1}^{j-1}\beta_{jk}G_k(t,z;h)\right), \qquad j = 2,3,\ldots,p.
\end{aligned} \tag{10.24}$$

A (10.18) és (10.24) képletekkel definiált módszerek osztályát *(explicit) Runge–Kutta-módszereknek* nevezzük. A cél úgy megválasztani a képletekben szereplő paramétereket, hogy a lehető legmagasabb rendű lokális képlethibát kapjuk.

Tekintsük most a $p = 2$ esetet. Erre

$$F(t,z;h) = \gamma_1 f(t,z) + \gamma_2 f(t+\alpha_1 h, z+\beta_{11}hf(t,z)).$$

(Ha $\gamma_1 = 0$, $\gamma_2 = 1$, $\alpha_1 = \beta_{11} = 1/2$, akkor visszakapjuk a felezőpont-módszert.) Próbáljuk meg úgy megválasztani a paramétereket, hogy harmadrendű lokális hibát kapjunk. Alkalmazzuk a másodrendű Taylor-formulát a jobb oldalra:

$$\begin{aligned}
F(t,z;h) &= (\gamma_1+\gamma_2)f(t,z) + h\gamma_2\left(\alpha_1\frac{\partial f}{\partial t}(t,z) + \beta_{11}f(t,z)\frac{\partial f}{\partial y}(t,z)\right)\\
&\quad + \frac{h^2}{2}\gamma_2\left(\alpha_1^2\frac{\partial^2 f}{\partial t^2}(t,z) + 2\alpha_1\beta_{11}f(t,z)\frac{\partial^2 f}{\partial t\,\partial y}(t,z)\right.\\
&\quad \left. + \beta_{11}^2(f(t,z))^2\frac{\partial^2 f}{\partial y^2}(t,z)\right) + E(t,z,\alpha_1 h,\beta_{11}hf(t,z)),
\end{aligned} \tag{10.25}$$

ahol E a harmadrendű hibatag. Hasonlítsuk ezt össze a harmadrendű Taylor-módszert definiáló

$$\begin{aligned}\tilde{F}(t,z;h) &= f(t,z)+\frac{h}{2}\left(\frac{\partial f}{\partial t}(t,z)+\frac{\partial f}{\partial y}(t,z)f(t,z)\right)\\ &+\frac{h^2}{6}\Big(\frac{\partial^2 f}{\partial t^2}(t,z)+2f(t,z)\frac{\partial^2 f}{\partial t\,\partial y}(t,z)\\ &+(f(t,z))^2\frac{\partial^2 f}{\partial y^2}(t,z)+\frac{\partial f}{\partial t}(t,z)\frac{\partial f}{\partial y}(t,z)+\left(\frac{\partial f}{\partial y}(t,z)\right)^2 f(t,z)\Big)\end{aligned} \tag{10.26}$$

függvénnyel. Láthatjuk, hogy F legfeljebb másodrendű tagjai mind szerepelnek $\tilde{F}$ képletében. Fordítva ez nem teljesül: a (10.26)-ban szereplő $\frac{\partial f}{\partial t}(t,z)\frac{\partial f}{\partial y}(t,z)$ és $\left(\frac{\partial f}{\partial y}(t,z)\right)^2 f(t,z)$ tagoknak nincs megfelelőjük (10.25)-ben. Ez azt jelenti, hogy nem tudunk minden h-ban másodrendű tagot helyettesíteni F másodrendű tagjaival. A kapott képlet így csak másodrendű lehet. Próbáljuk meg azért a lehető legtöbb másodrendű tagot előállítani. Olyan paramétereket keresünk, amelyeknél a (10.25) és (10.26) nullad- és elsőfokú tagjai megegyeznek, azaz:

$$\gamma_1+\gamma_2=1,\qquad \gamma_2\alpha_1=\frac{1}{2},\qquad \gamma_2\beta_{11}=\frac{1}{2}, \tag{10.27}$$

valamint a megfelelő másodrendű tagok együtthatói is megegyeznek:

$$\frac{\gamma_2}{2}\alpha_1^2=\frac{1}{6},\qquad \gamma_2\alpha_2\beta_{11}=\frac{1}{3},\qquad \frac{\gamma_2}{2}\beta_{11}^2=\frac{1}{6}. \tag{10.28}$$

Látható, hogy például $\gamma_1=\gamma_2=1/2$, $\alpha_1=\beta_{11}=1$ paraméterek megoldásai (10.27)-nek, de nem teljesítik a (10.28) egyenleteket. Viszont mivel a Taylor-módszer minden legfeljebb elsőrendű tagját visszakapjuk, így a felező-módszerhez hasonlóan belátható, hogy másodrendű módszert kapunk. Ezt a

$$z_{i+1}=z_i+\frac{h}{2}\Big(f(t_i,z_i)+f(t_{i+1},z_i+hf(t_i,z_i))\Big),\qquad i=0,1,2,\ldots,\qquad z_0=y_0 \tag{10.29}$$

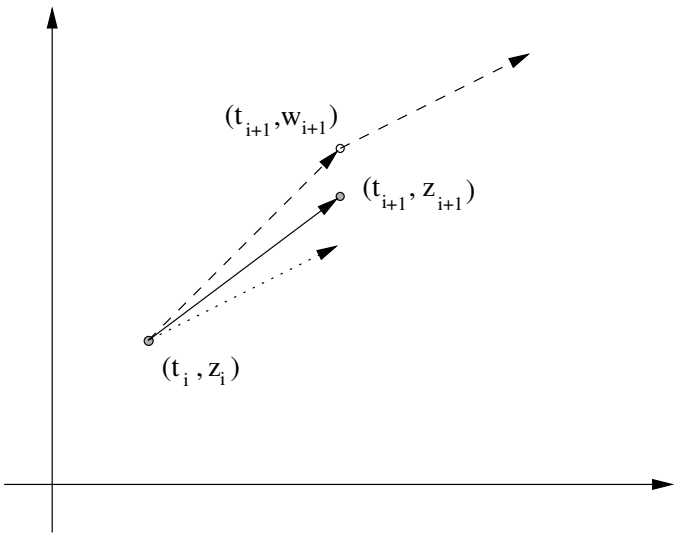
formulával definiált módszert *módosított Euler-módszernek* nevezzük.

Ha a paramétereknek a $\gamma_1=1/4$, $\gamma_2=3/4$ és $\alpha_1=\beta_{11}=2/3$ értékeket választjuk, akkor mind a (10.27) és (10.28) egyenletek teljesülnek. Az ehhez tartozó módszer, az ún. *Heun-módszer* definíciója tehát:

$$\begin{aligned}z_{i+1} &= z_i+\frac{h}{4}\left(f(t_i,z_i)+3f\left(t_i+\frac{2h}{3},z_i+\frac{2}{3}hf(t_i,z_i)\right)\right),\quad i=0,1,2,\ldots,\\ z_0 &= y_0.\end{aligned} \tag{10.30}$$

Mindkét módszer ún. másodrendű Runge–Kutta-képlet (mivel másodrendű lokális képlethibával rendelkeznek).

A módosított Euler-módszerhez is rendelhetünk geometriai tartalmat: tegyük fel, hogy az i-edik lépésben már kiszámítottuk a (t_i,z_i) pontot. Ha az Euler-lépéssel folytatnánk a generálást, akkor az $f(t_i,z_i)$ iránytangensű egyenes mentén a (t_{i+1},w_{i+1}) pontba lépnénk tovább, ahol $w_{i+1}:=z_i+hf(t_i,z_i)$. Ehelyett vesszük ebben a pontban is a ponton áthaladó pontos megoldás irántangensét, $f(t_{i+1},w_{i+1})$-et, és képezzük a $(f(t_i,z_i)+f(t_{i+1},w_{i+1}))/2$ átlagos iránytangenst, és az ez által meghatározott irányban lépünk (t_i,z_i)-ből a t_{i+1} első koordinátájú pontba. Lásd a 10.1. ábrát!



10.1. ábra. A módosított Euler-módszer geometriai interpretációja

Az eddig megadott néhány képlethez hasonló módon levezethető számos más Runge-Kutta típusú módszer. Belátható, hogy a különböző p értékekhez tartozó (10.24) képletekkel definiált Runge–Kutta-módszerekkel a következő maximális rendű lokális képlethibákat lehet elérni:

p	1	2	3	4	5	6	7	8	9	10
a módszer maximális rendje	1	2	3	4	4	5	6	6	7	7

Az egyik legnépszerűbb (10.24) típusú módszer, a „klasszikus” Runge–Kutta-módszer definíciója:

$$\begin{aligned} z_0 &= y_0, \\ w_{i,1} &= f(t_i, z_i), \\ w_{i,2} &= f\left(t_i + \frac{h}{2}, z_i + \frac{h}{2} w_{i,1}\right), \\ w_{i,3} &= f\left(t_i + \frac{h}{2}, z_i + \frac{h}{2} w_{i,2}\right), \\ w_{i,4} &= f\left(t_{i+1}, z_i + h w_{i,3}\right), \\ z_{i+1} &= z_i + \frac{h}{6}(w_{i,1} + 2w_{i,2} + 2w_{i,3} + w_{i,4}), \qquad i = 0, 1, 2, \ldots. \end{aligned} \tag{10.31}$$

Ez a módszer negyedrendű lokális képlethibával rendelkezik (feltéve, hogy $f \in C^5$). A módszer levezetését és a képlethiba rendjének bizonyítását itt nem közöljük.

10.9. példa. A (10.8) feladatra alkalmaztuk a módosított Euler-, Heun- és a klasszikus negyedrendű Runge–Kutta-módszereket a $h = 0.2$-es lépésközt használva. A kapott numerikus eredmények a 10.4. táblázatban találhatók. □

Feladatok

1. Ismételje meg a 10.2. szakasz 1. feladatát felezőpont-, módosított Euler-, Heun- és a klasszikus negyedrendű Runge–Kutta-módszereket használva!

10.4. táblázat. Runge–Kutta-módszerek

		módosított Euler		Heun		klasszikus	
t_i	$y(t_i)$	z_i	$\lvert y(t_i)-z_i\rvert$	z_i	$\lvert y(t_i)-z_i\rvert$	z_i	$\lvert y(t_i)-z_i\rvert$
0.0	1.0000	1.0000	0.0000e-01	1.0000	0.0000e-01	1.0000	0.0000e-01
0.2	1.5082	1.5005	7.6753e-03	1.5042	3.9753e-03	1.5082	1.1773e-05
0.4	2.1745	2.1570	1.7415e-02	2.1663	8.2078e-03	2.1744	2.6024e-05
0.6	2.8799	2.8505	2.9398e-02	2.8679	1.1995e-02	2.8798	4.2338e-05
0.8	3.4470	3.4035	4.3486e-02	3.4331	1.3882e-02	3.4469	5.9304e-05
1.0	3.6109	3.5521	5.8862e-02	3.5998	1.1100e-02	3.6109	7.3610e-05

2. Bizonyítsa be, hogy ha f Lipschitz-tulajdonságú a második változójában, akkor a felezőpont-módszert definiáló
$$F(t,z;h)=\frac{1}{2}f\left(t+\frac{h}{2},z+\frac{h}{2}f(t,z)\right)$$
függvény is Lipschitz-tulajdonságú a második változójában.
3. Az Euler-módszer 3. levezetéséhez hasonló módon vezesse le a (10.29) képletet!
4. Mutassa meg, hogy felezőpont-, módosított Euler- és a Heun-módszer minden lépésköz esetén ugyanazt a közelítő megoldást generálja az
$$y'=2-t-y,\qquad y(0)=1$$
kezdeti érték problémára!
5. Keressen geometriai jelentést a klasszikus negyedrendű Runge–Kutta-módszerhez!
6. Igazolja, hogy ha f csak t-től függ, akkor a klasszikus negyedrendű Runge–Kutta-módszer a Simpson-féle kvadratúra formulára redukálódik!
7. Fogalmazza meg a klasszikus negyedrendű Runge–Kutta-módszert differenciálegyenlet-rendszerekre!
8. Oldja meg a 10.2. szakasz 3. és 4. feladataiban szereplő kezdeti érték problémákat negyedrendű Runge–Kutta-módszerrel!

Irodalomjegyzék

[1] K. E. Atkinson, An Introduction to Numerical Analysis, Wiley, New York, 1978.

[2] R. L. Burden, J. D. Faires, Numerical Analysis, Brooks/Cole, Cengage Learning, 2011.

[3] J. E. Dennis Jr., R. B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentica-Hall, Englewood Cliffs, 1983.

[4] Dringó László, Numerikus Analízis I.–II., Tankönyvkiadó, Budapest, 1978.

[5] Stoyan Gisbert, Takó Galina, Numerikus módszerek I.–II., ELTE – TypoTeX, Budapest, 1993, 1995.

[6] E. Isaacson, H. B. Keller, Analysis of Numerical Methods, Wiley, New York, 1966.

[7] Móricz Ferenc, Bevezetés a numerikus matematikába, Polygon Jegyzettár, Szeged, 2008.

[8] Móricz Ferenc, Differenciálegyenletek numerikus módszerei, Polygon Jegyzettár, Szeged, 1998.

[9] Móricz Ferenc, Numerikus módszerek az algebrában és analízisben, Polygon Jegyzettár, Szeged, 1997.

[10] Popper György, Csizmás Ferenc, Numerikus módszerek mérnököknek, Akadémiai Kiadó – TypoTeX, Budapest, 1999.

[11] A. Ralston, Bevezetés a numerikus analízisbe, Műszaki Könyvkiadó, Budapest, 1969.

[12] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Springer-Verlag, New York, 1980.

[13] Virágh János, Numerikus matematika, JATE Press, Szeged, 2003.

Név- és tárgymutató