Numerikus módszerek - jegyzet

Kupán Pál

Tartalomjegyzék

1. FEJEZET

Számábrázolás, hibaszámítás.

1.1. Számítógépes számábrázolás

Tradicionálisan az emberi gondolkodás a számokat tízes (decimális) számrendszerben kezeli-értelmezi. Például $1205 = 1000 + 200 + 0 + 5 = 1 \cdot 10^3 + 2 \cdot 10^2 + 0 \cdot 10^1 + 5 \cdot 10^0$. Általánosabban, ha r jelöli a számrendszer alapját akkor egy a szám felírása a következő:

$$a = a_m \cdot r^m + a_{m-1} \cdot r^{m-1} + ...,$$

ahol $a_i \in \{0, 1, 2, ..., r-1\}$.

A számítógépek bináris, oktális, illetve hexadecimális számrendszert használnak.

Számrendszer	Alap=r	Szimbólumok=a_i
decimális	10	$0, 1, 2, 3, 4, 5, 6, 7, 8, 9$
bináris	2	$0, 1$
oktális	8	$0, 1, 2, 3, 4, 5, 6, 7$
hexadecimális	16	$0, 1, 2, \ldots, 8, 9, A\,(=10), B\,(=11), C\,(=12), D, E, F$

1. Példa.

$$10010110101_{(2)} = 1 \cdot 2^0 + 0 \cdot 2^1 + 1 \cdot 2^2 + 0 \cdot 2^3 + 1 \cdot 2^4 + 1 \cdot 2^5 + \\ + 0 \cdot 2^6 + 1 \cdot 2^7 + 0 \cdot 2^8 + 0 \cdot 2^9 + 1 \cdot 2^{10} = 1205_{(10)}$$

$$2265_{(8)} = 5 \cdot 8^0 + 6 \cdot 8^1 + 2 \cdot 8^2 + 2 \cdot 8^3 = 1205_{(10)}$$

$$4B5_{(16)} = 5 \cdot 16^0 + B \cdot 16^1 + 4 \cdot 16^2 = 5 \cdot 16^0 + 11 \cdot 16^1 + 4 \cdot 16^2 = 1205_{(10)}$$

Decimális számrendszerből egy r számrendszerbe való áttérést maradékos osztással kapjuk meg (mod r). Az első maradék a szám legkisebb helyiértékű számjegyet adja, stb. Például decimális számrendszerből binárisba való átszámításkor a keresett szám számjegyei a kettővel

való osztási maradékok, melyeket fordított sorrendbe olvasunk. Pl. $1205_{(10)} = 10010110101_{(2)}$ mert

osztó	maradék
1205	1
602	0
301	1
150	0
75	1
37	1
18	0
9	1
4	0
2	0
1	1↑

Az osztást addig végezük amíg a hányados nulla.

Hasonlóan $1205_{(10)} = 2265_{(8)}$ mert

osztó	maradék
1205	5
150	6
18	2
2	2 ↑

, $1205_{(10)} = 4B5_{(16)}$ mert

osztó	maradék
1205	5
75	11=B
4	4 ↑

.

Az oktális, illetve hexadecimális számrendszerek előnye a tömörség, vagyis egy szám ábrázolásához kevesebb szimbólumra van szükség. Ugyanakkor, a kettes számrendszerből nagyon könnyű az áttérés az oktális ($2^3 = 8$), illetve hexadecimális ($2^4 = 16$) számrendszerbe. Egy bináris szám oktális alakját úgy kapjuk meg, hogy a bináris szám számjegyeit hármasával csoportosítjuk (jobbról), majd a megfelelő oktális értéket

$*_{(2)}$	$*_{(8)}$
000	0
001	1
010	2
011	3
100	4
101	5
110	6
111	7

rendeljük hozzá. Például $10\underbrace{010}\underbrace{110}\underbrace{101}_{(2)} = 2265_{(8)}$. A bináris-hexadecimális átalakítás hasonlóan történik, csak a bináris számjegyeket négyesével csoportosítjuk, például $100\underbrace{1011}\underbrace{0101}_{(2)} = 4B5_{(16)}$.

Tizedes alakú számok átalakítása hasonlóan történik:

$$12.34_{(10)} = 1 \cdot 10^1 + 2 \cdot 10^0 + 3 \cdot 10^{-1} + 4 \cdot 10^{-4},$$
$$110.01_{(2)} = 1 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} = 6.25_{(10)}$$

A valós szám egész részének decimális számrendszerből binárisba már láttuk, hogy osztással történik. A szám tizedes részét osztással kapjuk, például 0.25 bináris alakja:

2. Példa.

$\times 2$	egész rész
0.25	$0 \downarrow$
0.5	1
0	

3. Példa. A $\frac{1}{10} = 0.1_{(10)}$ bináris alakja:

$$\frac{1}{10} = 0.000110011001\ldots = 1 \cdot 2^{-4} + 1 \cdot 2^{-5} + 0 \cdot 2^{-6} + 0 \cdot 2^{-7} + 1 \cdot 2^{-8} + 1 \cdot 2^{-9} + \ldots,$$
$$\frac{1}{10} = 1.10011001\ldots \cdot 2^{-4} = 2^{-4}\left(1 + 2^{-1} + 2^{-4} + 2^{-5} + \ldots\right)$$

Egy másik különbség az emberi, illetve a számítógép számábrázolása között a valós számok ábrázolásánál jelenik meg, nevezetesen az emberek fixpontosan értelmezik a számokat, a számítógépben pedig lebegőpontosan. A fixpontos módszert a gépek csak egész számok ábrázolására használják.

Egy valós (tizedes)

Ha egy a szám nem fér el a fixpontos ábrázolási tartományban akkor át kell térni lebegőpontos számábrázolásra. Egy decimális szám lebegőpontos felírása a következőképpen lehetséges:

$$a = (-1)^s \cdot 0.f \cdot r^e \tag{1.1.1}$$

ahol s =előjel (sign), f =a szám törtrésze (mantissza) normalizált alakban, r =a számrendszer alapja, e =a szám kitevője (exponent).

Például a $-1234.5 = (-1)^1 \cdot 0.12345 \cdot 10^4$ szám egy 8 bites regiszteren a következőképpen ábrázolható:

1	0	4	1	2	3	4	5

.

Tehát a 8 bitből egy bit a szám előjele (az első 1 vagyis negatív), két bit a kitevő (04), őt bit pedig a mantissza (12345). Mivel a kitevőt két biten ábrázoljuk az általa leírt tartomány $0 : 99$. További bitre volna szükség a kitevő előjelére. Ezt egyszerű eltolással megoldható, vagyis a kitevőhöz (04-hez) hozzáadunk 50-et, a kitevő tartomány felét, így az előbbi szám a következőképpen ábrázolható:

1	5	4	1	2	3	4	5

.

Tehát a kitevőt úgy számítjuk ki, hogy a regiszterben szereplő értékből kivonunk 50-t.

4. Példa. A

0	5	5	5	4	3	2	1

nyolc bites ábrázolás a $+0.54321 \cdot 10^{55-50} = 54321$ számot jelöli, míg

1	4	8	2	4	6	8	0

a $-0.24680 \cdot 10^{48-50} = -0.0024680$ számot.

A 357.02468 szám 8-bites ábrázolásához a következő lépéseket végezzük:

- normál alakba írjuk: $0.35702468 \cdot 10^{-3}$;
- a mantisszát öt jegyűre vágjuk: $0.35702 \cdot 10^{-3}$;
- ábrázoljuk:

0	4	7	3	5	7	0	2

.

A számítógépekben a valós számok ábrázolása hasonlóan történik a leírtakkal, csak nem tízes- hanem kettes számrendszerben, és nem 8 biten hanem (általában) 32 vagy 64. Kettes számrendszerben a szám lebegőpontos ábrázolása a következő:

$$a = (-1)^s \cdot (1+f) \cdot 2^e = (-1)^s \cdot 1.f \cdot 2^e. \tag{1.1.2}$$

ahol az s, f, e paraméterek jelentősége ugyanaz mint a (1.1.1) képletben.

5. Példa. $3.55_{(10)} = 11.10001100110011... = (-1)^0 \cdot 1.110001100110011... \cdot 2^1$

Ahhoz, hogy különböző processzorokon végzett számítások eredményei ne térjenek el egymástól számottevően, a lebegőpontos ábrázolást 1985-ben szabványosították (IEEE Standards 754).

A 64 bites regiszter esetében (double precision) egy bit a szám előjelét tárolja, 11 a kitevőt és 52 a mantisszát:

s	e	e	e	e	e	e	e	e	e	e	e	f	f	f	f	f	f	f	f	f	f	f	f	...

Mivel a kitevő tartománya 0 és $2^0 + 2^1 + ...2^{10} = 2^{11} - 1$ között van, a kitevő előjele szabályozható ha eltoljuk a tartomány felével vagyis 1023-mal. A binárisan ábrázolt mantisszát mindig 1-re normalizáljuk. Innen következik, hogy még egy bitet spórolhatunk meg ha ezt az 1-t nem ábrázoljuk (implicit-bites ábrázolás).

6. Példa. A $11.34_{(10)} = 1011.010101110..._{(2)} = 1.011010101110... \cdot 2^3$ szám előjele $+$ $(s = 0)$, a kitevője $3+1023 = 1026_{(10)} = 10000000010_{(2)}$, és a mantisszája 011010101110..., tehát a szám ábrázolása:

0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0	1	0	1	1	1	0	...

7. Példa. Az

1	1	0	0	0	0	0	0	0	1	0	1	1	1	1	0	1	1	0	1	1	1	0	0	...

ábrázolás esetén az előjel negatív $(s = 1)$, a kitevő $e = 10000000101_{(2)} =$

$2^{10} + 2^2 + 2^0 = 1029_{(10)}$, a mantissza $1.f = 1.111011011100...$ tehát az adott bináris szám $=-1.111011011100... \cdot 2^{1029-1023} = -123.45$ számot jelöli.

A decimális $a = 1$ felírása binárisan a következő:

$$1.0\ldots00 = 1 \cdot 2^0 + 0 \cdot 2^{-1} + 0 \cdot 2^{-2} + \ldots + 0 \cdot 2^{-52}.$$

A következő ábrázolható számot úgy kapjuk, hogy az utolsó bitet egyesre cseréljük. Az így előállított növekmény nagysága 2^{-52}. Ezzel a növekménnyel generálható a $[2^0, 2^1]$ bináris intervallum összes ábrázolható száma.

Gépi epszilonnak nevezzük az $a = 1$ és a rákövetkező ábrázolható szám különbségét:

$$eps = 2^{-52} \simeq 2.2 \cdot 10^{-16}\ _{(10)}$$

8. Példa. A gépi epszilonnak a generálására a következő műveletsort lehet használni:

$$a = 4/3,\ b = a - 1,\ c = 3 \cdot b,\ d = 1 - c.$$

Pontos számítással a d értéke nulla, viszont a közelítések miatt $d = eps$.

Különböző $[2^e, 2^{e+1}]$ intervallumokon a növekmények változnak. Például a $[2^1, 2^2]$ intervallumban, a 2 felírása:

$$2 = 10_{(2)} = 1.0\ldots00 \cdot 2^1 = 2^1 \cdot \left(1 \cdot 2^0 + 0 \cdot 2^{-1} + 0 \cdot 2^{-2} + \ldots + 0 \cdot 2^{-52}\right),$$

tehát az utolsó bit cseréjével a $[2^1, 2^2]$ intervallumban az ábrázolható számok 2^{1-52} növekménnyel generálhatóak. Általánosan, a $[2^e, 2^{e+1}]$ intervallumon a növekmény 2^{e-52}.

A kitevő (eltolt) két szélsőértéke: -1023 és 1024, különös jelentőséggel bírnak, ezért ezeket csak sajátos esetben használjuk. Tehát a számok ábrázolásához a kitevő használható tartománya:

$$-1022 \leq e \leq 1023.$$

Innen következik, hogy a legkisebb (pozitív) ábrázolható számot (jel. *realmin*) az $e = -1022$ és $f = 0$ értékekre kapjuk, vagyis

$$realmin = 2^{-1022}{}_{(2)} \approx 2.2251 \cdot 10^{-308}{}_{(10)}.$$

A legnagyobb számot $e = +1023$ és $f = 0.11\ldots1$ értékekre kapjuk, vagyis:

$$realmax = (2 - eps)\, 2^{1023} \approx 1.7977 \cdot 10^{308}{}_{(10)}.$$

Túlcsordulásról beszélünk ha a számítások során egy eredmény nagyobb *realmax*-nál, illetve alul-csordulásról ha az eredmény kisebb *realmin*-nél.

9. Példa. Ha egy vizsga végső jegye az írásbeli kétharmadából, illetve az szóbeli harmadából áll és a jegyek 7.75 illetve 7, akkor számítsuk ki (a megszokott kerekítéssel) a vizsga jegyet!

A pontos megoldás

$$\frac{2}{3} \cdot 7.75 + \frac{1}{3} \cdot 7 = \frac{1}{3} \cdot \frac{45}{2} = \frac{15}{2} = 7.5,$$

tehát kerekítve a vizsga jegye 8. Számítógéppel viszont az eredmény 7.4999999, tehát kerekítve 7.

1.2. A hibaszámítás alapfogalmai

A gyakorlatban, a műveleteket általában a számok közelítő értékével végezzük, és az eredmény is közelítő értékű lesz. Lényeges követelmény, hogy mindig becslést tudjunk adni a közelítés pontosságáról.

Hibaforrás lehet már a bemenő adat; ez bekövetkezhet a hibás mérés, vagy a pontatlan műszerek miatt.

1.2.1. A közelítő érték és hibája Tekintsük az A pontos értéket. A gyakorlatban A helyett ennek egy *közelítő érték*ével dolgozunk, jelöljük a-val.

10. Példa. A π közelítő értékének az $a = 3.14$ értéket lehet tekinteni.

A pontos es a közelítő érték abszolút különbségét az *a szám abszolút hibáj*ának nevezzük: $|A - a| = \Delta$.

Mivel a pontos érték legtöbbször ismeretlen ezért az abszolút hiba is ismeretlen. Helyette egy α legkisebb *hibakorlátot* fogunk használni:

$$\Delta = |A - a| \leq \alpha.$$

A fenti képlet felírható a következőképpen:

$$a - \alpha \leq A \leq a + \alpha.$$

11. Példa. Ha egy tetszőleges mérésnél egy bizonyos 0.1 grammért szavatolunk úgy 0.1 gramm a hibakorlát ($A = a \pm 0.1$).

Az abszolút hiba es a hibakorlát a mérés és számolás pontosságának a jellemzésére nem elegendő. Például ha mérések alapján $L_1 = 145cm \pm 0.1cm$, $L_2 = 6.2cm \pm 0.1cm$, habár a két hibakorlát azonos, az első mérés minősége jobb, mint a másodiké.

Az A érték *relatív (viszonylagos) hiba* értelmezése a következő:

$$\nabla = \frac{\Delta}{|a|} = \frac{|A - a|}{|a|}.$$

A gyakorlatban, hasonlóan az abszolút hibához, nem a relatív hibát vizsgáljuk hanem annak a legkisebb korlátját. Ezt nevezzük *relatív hibakorlátnak* és jelöljük δ-val:

$$\frac{\Delta}{|a|} = \nabla \leq \delta\,.$$

Ismerve az abszolút hibakorlátot mivel $\Delta \leq \alpha$ felírhatjuk:

$$\delta = \frac{\alpha}{|a|}.$$

12. Példa. Az előbbi példát felhasználva $\delta_1 = \frac{0.1}{145} = 6.896\,6 \times 10^{-4}$, $\delta_2 = \frac{0.1}{6.2} = 1.612\,9 \times 10^{-2}$.

A továbbiakban az abszolút, illetve relatív hibát Δ, illetve δ-vel azonosítjuk.

1.2.2. A helyes jegyek száma A számok relatív hibája könnyen kiszámítható ha ismert e számok tizedes alakja illetve a helyes jegyek száma.

*Első értékes számjegy*nek nevezzük az első nem nulla számjegyet amit balról jobbra haladva találunk.

13. Példa. A $3.14 = 3 \times 10^0 + 1 \times 10^{-1} + 4 \times 10^{-2}$ számnak az első értékes számjegye a 3, a $0.0027 = 2 \times 10^{-3} + 7 \times 10^{-4}$-nek pedig a 2.

Általánosan egy $a = a_m \times 10^m + a_{m-1} \times 10^{m-1} + ...$ számnak az első értékes számjegye a_m. Innen következik, hogy:

$$a_m 10^m \leq a \leq a_m 10^m + 10^m = (1 + a_m)10^m.$$

A tizedestört alakjában felírt a szám valamely meghatározott jegyének helyén álló mindegyik egységnek megvan a maga *helyi-értéke*: az első helyen álló egység 10^m-el egyenlő, a második 10^{m-1} ..., az n-edik helyen álló egység egyenlő 10^{m-n+1}.

14. Példa. A 3.141 számban a harmadik jegy helyi-értéke 10^{-2}, míg a 0.3141 számban ugyanannak a jegynek 10^{-3} helyi értéke van.

Hogy egyszerűbbé tegyünk egy tizedes számot használhatjuk a *csonkítási szabályt*, azaz, elhanyagolhatunk bizonyos számú számjegyet a végéről. Az abszolút hiba, amelyet ebben az esetben elkövetünk, nem fogja meghaladni a meghagyott jegyek utolsójának helyi értékét.

15. Példa. Ha a 2.10527 számnak az első három jegyére szorítkozunk, akkor a 2.10 számot kapjuk, az elkövetett hiba pedig kisebb mint 10^{-2}.

Ha nagyobb pontosságot szeretnénk elérni használhatjuk a *kerekítési szabály*t. Ez abban áll, hogy a meghagyott számjegyek utolsóját változatlanul hagyjuk vagy növeljük egy egységgel attól függően, hogy az első elhanyagolt számjegy kisebb ($<$), vagy nagyobb vagy egyenlő ($\geq$) 5-el. Ebben az esetben, az elkövetett abszolút hiba nem haladja meg a meghagyott jegyek utolsójának helyi értekének a felét.

16. Példa. Ha a 2.10527 számban, a kerekítési szabályt alkalmazva, az első három jegyet tartjuk meg, akkor kapjuk a 2.11 számot, az elkövetett hiba pedig kisebb mint $1/2 \times 10^{-2}$.

Általánosan, ha az A számot az a érték n helyes számjeggyel közelíti meg, akkor az abszolút hibára a következő egyenlőtlenségeket kapjuk:

$$\begin{aligned}\Delta &= |A-a| \leq 10^{m-n+1} \text{ csonkítási szabállyal ill.}\\ \Delta &= |A-a| \leq \frac{1}{2}10^{m-n+1} \text{ kerekítési szabállyal.}\end{aligned}$$

Ezen feltételek mellett kifejezhető természetesen a relatív hiba is:

$$\begin{aligned}\delta &= \frac{10^{m-n+1}}{|a|} \leq \frac{10^{m-n+1}}{a_m 10^m} \implies \delta \leq \frac{1}{a_m 10^{n-1}} \text{ csonkítás ill.}\\ \delta &\leq \frac{1}{2a_m 10^{n-1}} \text{ kerekítés esetben.}\end{aligned}$$

Nagy vagy igen kicsi számokat ajánlatos 10-nek a hatványai szerint írni.

17. Példa. A 12345678 számot három helyes számjeggyel a következőképpen írjuk: 123×10^5. Hasonlóan a 0.000001234 számot három helyes jeggyel $=12.3 \times 10^{-7}$.

A problémát fordítva is meg lehet fogalmazni, vagyis egy adott pontosságra hány helyes számjegyet kell számításba venni.

1.2.3. Hibaterjedés

1.2.3.1. Az alapműveletek hibái Legyen a_1, a_2 az A_1, A_2 pontos számok közelítő értékei. Ismerjük továbbá az eltérés hibakorlátjait: α_1, α_2 illetve δ_1, δ_2. Szeretnénk meghatározni a hibakorlátokat abban az esetben mikor az A_1, A_2 értékekkel a négy alapműveletet végezzük.

- Az összeg hibakorlátja

Ebben az estben az $A = A_1 + A_2$ mennyiség megközelítéséül szolgaló $a = a_1 + a_2$ összeg α abszolút hibakorlátot határozzuk meg.

$$|A-a| = |A_1 - a_1 + A_2 - a_2| \leq |A_1 - a_1| + |A_2 - a_2| \leq \alpha_1 + \alpha_2$$

vagyis az α hibakorlátot egyenlőnek lehet venni az α_1, α_2 hibakorlátok összegével:

$$\alpha = \alpha_1 + \alpha_2. \tag{1.2.1}$$

- A különbség hibakorlátja

Hasonlóan lehet eljárni mint az előbbi pontnál: $A = A_1 - A_2$, $a = a_1 - a_2 \Longrightarrow$

$$|A - a| = |A_1 - a_1 - A_2 + a_2| \leq |A_1 - a_1| + |A_2 - a_2| \leq \alpha_1 + \alpha_2$$

vagyis, mint az előbb, az α különbség hibakorlátját egyenlőnek lehet venni az α_1, α_2 hibakorlátok összegével:

$$\alpha = \alpha_1 + \alpha_2 \,. \tag{1.2.2}$$

Külön figyelmet kell szentelni ha A_1 es A_2 közeli értékek ugyanis, ebben az esetben sokat veszíthetünk a pontosságból.

18. Példa. Ha $x_{1,2} = 1 \pm \sqrt{1 + 10^{-16}}$ akkor:

(1) adjuk meg azt a másodfokú egyenletet amelynek gyökei x_1, x_2.
(2) számítsuk $\frac{x_1}{x_2}$ arányt

A Viéte összefüggések szerint $S = x_1 + x_2 = 2$, $P = x_1 \cdot x_2 = 0$, $(x_2 = 0!)$ tehát az egyenlet $x^2 - 2x = 0$ aminek a gyökei 0, illetve 2. Egyszerű behelyettesítéssel az $\frac{x_1}{x_2}$ arány inf értéket fog mutatni miközben a pontos értéke $\frac{1+\sqrt{1+10^{-16}}}{1-\sqrt{1+10^{-16}}} = \frac{\left(1+\sqrt{1+10^{-16}}\right)^2}{-10^{-16}} = -4 \cdot 10^{16}$.

19. Példa. Számítsuk ki két kocka V_1, V_2 térfogatainak különbségét ha oldalaik $0.120\,m$ ill. $0.121\,m$. A térfogatok négy tizedessel (csonkítás) a következők lesznek: $V_1 = 0.0017$, $V_2 = 0.0017$, ahonnan a különbségük $V_2 - V_1 = 0$. Hogy elkerüljük az értékes tizedesek elhagyását használjuk az $a^3 - b^3 = (a - b)\,(a^2 + a \cdot b + b^2)$ képletet ahol $a = 0.121$ $b = 0.120$, ahonnan

$$V_2 - V_1 = 0.001 \cdot \left(0.121^2 + 0.121 \cdot 0.120 + 0.12^2\right) = 10^{-3} \cdot 0.0308 m^3 = 30.8 cm^3.$$

Egy másik lehetőség az értékes jegyek megtartására a Taylor képlet használata:

$$f(x + \Delta x) - f(x) = f'(x) \cdot \Delta x,$$

ahol $f(x) = x^3$, $x = 0.120$, $\Delta x = 0.001$ azt kapjuk, hogy: $V_2 - V_1 = 3 \cdot 0.120^2 \cdot 0.001 = 0.0432 \cdot 10^{-3}\, m^3 = 43.2\, cm^3$.

- A szorzat hibakorlátja

$$\begin{aligned} |A_1A_2 - a_1a_2| &= |A_1A_2 - A_1a_2 + A_1a_2 - a_1a_2| = |A_1\,(A_2 - a_2) + (A_1 - a_1)\,a_2| \leq \\ &\leq |A_1|\,|A_2 - a_2| + |a_2|\,|A_1 - a_1| \leq |A_1|\,\Delta_2 + |a_2|\,\Delta_1\ . \end{aligned}$$

Az ismeretlen $|A_1|$ tagot majoráljuk a következő taggal:

$$|A_1| = |A_1 - a_1 + a_1| \leq |A_1 - a_1| + |a_1| \leq \Delta_1 + |a_1|$$

tehát $\Delta \leq (\Delta_1 + |a_1|)\,\Delta_2 + |a_2|\,\Delta_1 = \Delta_1\Delta_2 + |a_1|\,\Delta_2 + |a_2|\,\Delta_1$. Elosztva az egyenlőtlenség mindkét oldalát $|a_1a_2|$ taggal kapjuk, hogy:

$$\frac{\Delta}{|a_1a_2|} \leq \frac{\Delta_1\Delta_2}{|a_1a_2|} + \frac{\Delta_1}{|a_1|} + \frac{\Delta_2}{|a_2|}$$

vagyis:

$$\delta = \delta_1\delta_2 + \delta_1 + \delta_2.$$

Mivelhogy a relatív hiba kicsi szokott lenni ezért a $\delta_1\delta_2$ szorzat elhanyagolható és a szorzat relatív hibájának vehetjük a δ_1, δ_2 relatív hibák ősszegét:

$$\delta = \delta_1 + \delta_2. \tag{1.2.3}$$

- A hányados hibakorlátja

$\left|\frac{A_1}{A_2} - \frac{a_1}{a_2}\right| = \left|\frac{A_1a_2 - A_2a_1}{A_2a_2}\right| = \frac{|A_1a_2 - a_1a_2 + a_1a_2 - A_2a_1|}{|A_2a_2|} = \frac{|(A_1 - a_1)a_2 + a_1(a_2 - A_2)|}{|A_2a_2|} \leq \frac{|A_1 - a_1||a_2| + |a_1||a_2 - A_2|}{|A_2||a_2|}$

Az $|A_2|$ ismeretlent minoráljuk a következő képpen:

$|a_2| = |a_2 - A_2 + A_2| \leq |a_2 - A_2| + |A_2| \Longrightarrow\ |A_2| \geq |a_2| - |a_2 - A_2|$ ahonnan $\Delta = \left|\frac{A_1}{A_2} - \frac{a_1}{a_2}\right| \leq \frac{\Delta_1|a_2| + |a_1|\Delta_2}{|a_2|(|a_2| - \Delta_2)}$.

Hogy a relatív hibakorlátot megkapjuk, elosztjuk Δ-t $\left|\frac{a_1}{a_2}\right|$-vel es azt kapjuk, hogy:

$$\frac{\Delta}{\left|\frac{a_1}{a_2}\right|} \leq \frac{\Delta_1\,|a_2| + |a_1|\,\Delta_2}{|a_2|\,(|a_2| - \Delta_2)}\,\frac{|a_2|}{|a_1|} = \frac{\frac{\Delta_1}{|a_1|} + \frac{\Delta_2}{|a_2|}}{1 - \frac{\Delta_2}{|a_2|}}.$$

Figyelembe véve, hogy $\frac{\Delta_2}{|a_2|}$ elhanyagolható 1-hez képest azt kapjuk, hogy:

$$\delta = \delta_1 + \delta_2. \tag{1.2.4}$$

20. Példa. Egy fém egyeneshenger magassága $h = 8.5 \pm 0.01cm$, átmérője $d = 4.04 \pm 0.01cm$. Számítsuk ki a henger térfogatát és becsüljük meg az elkövetett hibát! Ha a henger tömege $m = 2103 \pm 1g$, számítsuk ki a henger sűrűségét és becsüljük meg a hibát!

Bizonyítás. A megadott adatok abszolút és relatív hibái:

$$\begin{aligned} \triangle_h &= 0.01cm, \quad \delta_h = \frac{0.01}{8.5} = 0.0012 = 1.2 \cdot 10^{-3}, \\ \triangle_d &= 0.01cm, \quad \delta_d = \frac{0.01}{4.04} \approx 0.002475 = 2.475 \cdot 10^{-3}, \end{aligned}$$

A

$$V = \pi \frac{d^2}{4} h,$$

térfogat képlethez szükséges π értékét az eddigi értékek közeli nagyságrenddel (csonkítással) közelítjük meg: $\pi \approx 3.141$, ekkor

$$\triangle_\pi = 10^{-3}, \quad \delta_\pi = \frac{0.001}{3.141} = 0.318 \cdot 10^{-3},$$

tehát

$$V = 3.141 \frac{4.04^2}{4} 8.5 \approx 108.94 cm^3,$$

és a hiba:

$$\delta_V = \delta_\pi + 2\delta_d + \delta_h = 6.468 \cdot 10^{-3}, \quad \triangle_V = V \cdot \delta_V = 1.036548744 cm^3.$$

amit $V = 108.94 \pm 1.0365 cm^3$ alakban írhatunk. A sűrűség

$$\rho = \frac{m}{V} = \frac{2103}{108.94} \approx 19.3042 \frac{g}{cm^3} = 19300 \frac{kg}{m^3}.$$

□

1.2.3.2. Függvény hibakorlátja Ebben az esetben szeretnénk megvizsgálni hogyan befolyásolja a változó hibakorlátja a függvény hibakorlátját.

Legyen $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ egy deriválható függvény, illetve $X^0 = (X_1^0, X_2^0, ..., X_n^0) \in \mathbb{R}^n$ egy pontos érték. Kérdés mennyi $f(X^0) =?$

Az X^0 ideális értéket megközelítjük egy $x^0 = (x_1^0, x_2^0, ..., x_n^0) \in \mathbb{R}^n$ ponttal, illetve az $f(X^0)$ értéket az $f(x^0)$ értékkel. A függvény megközelítésében elkövetett abszolút hibát jelöljük $\Delta f(x^0) = |f(X^0) - f(x^0)|$.

A középérték tételből:

$$f\left(X^0\right) - f\left(x^0\right) = df\left(\xi\right)\left(X^0 - x^0\right),\ \xi \in \left(X^0, x^0\right)$$

következik, hogy:

$$\Delta f\left(x^0\right) = \left|\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\left(\xi\right)\left(X_i^0 - x_i^0\right)\right| \leq \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\left(\xi\right)\right| \left|X_i^0 - x_i^0\right| \ .$$

Tételezzük továbbá fel, hogy X^0 pontnak létezik egy olyan Ω környezete amelyre $x^0 \in \Omega$ es f parciális deriváltjai korlátosak, vagyis:

$$\sup_{x\in\Omega} \left|\frac{\partial f}{\partial x_i}\left(\xi\right)\right| = M_i \in \mathbb{R}.$$

Következik, hogy:

$$\left|f\left(X^0\right) - f\left(x^0\right)\right| = \Delta f\left(x^0\right) \leq \sum_{i=1}^{n} M_i \left|X_i^0 - x_i^0\right|.$$

Ha ϵ egy előre megadott pontosság akkor kikötjük, hogy:

$$\left|X_i^0 - x_i^0\right| \leq \frac{\epsilon}{nM_i}, \forall i$$

ahonnan az előbbi egyenlőtlenség alapján kapjuk:

$$\left|f\left(X^0\right) - f\left(x^0\right)\right| \leq \epsilon.$$

Ha $f\left(x\right) = x_1 + x_2$, akkor $M_i = 1$ és visszakapjuk az összegre vonatkozó hibakorlátot, $x = (x_1, x_2) \in \mathbb{R}^2$. Hasonlóan, visszakapjuk a hibabecslést a kivonásra, szorzásra, osztásra ha $f(x)$ rendre $f(x) = x_1 - x_2$, $f(x) = x_1 x_2$, $f(x) = \frac{x_1}{x_2}$.

21. Példa. Legyen $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, $f\left(x\right) = x_1^2 + 2x_2^2 + x_1 x_2$, ahol $x = (x_1, x_2) \in \mathbb{R}^2$.

$$\Delta f\left(x^0\right) = M_1 \Delta x_1 + M_2 \Delta x_2$$

ahol $M_1 = \sup\left|\frac{\partial f}{\partial x_1}\right| = \sup\left|2x_1 + x_2\right|$, $M_2 = \sup\left|\frac{\partial f}{\partial x_2}\right| = \sup\left|4x_2 + x_1\right|$. Konkrétan ha $x^0 = (0,0)$, $X^0 = (0.1, -0.025)$, $\Omega = [-0.1, 0.1] \times [-0.025, 0.025]$, akkor $M_1 = \sup_{x\in\Omega}\left|2x_1 + x_2\right| = 0.225$, $M_2 = \sup_{x\in\Omega}\left|4x_2 + x_1\right| =$

0.2 ahonnan

$$\Delta f\left(x^0\right) = 0.225 \cdot 0.1 + 0.2 \cdot 0.025 = 0.027\,5.$$

2. FEJEZET

Számsorok

22. Definíció. Számsornak nevezzük az alábbi végtelen összeget:

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + ... + a_n + ... \tag{2.0.1}$$

A sor természetét -konvergencia, divergencia- az $(S_n)_n$ részletösszeg sorozat segítségével tanulmányozzuk

$$S_n = a_1 + a_2 + ... + a_n. \tag{2.0.2}$$

Ha az $(S_n)_n$ sorozatnak van véges $S = \lim_{n\to\infty} S_n$ határértéke, akkor a $\sum_{n=1}^{\infty} a_n$ sort konvergensnek nevezzük és a sor összege egyenlő S-el; ha viszont ez a határérték végtelen, vagy a sorozatnak nincs határértéke, akkor a sort divergensnek nevezzük.

Könnyen belátható hogy ha az $(a_n)_n$ sorozat nem konvergál zéróhoz akkor a sor divergens, tehát szükséges (de nem elégséges) feltétele annak hogy a $\sum_{n=1}^{\infty} a_n$ sor konvergens legyen az hogy:

$$\lim_{n\to\infty} a_n = 0. \tag{2.0.3}$$

23. Példa. A

$$0.01 + \sqrt[2]{0.01} + ... + \sqrt[n]{0.01} + ...$$

sor általános tagja $a_n = \sqrt[n]{0.01} \to 1 \neq 0$, tehát a sor divergens.

24. Példa. Az úgynevezett harmonikus sor esetén:

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + ...,$$

annak ellenére hogy $a_n \to 0$, a sor divergens. Az általánosított harmonikus sor:

$$\sum_{n=1}^{\infty} \frac{1}{n^{\alpha}} \tag{2.0.4}$$

konvergens ha $\alpha > 1$ és divergens ha $\alpha \leq 1$.

A konvergens sorok összegének a megközelítéséhez a sor első n tag összegét használjuk:

$$\sum_{n=1}^{\infty} a_n \simeq a_1 + a_2 + ... + a_n, \tag{2.0.5}$$

és ekkor az elkövetett hiba a következő lesz:

$$R_n = |a_{n+1} + a_{n+2} + ...| \tag{2.0.6}$$

Természetesen, minél több tagot veszünk figyelembe a (2.0.5) közelítéshez, annál pontosabb eredményt kapunk, de az összegezendő tagok számát úgy fogjuk meghatározni hogy az abszolút hiba egy bizonyos $\epsilon > 0$ ponotossági küszöbön belül maradjon

$$R_n < \epsilon. \tag{2.0.7}$$

2.1. Pozitív tagú sorok kiszámítása

Tekintsük a $\sum_{n=1}^{\infty} a_n,\ a_n > 0$, pozitív tagú sort amelyre teljesül az alábbi D'Alembert-féle konvergencia kritérium:

$$\frac{a_{n+1}}{a_n} \leq q < 1,\ \forall n \geq N. \tag{2.1.1}$$

Tehát:

$$a_{N+1} \leq q a_N,\ a_{N+2} \leq q a_{N+1} \leq q^2 a_N,\ ...$$

és az eredeti sort majorálhatjuk a következőképpen:

$$\begin{aligned}
\sum_{n=1}^{\infty} a_n &= a_1 + ... + a_N + a_{N+1} + ... \leq a_1 + ... + a_N + q a_N + q^2 a_N + ... \\
&= a_1 + ... a_{N-1} + a_N(1 + q + q^2 + ...) = a_1 + ... a_{N-1} + a_N \lim_n (\frac{1 - q^n}{1 - q}) = \\
&= a_1 + ... a_{N-1} + a_N (\frac{1}{1 - q}),\ \text{mert } q \in (0, 1)\,.
\end{aligned}$$

Tehát, a sor összegének a közelítésére használjuk a

$$\sum_{n=1}^{\infty} a_n \simeq a_1 + a_2 + ... + a_{N-1}, \tag{2.1.2}$$

és a hiba

$$R_N = a_N(\frac{1}{1-q}). \tag{2.1.3}$$

Az N küszöbszám kiszámításához megoldjuk a (2.0.7) kikötésből eredő egyenlőtlenséget:

$$a_N(\frac{1}{1-q}) \leq \epsilon. \tag{2.1.4}$$

Mivel numerikusan az (2.1.2) összeget egy DO-WHILE ciklusban számítjuk ki, a (2.1.4) egyenlőtlenség konkrét megoldására nincs szükség, ugyanis az index számot addig növeljük (és összegezünk egyúttal) ameddig az (2.1.4) egyenlőtlenség igaz lesz.

25. Példa. A $\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} \frac{1}{2^n n^2}$ sor esetében:

$$\frac{a_{n+1}}{a_n} = \frac{\frac{1}{2^{n+1}(n+1)^2}}{\frac{1}{2^n n^2}} = \frac{2^n n^2}{2^{n+1}\left(n+1\right)^2} = \frac{1}{2\left(1+\frac{1}{n}\right)^2} \leq \frac{1}{2} = q < 1.$$

Tehát, ha például $\epsilon = 10^{-2}$ akkor a (2.1.4) szerint $\frac{2}{2^N N^2} \leq 10^{-2} \Longrightarrow N = 4$, és a sor közelítő értéke (2.1.2) $\sum_{n=1}^{\infty} \frac{1}{2^n n^2} \simeq \frac{1}{2} + \frac{1}{2^2 2^2} + \frac{1}{2^3 3^2} = 0.576\,39$.

2.2. Váltakozó előjelű sorok

A váltakozó előjelű sorok az alábbi alakban irhatók:

$$\sum_{n=1}^{\infty} (-1)^{n+1} a_n = a_1 - a_2 + a_3 - ... \ , \ a_n > 0 \tag{2.2.1}$$

és esetükben alkalmazhatjuk a Leibniz-féle kritériumot, azaz ha $(a_n)_n$ monoton csökkenő és zéróhoz konvergál, akkor a (2.2.1) sor konvergens.

A sor összegének a közelítéséhez újból az első N tag összegét vesszük

$$\sum_{n=1}^{\infty} (-1)^{n+1} a_n \simeq a_1 - a_2 + a_3 - ... + (-1)^{N+1} \ a_N, \tag{2.2.2}$$

és a hiba:

$$R_N = \left|(-1)^{N+2}\ a_{N+1} + (-1)^{N+3}\ a_{N+2} + ...\right| \tag{2.2.3}$$

Feltételezve hogy teljesül a Leibniz-féle kritérium, a (2.2.3) hiba majorálható az első elhanyagolt taggal (moduluszban)

$$R_N \leq a_{N+1}. \tag{2.2.4}$$

Valóban, ha N páros azt kapjuk hogy

$$\begin{aligned} R_N &= |\ a_{N+1} -\ a_{N+2} + a_{N+3} -\ a_{N+4} + ...| = \\ &= a_{N+1} -\ a_{N+2} + a_{N+3} -\ a_{N+4} + ... \leq a_{N+1}, \end{aligned}$$

ha pedig N páratlan

$$\begin{aligned} R_N &= |\ -a_{N+1} +\ a_{N+2} - a_{N+3} +\ a_{N+4} + ...| = \\ &= a_{N+1} -\ a_{N+2} + a_{N+3} -\ a_{N+4} + ... \leq a_{N+1}. \end{aligned}$$

A fenti képletekben felhasználtuk hogy $(a_n)_n$ monoton csökkenő: $a_{N+1} \geq a_{N+2} \geq a_{N+3} \geq$ ....

Az összegezendő tagok számát N-et, újból a $\epsilon > 0$ pontosságtól tesszük függővé, vagyis kikötjük hogy

$$R_N \leq a_{N+1} \leq \epsilon, \tag{2.2.5}$$

bebiztosítva ezáltal hogy az összegzésből eredő hiba nem lépi túl a megengedett határt.

26. Példa. Tekintsük a Leibniz sort

$$1 - \frac{1}{2} + \frac{1}{3} - ... = \sum_{n=1}^{\infty} (-1)^{n+1}\, \frac{1}{n}.$$

Az $a_n = \frac{1}{n}$ sorozat telejesti a Leibniz feltételeket, tehát a sor konvergens. A sor összegének $\epsilon = 10^{-2}$ pontossággal való kiszámításához felhasználjuk a (2.2.5) kikötést

$$\frac{1}{N+1} \leq \epsilon,$$

ahonnan $N = 99$. Tehát

$$\sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n} \simeq S_{99} = 1 - \frac{1}{2} + \frac{1}{3} - ... + \frac{1}{99} = 0.6982$$

$\epsilon = 10^{-2}$ pontossággal. Ismerve, hogy a sor pontos összege=$\ln 2$, kiszámítható a hiba: $R_N = |\ln 2 - 0,6982| \simeq 5 * 10^{-3} < \epsilon$.

A fenti összegzést egy FOR ciklusban oldjuk meg; ugyanakkor, ha egy DO WHILE ciklust használunk akkor a (2.2.5) egyenlőtlenséget nem kell előzetesen megoldani mert az összegzést addig folytatjuk míg ez teljesül (tehát egyesével lépegetve csak leellenőrizzük).

27. DEFINÍCIÓ. Egy $\sum_{n=1}^{\infty} a_n$ sort abszolút konvergensnek nevezünk ha $\sum_{n=1}^{\infty} |a_n|$ sor konvergens.

A pozitív tagú sorok esetében az abszolút konvergenciának semmi jelentősége nincs, viszont a váltakozó sorok esetében ha egy sor abszolút konvergens akkor az konvergens is lesz. A fordított állítás nem igaz. Például a Leibniz-féle sor $\sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n}$ konvergens, de $\sum_{n=1}^{\infty} \left|(-1)^{n+1} \frac{1}{n}\right| = \sum_{n=1}^{\infty} \frac{1}{n}$ divergens.

2.3. A sorok konvergenciájának a javítása

A sorok összegének a kiszámításánál a pontos közelítés mellett fontos a gyorsaság is. A fenti példán láttuk hogy a Leibniz sor $\epsilon = 10^{-2}$ pontosságú kiszámításához $N = 99$ tagot kellett összegezni. A pontosság növelésével a N is arányosan nő. Például ha $\epsilon = 10^{-9}$ akkor $N = 10^{-9} - 1$. Ahhoz hogy ugyanazt az eredményt kevesebb tag összegezésével érjük el szükséges az eredeti sor átalakítása.

28. DEFINÍCIÓ. Azt mondjuk hogy a $\sum_{n=1}^{\infty} b_n$ sor gyorsabban konvergál mint a $\sum_{n=1}^{\infty} a_n$ sor ha

$$\lim_{n\to\infty} \frac{b_n}{a_n} = 0. \tag{2.3.1}$$

2.3.1. Az Euler transzformáció A váltakozó előjelű sorok esetében

$$S = \sum_{n=1}^{\infty} (-1)^{n+1} a_n = a_1 - a_2 + a_3 - ..., \ a_i > 0$$

az S_n részletösszeg helyett használjuk ezek középértékét:

$$T_n = \frac{S_{n-1} + S_n}{2} \tag{2.3.2}$$

és így a következő sorhoz jutunk:

$$T = \frac{a_1}{2} - \frac{a_2 - a_1}{2} + \frac{a_3 - a_2}{2} + ... \tag{2.3.3}$$

E sor összege megegyezik az eredeti sor összegével:

$$T_n - S_n = (-1)^n \frac{a_n}{2} \to 0, \text{ ha } n \to \infty, \text{ tehát } T = S,$$

viszont gyorsabban konvergál:

$$\lim_n \frac{(-1)^{n+1} \frac{a_n - a_{n-1}}{2}}{(-1)^{n+1} a_n} = \frac{1}{2} \lim_n \left(1 - \frac{a_{n-1}}{a_n}\right) = 0.$$

29. Példa. A Leibniz sorra alkalmazzuk az (2.3.3) átalakítást

$$\begin{aligned} \sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n} &= \frac{1}{2} - \frac{\frac{1}{2} - 1}{2} + \frac{\frac{1}{3} - \frac{1}{2}}{2} + ... + (-1)^{n+1} \frac{\frac{1}{n} - \frac{1}{n-1}}{2} + ... \\ &= \frac{1}{2} + \frac{1}{4} - \frac{1}{12} + ... + (-1)^n \frac{1}{2n\,(n-1)} + ... \end{aligned}$$

Figyelembe véve a (2.2.5) hiba majorálási kritériumot:

$$\frac{1}{2N\,(N+1)} \leq \epsilon$$

azt kapjuk hogy a sor $\epsilon = 10^{-2}$ pontossággal való kiszámításához $N = 7$ tagra van szükség és

$$\sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n} = \ln 2 \simeq \frac{1}{2} + \frac{1}{4} - \frac{1}{12} + ... - \frac{1}{84}.$$

Az átalakítás többször is alkalmazható.

2.3.2. A Kummer transzformáció Kiindulva $\sum_{n=1}^{\infty} a_n,\ a_n > 0$ pozitív tagú sorból, válasszunk egy $\sum_{n=1}^{\infty} u_n$ sort melynek u összege ismert és létezik a

$$\lim_n \frac{a_n}{u_n} = \lambda \in \mathbb{R}, \tag{2.3.4}$$

határérték. Ekkor

$$\sum_{n=1}^{\infty} a_n = \lambda \sum_{n=1}^{\infty} u_n + \sum_{n=1}^{\infty} (a_n - \lambda u_n) = \lambda u + \sum_{n=1}^{\infty} (a_n - \lambda u_n). \tag{2.3.5}$$

Tehát, a $\sum_{n=1}^{\infty} a_n$ sor tanulmányozása visszavezethető az alábbi sor elemzésére:

$$\sum_{n=1}^{\infty} (a_n - \lambda u_n), \tag{2.3.6}$$

ami viszont gyorsabban konvergál az eredeti sornál:

$$\lim_n \frac{a_n - \lambda u_n}{a_n} = 1 - \lambda \lim_n \frac{u_n}{a_n} = 0.$$

30. Példa. Alkalmazzuk az átalakítást az alábbi konvergens sorra

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

Segédsornak a $\sum_{n=1}^{\infty} u_n = \sum_{n=1}^{\infty} \frac{1}{n(n+1)}$ sort vesszük aminek az összege 1-el egyenlő és

$$\lambda = \lim_n \frac{a_n}{u_n} = \lim_n \frac{n(n+1)}{n^2} = 1.$$

Tehát, a (2.3.5) képlet szerint

$$\sum_{n=1}^{\infty} a_n = 1 + \sum_{n=1}^{\infty} \left(\frac{1}{n^2} - \frac{1}{n(n+1)} \right) = 1 + \sum_{n=1}^{\infty} \frac{1}{n^2(n+1)}$$

ami nyilván gyorsabban konvergál.

A gyorsító módszerek többször is alkalmazható egymásután.A módszer többször is alkalmazható egymásután.

2.3.3. Függvénysorok, hatványsorok Az olyan sorokat, amelyeknek tagjai egy (vagy több) változó függvényei, függvénysoroknak nevezzük. Jelölésük:

$$\sum_{n=1}^{\infty} f_n(x). \tag{2.3.7}$$

31. Példa. a) $1 + x + x^2 + ... + x^n + ... = \sum_{n=0}^{\infty} x^n$
b) $\sin x + \frac{\sin 2x}{2^2} + ... + \frac{\sin nx}{n^2} + ... = \sum_{n=1}^{\infty} \frac{\sin nx}{n^2}$.

A legfontosabb függvénysorok a hatványsorok:

$$\sum_{n=1}^{\infty} a_n (x - x_0)^n , \tag{2.3.8}$$

vagy $x := x - x_0$ transzformációt használva

$$\sum_{n=1}^{\infty} a_n x^n. \tag{2.3.9}$$

A hatványsorok előnye hogy a részletösszeg sorozat egy polinom függvény: $S_n = \sum_{k=1}^{n} a_k x^k$.

Azt a legnagyobb $(-R, R)$ intervallumot amelyre a (2.3.9) hatványsor abszolút konvergens konvergencia intervallumnak nevezzük, az $R \epsilon \overline{\mathbb{R}}$-et pedig konvergencia sugárnak. Ha $R = 0$ akkor a hatványsor csak $x = 0$-ban konvergens, ha pedig $R = \infty$ akkor az egész $\mathbb{R}$ tengelyen. Ha x eleme a konvergencia intervallumnak akkor a függvénysor egy f függvényt fog előállítani

$$f(x) = \sum_{n=1}^{\infty} a_n x^n,$$

vagyis minden $\epsilon > 0$ számhoz létezik egy $N(\epsilon, x) \in \mathbb{N}$ küszöbszám (ami függ ϵ és x-től) amelyre ha $n > N$, akkor

$$|f(x) - S_n(x)| < \epsilon. \tag{2.3.10}$$

Ha az N küszöbszám nem függ x-től akkor a sorról azt mondjuk hogy egyenletesen konvergál.

32. Tétel. *Ha* $0 < R < \infty$ *akkor a (2.3.9) hatványsor egyenletesen konvergál bármilyen* $[-r, r]$ *intervallumon ahol* $0 < r < R$.

Az R konvergencia sugárnak a kiszámításához alkalmazhatjuk a D'Alembert (hányados), illetve Cauchy (gyök) kritériumot: ha

$$l = \lim_n \left| \frac{a_{n+1}}{a_n} \right|, \tag{2.3.11}$$

vagy

$$l = \lim_n \sqrt[n]{|a_n|}, \tag{2.3.12}$$

határérték létezik, akkor $R = \frac{1}{l}$.

33. TÉTEL. *Ha* $\sum_{n=1}^{\infty} a_n x^n$ *hatványsor konvergencia sugara* R, *akkor* $f(x) = \sum_{n=1}^{\infty} a_n x^n$, $f : (-R, R) \to \mathbb{R}$ *összegfüggvény*
(i) $C^{\infty}(-R, R)$ *osztályú és*

$$a_n = \frac{f^{(n)}(0)}{n!}, \ \forall n \geq 0, \tag{2.3.13}$$

(ii) bármilyen $[a, b] \subset (-R, R)$ *intervallumon*

$$\int_a^b f(x)\, dx = \sum_{n=1}^{\infty} a_n \int_a^b x^n dx, \tag{2.3.14}$$

vagyis a sor tagonként integrálható.

34. PÉLDA. a)

$$1 - x + x^2 - ... = \lim_n \frac{1 - (-x)^n}{1 - (-x)} = \frac{1}{1 + x}, \ \forall x \epsilon (-1, 1).$$

Tagonként integrálva azt kapjuk hogy

$$c + x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - ... = \ln(x + 1)$$

és $x = 0$ -ra következik $c = 0$. Tehát $x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - ... = \ln(x + 1)$ és $x \to 1$-re $1 - \frac{1}{2} + \frac{1}{3} - ... = \ln 2$.
b)

$$1 - x^2 + x^4 - ... = \frac{1}{1 + x^2}, \ \forall x \epsilon (-1, 1).$$

Tagonként integrálva kapjuk hogy $x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - ... = arctgx$, és $x \to 1 \Rightarrow 1 - \frac{1}{3} + \frac{1}{5} - ... = \frac{\pi}{4}$, ahonnan π értéke kiszámítható bármilyen pontossággal.

2.3.4. Taylor sor Az előbbi fejezetben láttuk miként rendelhető hozzá egy konvergens sorhoz egy függvény.

A kérdés fordítva is felmerül, vagyis ha adott egy f függvény előállítható-e hatványsorral?

Legyen $I \subset \mathbb{R}$ egy nyílt intervallum, $x_0 \in I$ és $f : I \to \mathbb{R}$ egy végtelenszer deriválható függvény: $f \in C^{\infty}(I)$.

Minden f függvényhez hozzárendelhetünk az úgy nevezett Taylor sorát:

$$f(x) \approx f(x_0)+\frac{1}{1!}f'(x_0)(x-x_0)+\frac{1}{2!}f''(x_0)(x-x_0)^2+...+\frac{1}{n!}f^{(n)}(x_0)(x-x_0)^n+...$$

de ez nem jelent automatikus egyenlőséget a függvény és a Taylor sora között.

35. Definíció. Egy f függvényt analitikusnak nevezzünk ha bármely $x_0 \in I$-re, f előállítható hatványsorként x_0 környezetében (lokálisan):

$$f(x) = \sum_{n=0}^{\infty} \frac{1}{n!}f^{(n)}(x_0)(x-x_0)^n, \ x \in V_{x_0}. \tag{2.3.15}$$

Minden $\mathbb{R}$-en értelmezett elemi függvény analitikus, de nem minden végtelenszer deriválható függvény analitikus.

36. Példa. Az $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = \begin{cases} \frac{1}{e^{\frac{1}{x}}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

függvény végtelenszer deriválható: $f \in C^{\infty}(\mathbb{R})$ és $f^{(n)}(0) = 0, \ n \geq 0$ de nem analitikus. Különben (2.3.13) szerint $f(x) = 0$, minden x értékre 0 környezetében ami hamis.

37. Tétel. *Ha $f : I \to \mathbb{R}$ egy végtelenszer deriválható függvény az $I \subset \mathbb{R}$ nyílt intervallumon: $f \epsilon C^{\infty}(I)$, és $\exists M > 0$ úgy, hogy $\forall x \in I, \ \forall n \geq 0$,*

$$\left|f^{(n)}(x)\right| \leq M, \tag{2.3.16}$$

akkor az f függvény analitikus.

Numerikusan az f függvényt véges tagú összegként számítjuk ki, nevezetesen az n-ed fokú T_n Taylor polinom segítségével:

(2.3.17)

$$T_n(x) = f(x_0)+\frac{1}{1!}f'(x_0)(x-x_0)+\frac{1}{2!}f''(x_0)(x-x_0)^2+...+\frac{1}{n!}f^{(n)}(x_0)(x-x_0)^n.$$

38. TÉTEL. *Ha* $f \in C^{(n+1)}(I)$, $x_0 \in I$ *és* $M = \sup_{x \in I} \left\| f^{(n+1)}(x) \right\|$, *akkor:*

$$f(x) = T_n(x) + R_n(x) \tag{2.3.18}$$

ahol az R_n *hibatagra igaz az alábbi egyenlőtlenség:*

$$|R_n(x)| \leq M \frac{1}{(n+1)!} |x - x_0|^{n+1}. \tag{2.3.19}$$

Ha $f \epsilon C^{n+1}(I)$ *és* $x_0 \epsilon I$, *akkor az* R_n *maradéktag felírható az alábbi Lagrange--féle alakban:*

$$R_n(x) = \frac{(x - x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi), \ \xi \epsilon (x_0, x) \ \ (\textit{vagy } \xi \epsilon (x, x_0)\ ). \tag{2.3.20}$$

Ha $x_0 = 0$, akkor a Taylor sorfejtés MacLaurin nevet viseli:

$$T_n(x) = f(0) + \frac{x}{1!} f'(0) + \frac{x^2}{2!} f''(0) + ... + \frac{x^n}{n!} f^{(n)}(0). \tag{2.3.21}$$

Ha a $h = x - x_0$ jelölést használjuk, akkor a (2.3.18) Taylor képlet az alábbi alakban adható meg:

(2.3.22)

$$f(x_0 + h) = f(x_0) + \frac{h}{1!} f'(x_0) + \frac{h^2}{2!} f''(x_0) + ... + \frac{h^n}{n!} f^{(n)}(x_0) + \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(x_0 + \theta h)$$

ahol $\theta \epsilon (0, 1)$.

39. PÉLDA. a) Az $f, g : \mathbb{R} \to \mathbb{R}$, $f(x) = \sin(x)$, $g(x) = \cos(x)$ függvényekre $M = 1$ és a (2.3.21) MacLaurin sorfejtésük

$$\sin(x) = \frac{x}{1!} - \frac{x^3}{3!} + ... = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!},$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - ... = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}.$$

A sorok konvergencia sugara $R = \infty$.

b) Az $f : (-1, 1) \to \mathbb{R}$, $f(x) = e^x$ függvényre $M = 3$ és a (2.3.21) MacLaurin sorfejtése

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + ... = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

A sor konvergencia sugara $R = \infty$.

c) Az $f : (-1, 1) \to \mathbb{R}$, $f(x) = (1+x)^{\alpha}$, $\alpha\epsilon \in \mathbb{R} - \mathbb{N}$ függvényre $M =$ és a (2.3.21) MacLaurin sorfejtése

$$(1+x)^{\alpha} = 1 + \frac{\alpha}{1!}x + \frac{\alpha(\alpha-1)}{2!}x^2 + ...$$

A sor konvergencia sugara $R = 1$.

2.3.5. Hibabecslés $O-$nagy ordó segítségével A függvények aszimptotikus viselkedésének a tanulmányozására használjuk az úgy nevezett nagy-ordó O, kis-ordó o (Landau-féle) jelöléseket.

40. Definíció. Ha V_a az $a \in \mathbb{R}$ egy környezete és $f, g : V_a \to \mathbb{R}$ két függvény, akkor $f = O(g)$ (f egyenlő nagy-ordó g-vel), ha $\exists c > 0$ konstans ú.h.

$$|f(x)| \leq c \cdot |g(x)|, \ x \in V_a. \tag{2.3.23}$$

41. Példa. Ha $f(x) = 3x^2 - x + 5$ akkor $f = O(x^2)$ mert $|3x^2 - x + 5| < 4x^2$, $\forall x \geq 2$. Hasonlóan $\sin(x) = O(1)$, $\forall x \in \mathbb{R}$.

A O használata egyszerűsíti a függvények $f \simeq \overline{f}$ közelítéséből létrejövő hiba tanulmányozását, ugyanis a hibafüggvényt egy egyszerűbb (általában polinom) függvény viselkedésével helyettesítjük, ugyanakkor a c konstans kevésbé érdekel.

42. Definíció. Az $\overline{f} : V_a \to \mathbb{R}$ függvény $O(x^n)$ renddel közelíti meg az $f : V_a \to \mathbb{R}$ függvényt ha

$$\left|f(x) - \overline{f}(x)\right| \leq c \cdot |x^n|, \ x \in V_a \tag{2.3.24}$$

vagyis

$$hiba = \left|f(x) - \overline{f}(x)\right| = O(x^n).$$

A közelítés, illetve az abból eredő hibarend az alábbi egyszerű kifejezéssel írható le:

$$f(x) = \overline{f}(x) + O(x^n). \tag{2.3.25}$$

Például, a MacLaurin sorfejtésből

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$$

és az

$$e^x \simeq 1 + x + \frac{x^2}{2}$$

közelítésből eredő abszolút hiba: $\left|e^x - \left(1 + x + \frac{x^2}{2}\right)\right|$, zéró környezetében kisebb mint $|x^3|$ (egy konstanstól eltekintve). Tehát

$$e^x = 1 + x + \frac{x^2}{2} + O\left(x^3\right),\ x \in V_0.$$

Hasonlóan

$$\begin{aligned} e^x &= 1 + \frac{x}{1!} + \frac{x^2}{2!} + ... + \frac{x^n}{n!} + O\left(x^{n+1}\right),\ x \in V_0, \\ \sin(x) &= \frac{x}{1!} - \frac{x^3}{3!} + ... + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + O\left(x^{2n+3}\right),\ x \in V_0. \end{aligned}$$

Általánosan, a (2.3.18),(2.3.19) képletekből

$$f(x) = T_n(x) + R_n(x) = T_n(x) + O\left(x^{n+1}\right)$$

ahol T_n az f függvény n-ed rendű Taylor polinomja.

A nagy-ordóra a következő összeadási, illetve szorzási szabályok érvényesek:

- $O(x^p) + O(x^q) = O(x^r)$ ahol $r = \min\{p, q\}$;
- $O(x^p) \cdot O(x^q) = O(x^s)$ ahol $s = p + q$.

43. Példa. A MacLaurin sorfejtéseket használva $x \in V_0$:

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + O\left(x^4\right), \cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} + O\left(x^6\right) \Rightarrow$$

$e^x + \cos(x) = \left(1 + x + \frac{x^2}{2} + \frac{x^3}{6}\right) + \left(1 - \frac{x^2}{2} + \frac{x^4}{24}\right) + O\left(x^{\min\{4,6\}}\right) =$
$2 + x + \frac{x^3}{6} + \frac{x^4}{24} + O(x^4) = 2 + x + \frac{x^3}{6} + O(x^4)$ mert $\frac{x^4}{24} + O(x^4) = O(x^4)$;
$e^x \cdot \cos(x) = \left(1 + x + \frac{x^2}{2} + \frac{x^3}{6}\right) \cdot \left(1 - \frac{x^2}{2} + \frac{x^4}{24}\right) + \left(1 + x + \frac{x^2}{2} + \frac{x^3}{6}\right) O(x^6) +$
$+ \left(1 - \frac{x^2}{2} + \frac{x^4}{24}\right) O(x^4) + O(x^{4+6}) = (1 + x - \frac{1}{3}x^3 - \frac{5}{24}x^4 - \frac{1}{24}x^5 + \frac{1}{48}x^6 +$
$\frac{1}{144}x^7) + O\left(x^{\min(4,6)}\right) = 1 + x - \frac{1}{3}x^3 + O(x^4)$.

Összefoglalva, ha $f(x) = \overline{f}(x) + O(x^p)$, $g(x) = \overline{g}(x) + O(x^q)$, és $r = \min\{p, q\}$, akkor:

- $f(x) + g(x) = \overline{f}(x) + \overline{g}(x) + O(x^r)$;
- $f(x)\, g(x) = \overline{f}(x)\, \overline{g}(x) + O(x^{p+q})$;

- $\frac{f(x)}{g(x)} = \frac{\overline{f}(x)}{\overline{g}(x)} + O\left(x^{p-q}\right),\ \left(g\left(x\right), \overline{g}\left(x\right) \neq 0\right).$

A sorozatok sajátos függvények, ezért az esetükben is használhatók a O-ra említett tulajdonságok.

44. Definíció. Adott $(x_n)_{n=1}^{\infty}, (y_n)_{n=1}^{\infty}$ sorozatok esetében, $(x_n) = O\left((y_n)\right)$ (az (x_n) sorozat egyenlő nagy O rendű (y_n)), ha $\exists c > 0$ és N számok úgy, hogy

$$|x_n| \leq c \cdot |y_n|, \ \forall n \geq N.$$

45. Példa. Ha $x_n = \frac{n-1}{n^2}$ és $y_n = \frac{1}{n}$ akkor $|x_n| \leq |y_n|, \ \forall n \geq 1$, tehát $x_n = O\left((y_n)\right)$.

46. Definíció. Az $(x_n)_n$, $\lim_n x_n = x^*$, konvergens sorozatról azt mondjuk, hogy konvergenciarendje $p \in \mathbb{R}$, $p \geq 1$, ha

$$\lim_n \frac{|x_{n+1} - x^*|}{|x_n - x^*|^p} = c, \tag{2.3.26}$$

vagy

$$e_{n+1} = O\left(e_n^p\right), \tag{2.3.27}$$

ahol $e_n = |x_n - x^*|$ az x_n hibája.

A konvergenciát lineárisnak nevezzük ha $p = 1$, szuper-lineárisnak ha $p > 1$, és négyzetesnek ha $p = 2$.

Az algoritmusok komplexitásának az elemzéséhez úgyszintén használatos az O jelölés. Igy például két $n \times n$-es mátrix szorzásának a komplexitása $O\left(n^3\right)$. Az $O\left(n^p\right),\ p \geq 1$ algoritmusokat polinomiális típusúnak nevezzük ($p = 1$ lineáris, $p = 2$ négyzetes, $p = 3$ köbös). Gyakran fordul elő még a logaritmikus $O\left(\log n\right)$, loglineáris $O\left(n \log n\right) = O\left(\log n^n\right) = O\left(n!\right)$, exponenciális $O\left(c^n\right)$, $c > 1$, illetve faktoriális $O\left(n!\right)$ típusú algoritmusok.

A függvények aszimptotikus vizsgálatát más mérőszámmal is elvégezhető, például a kis-ordó feltétellel.

Ha $f, g : V_a \to \mathbb{R}$, akkor $f = o\left(g\right)$ ha $\lim_{x \to a} \frac{f(x)}{g(x)} = 0$.

Az értelmezésből látszik, hogy $f = o\left(g\right)$ feltétel erősebb mint a $f = O\left(g\right)$.

3. FEJEZET

Egyenletek numerikus megoldása

Az $f(x) = 0$, $f : D \subset \mathbb{R} \longrightarrow \mathbb{R}$ egyenleteket algebrainak, illetve transzcendensnek nevezzük attól függően hogy az f polinom e (esetleg azzá alakítható) vagy sem.

Pl. $f(x) = 4x^3 + 16x - 3 = 0$ egyenlet algebrai, míg az $f(x) = \sin x + 1 - x = 0$ transzcendens egyenlet.

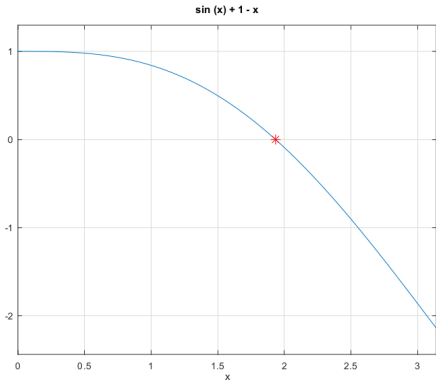
Az egyenletek numerikus megoldása két lépésben történik: az első lépés a gyökök elkülönítése (szeparálása), majd következik, a meghatározott intervallumokban, a gyökök kiszámítása.

3.1. A gyökök elkülönítése

A gyökök elkülönítése egy olyan eljárás amellyel a D értelmezési tartományt diszjunkt intervallumokra bontjuk úgy, hogy minden intervallum az egyenlet egyetlenegy gyökét tartalmazza. A módszerek között meg lehet említeni a grafikus, illetve az analitikus eljárást.

A grafikus módszer

Az eljárás a következő: vázoljuk az $y = f(x)$ függvény görbéjét, majd az Ox tengellyel való metszéspontok szolgálják az $f(x) = 0$ egyenlet közelítő gyökeit.



3.1.1. ábra. A gyök a függvény és az Ox tengely metszeténél található

Gyakran előnyösebb felbontani az $f(x) = 0$ egyenletet $f_1(x) - f_2(x) = 0$ alakra, es ezután keresni az f_1 illetve f_2 függvények metszéspontját.

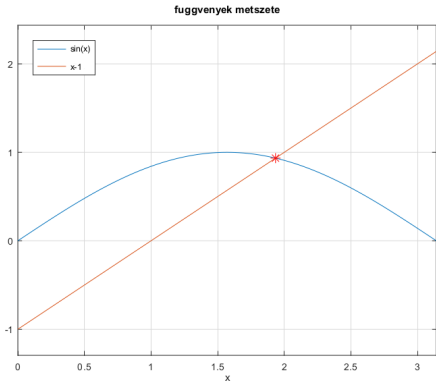
Például a

$$f(x) = \sin x + 1 - x = 0$$

egyenletet előnyösebb (és könnyebb) felbontani a következőképpen:

$$\sin x = x - 1.$$

Az egyenlet gyöke a két függvény metszeténél található: $x^* \in (0, \pi)$.



3.1.2. ábra. A gyök mint függvények metszete

Analitikus módszer

A módszer a Rolle tétel két folyományán alapszik.

Folyomány1. Ha a es b az $f \in C^1(D)$ függvénynek két egymásutáni stacionárius pontja: $f'(a) = f'(b) = 0$, akkor f-nek legfeljebb egy gyöke van az (a, b) intervallumban.

Valóban, ha $\exists c_1, c_2 \in (a, b)$ ugy hogy $f(c_1) = f(c_2) = 0$ akkor Rolle tétele szerint $\exists \overline{x} \in (c_1, c_2)$ u.h. $f'(\overline{x}) = 0$, de akkor a es b nem egymásutáni.

Folyomány2. Ha az $f \in C(D)$ folytonos függvényre teljesül a következő állítás: $f(a) f(b) < 0$, akkor f-nek legalább egy gyöke van az (a, b) intervallumban.

Ha a két folyomány igaz, akkor f-nek egyetlen egy gyöke lesz az (a, b) intervallumban.

Tehát:

- meghatározzuk az f függvény stacionárius pontjait: $x_1, x_2, ..., x_n$ es ezekkel

-kepézzük a Rolle-féle sorozatot: $f(x_1), f(x_2), ..., f(x_n)$.

Ha a Rolle-féle sorozatban előjel váltás van, akkor az illető intervallumban egyetlen egy gyöke van az egyenletnek.

Pl. Különítsük el az $2x^3 - 9x^2 + 12x - 4.5 = 0$ egyenlet gyökeit.

$f(x) = 2x^3 - 9x^2 + 12x - 4.5 \Rightarrow f'(x) = 6x^2 - 18x + 12$.

$f'(x) = 0 \implies x_1 = 1$ és $x_2 = 2$. A Rolle sorozatot táblázatba írjuk:

x	-∞	1	2	$+\infty$
$f(x)$	-∞	0.5	0.5	$+\infty$
	$-$	$+$	$-$	$+$

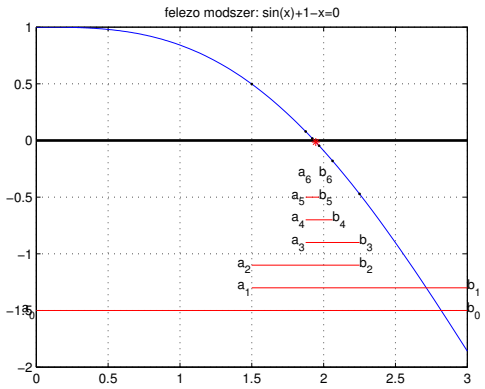
Tehát az egyenlet gyökeit elkülönítettük a $(-\infty, 1), (1, 2), (2, +\infty)$ intervallumokban.

3.2. A felező módszer

Tételezzük fel, hogy az $f(x) = 0$ egyenletnek elkülönítettük a gyökét az $[a, b]$ intervallumban és $f(a) \cdot f(b) \leq 0$, $f \in C[a, b]$.

Az $[a, b]$ intervallumot felbontjuk két egyforma hosszúságú részintervallumra: $\left[a, \frac{a+b}{2}\right], \left[\frac{a+b}{2}, b\right]$. Ezek közül csak az egyik tartalmazza

az x^* gyököt (hacsak nem éppen $\frac{a+b}{2}$ a gyök, de akkor a feladat meg van oldva). Ezt az intervallumot újból felezzük, stb.



3.2.1. ábra. Felező módszer

Az algoritmus két sorozatnak $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$ a felépítéséből áll. A sorozatok tagjait a következőképpen értelmezzük: $a_0 = a$, $b_0 = b$. Ekkor vagy $f(a) f\left(\frac{a+b}{2}\right) \leq 0$, vagy $f(a) f\left(\frac{a+b}{2}\right) \geq 0$; ennek megfelelően $a_1 = a$ és $b_1 = \frac{a+b}{2}$, vagy $a_1 = \frac{a+b}{2}$ és $b_1 = b$. Az n.-ik lépésnél meghatározzuk az $[a_n, b_n]$ intervallumot mely tartalmazza az x^* gyököt és amit újból felezünk. Ha $f(a_n) f\left(\frac{a_n+b_n}{2}\right) \leq 0 \Rightarrow a_{n+1} = a_n$ és $b_{n+1} = \frac{a_n+b_n}{2}$, ha pedig $f(a_n) f\left(\frac{a_n+b_n}{2}\right) \geq 0 \Rightarrow a_{n+1} = \frac{a_n+b_n}{2}$ és $b_{n+1} = b_n$.

47. Tétel. *Az $(a_n)_n$ sorozat monoton növekvő, a $(b_n)_n$ sorozat pedig monoton csökkenő. A sorozatok határértéke megegyezik:*

$$\lim_n a_n = \lim_n b_n = x^*, \text{ ahol } f(x^*) = 0.$$

Bizonyítás. A sorozatok felépítéséből következik a monotonitás. Mivel $a_n \leq x^* \leq b_n$ következik hogy a sorozatok korlátosak. Tehát a sorozatok konvergensek. Határértéket számítva a

$$b_n - a_n = \frac{b-a}{2^n}$$

képletből, a fogó szabályt alkalmazva, azt kapjuk, hogy:

$$\lim_n b_n = \lim_n a_n = x^*.$$

□

A gyakorlatban természetesen csak véges lépést fogunk végrehajtani, viszont minden lépésben csak egy intervallumot kapunk (kivéve ha a gyök az intervallum közepére esik, ekkor viszont az algoritmus véget ér). A keresett gyök közelíthető az n-ik intervallum közepével

$$x^* \approx c_n = \frac{a_n + b_n}{2}. \tag{3.2.1}$$

Mivel minden lépésben az intervallum feleződik, az $[a_n, b_n]$ intervallum hossza

$$b_n - a_n = \frac{b-a}{2^n}.$$

vagyis, n lépés után az abszolút hiba e_n kisebb mint

$$e_n = |c_n - x^*| < \frac{b-a}{2^n}. \tag{3.2.2}$$

48. Tétel. *A módszer konvergenciarendje lineáris:*

$$e_{n+1} = O(e_n).$$

Bizonyítás. e_n és e_{n+1}-vel jelölve a hibákat az n, illetve $(n+1)$-ik intervallumban és ismerve, hogy az intervallumok feleződnek következik, hogy

$$e_{n+1} = \frac{1}{2} e_n$$

vagyis

$$e_{n+1} = O(e_n).$$

□

Ha adott egy $\epsilon > 0$ pontosság, akkor a $\frac{b-a}{2^n} \leq \epsilon$ egyenlőtlenségből következik, hogy

$$n \geq \frac{\ln\left(\frac{b-a}{\epsilon}\right)}{\ln 2}$$

lépésre van szükség a pontosság eléréséhez.

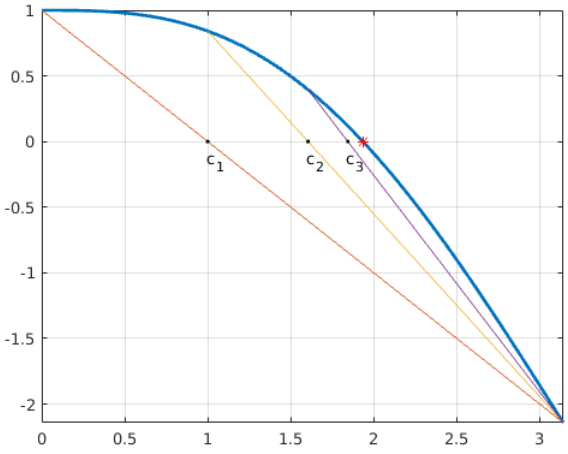
49. Példa. $f(x) = \sin x + 1 - x = 0$. A $[0, \pi]$ intervallumon az f függvény előjelt vált, közepe $c_1 = \frac{\pi}{2}$ és $f(0) = 1 > 0$, $f(\pi) = 1 - \pi < 0$, $f\left(\frac{\pi}{2}\right) = 2 - \frac{\pi}{2} > 0$, tehát $x^* \in \left[\frac{\pi}{2}, \pi\right]$ mert itt vált előjelt a függvény. A $\left[\frac{\pi}{2}, \pi\right]$ intervallum kőzepe $c_2 = \frac{3\pi}{4}$ és $f\left(\frac{3\pi}{4}\right) < 0$ tehát $x^* \in \left[\frac{\pi}{2}, \frac{3\pi}{4}\right]$. Ha a

gyököt $x^* \approx c_2$-vel közelítjük akkor a hiba kisebb mint $e_2 < \frac{\pi}{2^2}$. Ahhoz, hogy $\epsilon = 10^{-4}$ pontosságot elérjünk $\ln\left(\frac{\pi}{10^{-4}}\right) / \ln(2) = 14.939 \leq n = 15$ felezést kell végrehajtani.

A módszer nem alkalmazható ha az f függvény nem vált előjelt a gyök két oldalán, például az $f(x) = x^2 = 0,\ x \in [-1, 3]$ egyenlet nem oldható meg felezési módszerrel.

3.3. A húr módszer (regula falsi)

A felező módszerhez hasonlóan, feltételezzük, hogy $f(a) \cdot f(b) < 0$, vagyis a gyök el van különítve az (a, b) intervallumba. A húr módszer alapgondolata hasonló a felező módszeréhez, a különbség az, hogy a gyököt nem az (a, b) intervallum közepével közelítjük, hanem az $A = (a, f(a))$, $B = (b, f(b))$ pontokat összekötő húr az Ox tengellyel való metszetével: $(c, 0)$.



3.3.1. ábra. Húr módszer

A húr egyenlete

$$y - f(a) = \frac{f(b) - f(a)}{b - a}(x - a),$$

ahonnan

$$c = a - \frac{f(a)}{\frac{f(b)-f(a)}{b-a}}. \tag{3.3.1}$$

Az eljárást folytatjuk az $[a, b] := [a, c]$, vagy az $[a, b] := [c, b]$ intervallummal attól függően, hogy az első vagy a második intervallumban található-e a gyök. A felező módszertől eltérően a húr módszernél az $[a, b]$ intervallum nem feltétlenül konvergál nullához. Ha $\epsilon > 0$ egy adott pontosság és x_n a módszerrel előállított sorozat (közelítő) tagjai, akkor kilépésként használhatjuk az alábbi kritériumokat:

(1) az eljárás utolsó két tag abszolút eltérése kisebb mint ϵ:

$$|x_{n+1} - x_n| < \epsilon, \tag{3.3.2}$$

(2) az eljárás utolsó két tag relatív eltérése kisebb mint ϵ:

$$\frac{|x_{n+1} - x_n|}{|x_{n+1}|} < \epsilon, \tag{3.3.3}$$

(3) mivel (x_n) konvergál az f gyökéhez következik, hogy $f(x_n) \approx 0$, tehát:

$$|f(x_n)| < \epsilon. \tag{3.3.4}$$

A fenti kilépési feltételek nem garantálják, hogy az x_n eltérése a pontos gyöktől kisebb mint ϵ.

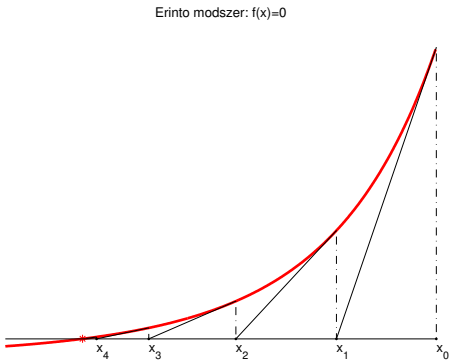
A húr módszer konvergenciarendje lineáris.

50. Példa. $f(x) = \sin(x) + 1 - x$, $x \in [0, \pi]$. A húr metszete az Ox tengellyel $c_1 = 1$ és $f(0) = 1 > 0$, $f(\pi) = 1 - \pi < 0$, $f(1) = \sin(1) > 0$, tehát $x^* \in [1, \pi]$ mert itt vált előjelt a függvény. Az $[1, \pi]$ intervallumon a húr metszete az Ox tengellyel $c_2 = 1.6$.

3.4. Az érintő (Newton) módszer

Tételezzük fel, hogy az $f(x) = 0$ egyenletnek, $f \in C^1([a, b])$, elkülönítettük a gyökét az $[a, b]$ intervallumban.

A módszer segítségével egy olyan $(x_n)_n$ konvergens sorozatot állítunk elő amely konvergál az adott egyenlet x^* megoldásához. A módszer mértani jelentése és a sorozat előállításának lépései az alábbi ábrán láthatók.



3.4.1. ábra. Érintő módszer

Legyen $x_0 = b$. Az $M_0\left(x_0, f\left(x_0\right)\right)$ pontból érintőt húzzunk a függvény görbéjéhez, ami az Ox tengelyt az x_1-ben metszi. Az érintők szerkesztését folytatjuk az $M_1\left(x_1, f\left(x_1\right)\right)$ ponttal, illetve az n-ik lépésben az $M_n\left(x_n, f\left(x_n\right)\right)$ ponttal. Ebben a pontban az érintő egyenlete a következő:

$$y - f\left(x_n\right) = f'\left(x_n\right)\left(x - x_n\right).$$

Az érintő metszete az Ox tengellyel adja az $(x_{n+1}, 0)$ pontot:$\Longrightarrow$

$$-f\left(x_n\right) = f'\left(x_n\right)\left(x_{n+1} - x_n\right)$$

ahonnan

$$x_{n+1} = x_n - \frac{f\left(x_n\right)}{f'\left(x_n\right)}. \tag{3.4.1}$$

3.4.1. A kiindulópont megválasztása és az érintő módszer konvergenciája Az iteratív módszereknél nagyon fontos a kezdeti pont helyes kiválasztva, ugyanis gyakran függ ettől a módszer konvergenciája.

A fenti ábrán is érzékelhető: ha az $[a,b]$ intervallum másik végpontját választottuk volna, akkor x_1 kívül esett volna ennek az intervallumon.

A legegyszerűbb ha x_0 az $[a,b]$ intervallum egyik végpontja, de ez nem szükségszerű.

Feltételezzük, hogy $f \in C^2([a,b])$ és az f', f'' deriváltak előjeltartóak az adott intervallumban. Ebben az esetben a következő esetek lehetségesek:

ABRAK

$$f(x_0) f''(x_0) > 0 \tag{3.4.2}$$

51. Tétel. *Az érintő módszerrel (3.4.1) kapott sorozat* $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, *ahol* x_0 *a fentiek szerint van meghatározva, konvergens.*

Bizonyítás. Feltételezzük, hogy $f'(x) > 0, f''(x) > 0$ (a többi eset tárgyalása hasonlóan történik). Igazoljuk, hogy az $(x_n)_n$ sorozat monoton és korlátos, tehát konvergens. A (3.4.2)-ból következik, hogy $f(x_0) > 0$. Mivel $f'(x) > 0$ következik, hogy f növekvő, és negatív az $[a, x^*)$ illetve pozitív az $(x^*, b]$ intervallumon, tehát $x^0 \in (x^*, b]$, ahol x^* az f gyöke $f(x^*) = 0$. Feltételezve, hogy $x^* < x_n \leq b$ a Taylor képletből következik, hogy

$$0 = f(x^*) = f(x_n) + \frac{1}{1!} f'(x_n)(x^* - x_n) + \frac{1}{2!} f''(\xi_n)(x^* - x_n)^2,\ \xi_n \in (x^*, x_n)$$

és mivel $f''(\xi_n) > 0$ következik, hogy

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} > x^*,$$

tehát x^* az $(x_n)_n$ sorozat alsó korlátja. Mivel $f(x_n) > 0$, $\forall x_n \in (x^*, b]$

$$x_{n+1} - x_n = -\frac{f(x_n)}{f'(x_n)} < 0,$$

tehát az $(x_n)_n$ sorozat monoton csökkenő. λ-val jelölve az $(x_n)_n$ sorozat határértékét a (3.4.1)-ből

$$\lambda = \lambda - \frac{f(\lambda)}{f'(\lambda)}$$

ahonnan $f(\lambda)=0$, vagyis $\lambda=x^*$. □

52. TÉTEL. *Az érintő módszer konvergenciarendje négyzetes*

$$e_{n+1}=O\left(e_n^2\right),$$

ahol $e_n=|x_n-x^*|$.

BIZONYÍTÁS. Taylor sorba fejtve az f függvényt következik, hogy

$$0=f(x^*)=f(x_n)+\frac{1}{1!}f'(x_n)(x^*-x_n)+\frac{1}{2!}f''(\xi_n)(x^*-x_n)^2,\ \xi_n\in(x^*,x_n)$$

ahonnan figyelembe véve, hogy $f(x^*)=0$ következik, hogy

$$x_n-\frac{f(x_n)}{f'(x_n)}-x^*=\frac{f''(\xi_n)}{2f'(x_n)}(x^*-x_n)^2,\ \xi_n\in(x^*,x_n).$$

Felhasználva a Newton iterációt (3.4.1) következik, hogy

$$x_{n+1}-x^*=\frac{f''(\xi_n)}{2f'(x_n)}(x^*-x_n)^2$$

$\Rightarrow$

$$\frac{|x_{n+1}-x^*|}{\left|(x^*-x_n)^2\right|}=\left|\frac{f''(\xi_n)}{2f'(x_n)}\right|\xrightarrow{n\to\infty}\left|\frac{f''(\xi_n)}{2f'(x^*)}\right|=c,$$

vagyis $e_{n+1}=O\left(e_n^2\right)$. □

53. PÉLDA. $f(x)=\sin x+1-x=0$, $x\in(0,\pi)\Rightarrow f'(x)=\cos x-1$, $f''(x)=-\sin x<0$ ha $x\in(0,\pi)$, tehát $x_0=\pi$. Az (3.4.1) iterációból $\Rightarrow x_1=x_0-\frac{f(x_0)}{f'(x_0)}=\pi-\frac{1-\pi}{-2}\approx 2.0708$, $x_2=x_1-\frac{f(x_1)}{f'(x_1)}\approx 1.9402$.

Az n.-ik lépésben az iterálást abba lehet hagyni ha a függvény értéke $f(x_n)<\epsilon$, vagy két egymásutáni tag abszolút $|x_n-x_{n-1}|$ (relatív $\frac{|x_n-x_{n-1}|}{|x_n|}$) eltérése elég kicsi ($<\epsilon$).

54. PÉLDA. A Newton módszer jól alkalmazható egész kitevőjű gyökvonásra $\sqrt[p]{a}$. Például $\sqrt{a}$ kiszámítása ekvivalens az $x^2-a=0$ egyenlet (pozitív) megoldásával, amit Newton módszerrel határozunk meg. A sorozat rekurziós képlete a következő lesz:

$$x_{n+1}=x_n-\frac{x_n^2-a}{2x_n}=\frac{1}{2}\left(x_n+\frac{a}{x_n}\right), \tag{3.4.3}$$

ahonnan $a=2$-re $\Rightarrow$ $\sqrt{2}\approx x_0=2$, $\sqrt{2}\approx\frac{3}{2}$, $\sqrt{2}\approx\frac{17}{12}$, $\sqrt{2}\approx\frac{577}{408}=1.4142$, ...

A Newton módszer hátránya, hogy minden iterációban szükséges az f függvény deriváltja. A következő módszerek ezt a hátrányt küszöbölik ki a konvergenciarend rovására.

3.5. A szelő módszer

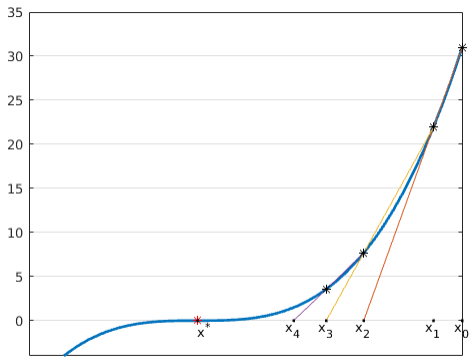
A (3.4.1) képletben az $f'(x_n)$ tagot felcserélve a közelítő értékkel

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

a következő iterációs képletet kapjuk:

$$x_{n+1} = x_n - \frac{f(x_n)}{\frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}},\ n \geq 1,\ x_0, x_1 = \text{kezdeti értékek} \tag{3.5.1}$$

A mértani jelentése a szelő módszernek a következő: kiindulva x_0, x_1 kezdeti értékekből az $(x_{n-1}, f(x_{n-1})), (x_n, f(x_n))$ pontokon áthaladó szelő metszete az Ox tengellyel adja a közelítő sorozat x_{n+1} tagját.



3.5.1. ábra. Szelő módszer

Az $(x_n)_n$ sorozat első tagját x_0-t ugyanúgy választjuk meg mint az érintő módszernél (3.4.2), x_1 pedig egy pont x_0 es x^* között: $x_1 \in (x^*, x_0)$. Az x_0, x_1 pontoknak megfelel a függvény görbéjén az $M_0(x_0, f(x_0))$, és $M_1(x_1, f(x_1))$. Az M_0M_1 szelő metszi az Ox tengelyt x_2-ben. Az eljárást folytatjuk M_1M_2 szelővel. Az n-edik lépésnél $M_{n-1}M_n \cap Ox =$

x_{n+1}. Az $M_{n-1}M_n$ szelő egyenlete:

$$f(x) - f(x_n) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x - x_n)$$

amelybe behelyettesítjük a metszéspont koordinátáit: $(x_{n+1}, 0) \Longrightarrow$

$$-f(x_n) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x_{n+1} - x_n)$$

ahonnan kifejezve x_{n+1}-t kapjuk a (3.5.1) képletet.

55. TÉTEL. *A szelő módszer konvergenciarendje* $p = \frac{1+\sqrt{5}}{2} (= \Phi$ *aranyszám*).

3.6. A Steffensen-féle módszer

Az $f'(x_n)$ derivált közelítéséhez felhasználjuk, hogy a gyök közelében $f(x_n) \approx f(x^*) = 0$, tehát a derivált

$$f'(x_n) = \lim_{h \to 0} \frac{f(x_n + h) - f(x_n)}{h} \approx \frac{f(x_n + h) - f(x_n)}{h}$$

átírható a következő alakban:

$$f'(x_n) \approx \frac{f(x_n + f(x_n)) - f(x_n)}{f(x_n)}.$$

Innen az iteratív eljárás a következő lesz:

$$x_{n+1} = x_n - \frac{f(x_n)}{\frac{f(x_n + f(x_n)) - f(x_n)}{f(x_n)}}. \tag{3.6.1}$$

56. TÉTEL. *A Steffensen módszer konvergenciarendje (a gyök környezetében) négyzetes* $p = 2$.

3.7. A fokozatos közelítések módszere (fixpont módszer)

A fokozatos közelítések (szukcesszív approximációk) módszere a numerikus eljárások egyik alapvető és gyakran használt módszere.

A megoldandó egyenletet $f(x) = 0$ átírjuk a következő alakra:

$$x = \varphi(x). \tag{3.7.1}$$

Erre több lehetőség is van, például $\varphi(x) = x + f(x)$, $\varphi(x) = x - f(x)$, vagy általánosabban $\varphi(x) = x + \omega f(x)$.

57. DEFINÍCIÓ. Az x értéket amelyre $x = \varphi(x)$, a φ függvény fixpontjának nevezzük.

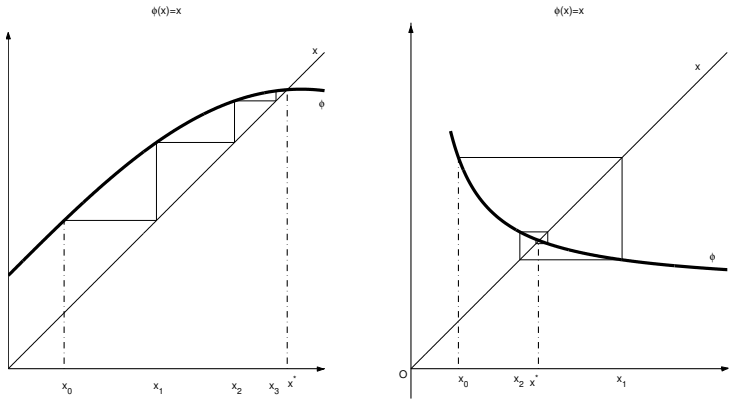
Az $y \to x$ függvény (első szögfelező) tulajdonsága, hogy az abszcisszája és ordinátája azonos.

Az $(x_n)_n$ sorozatot ahol $x_{n+1} = \varphi(x_n)$, az approximációk sorozatának nevezzük.

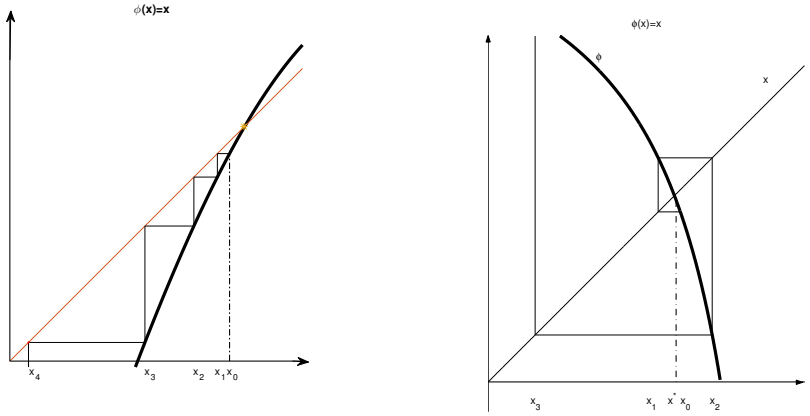
58. TÉTEL. *Ha az $(x_n)_n$ sorozat konvergens és φ folytonos, akkor a határértéke megegyezik az egyenlet gyökével.*

BIZONYÍTÁS. $x^* = \lim x_{n+1} = \lim \varphi(x_n) = \varphi(\lim x_n) = \varphi(x^*)$ vagyis x^* gyöke az (3.7.1) egyenletnek ami ekvivalens $f(x) = 0$ egyenlettel. □

A φ-től függően az eljárás lehet konvergens vagy divergens. Az alábbi ábrákon szemléltettünk egy pár esetet:



3.7.1. ábra. Vonzó gyök



3.7.2. ábra. Taszító gyök

A gyök az $y = x$ illetve $y = \varphi(x)$ görbék metszeténél található. Kiindulva az $(x_n, x_{n+1}) = (x_n, \varphi(x_n))$ pontból a (x_{n+1}, x_{n+2}) pontot a következőképpen szerkesztjük meg. Az (x_n, x_{n+1}) ponton keresztül húzzuk az $y = x_{n+1}$ egyenest az $y = x$ egyenessel való metszésig, majd berajzolva az $x = x_{n+1}$ egyenest az $y = \varphi(x)$ görbével való metszésig megkapjuk a kívánt (x_{n+1}, x_{n+2}) pontot.

(1) Legyen $\varphi : [a, b] \to [a, b]$ egy deriválható függvény. Ha teljesül a következő feltétel:

$$|\varphi'(x)| \leq q < 1, \ \forall x \in (a, b),$$

akkor:
a) az $(x_n)_n$ sorozat: $x_{n+1} = \varphi(x_n)$ konvergens tetszőleges x_0-ra,
b) a sorozat x^* határértéke az $\varphi(x) = x$ egyenlet egyetlen gyöke az $[a, b]$ intervallumban.

Bizonyítás. a)

$$|x_{n+1} - x_n| =^{def} |\varphi(x_n) - \varphi(x_{n-1})| =^{Lagrange} |\varphi'(\xi_n)(x_n - x_{n-1})| \leq$$
$$\leq^{Tetel} q|x_n - x_{n-1}| \leq ... \leq q^n |x_1 - x_0|,$$

ahol $\xi_n \in (x_n, x_{n+1}) \Longrightarrow$

$$\begin{aligned}|x_{n+p} - x_n| &= |x_{n+p} - x_{n+p-1} + x_{n+p-1} - ... - x_n| \leq \\ &\leq |x_{n+p} - x_{n+p-1}| + ... + |x_{n+1} - x_n| \leq \\ &\leq q^{n+p-1} |x_1 - x_0| + ... + q^n |x_1 - x_0| = \\ &= \left(q^{n+p-1} + ... + q^n\right) |x_1 - x_0| = q^n \frac{1-q^p}{1-q} |x_1 - x_0| .\end{aligned}$$

Tehát:

$$|x_{n+p} - x_n| \leq \frac{q^n}{1-q} |x_1 - x_0| , \ n, p \in \mathbb{N}. \tag{3.7.2}$$

Mivel $q \in (0,1)$ a jobboldal tetszőlegesen lekicsinyíthető, vagyis az $(x_n)_n$ Cauchy-féle sorozat. Az $\mathbb{R}$-en minden Cauchy sorozat konvergens, tehát $(x_n)_n$ konvergens.
b) Jelöljük az $(x_n)_n$ sorozat határértékét x^*-el:

$$\lim_n x_n = x^*.$$

Következik, hogy

$$x^* = \lim_n (x_n) = \lim_n (\varphi (x_{n-1})) = \varphi \left(\lim_n (x_{n-1})\right) = \varphi (x^*) ,$$

tehát x^* az $x = \varphi (x)$ egyenlet gyöke. Kimutatjuk hogy x^* az egyedüli gyöke. Feltételezzük, hogy az egyenletnek még létezik egy gyöke:

$$x^{**} = \varphi (x^{**}) , x^{**} \neq x^*,$$

tehát

$$|x^* - x^{**}| = |\varphi (x^*) - \varphi (x^{**})| = |\varphi' (\xi^*) (x^* - x^{**})| \leq q |x^* - x^{**}| ,$$

ahol $\xi^* \in (x^*, x^{**}) \Longrightarrow$

$$(1-q) |x^* - x^{**}| \leq 0,$$

és mivel $(1-q) > 0 \Longrightarrow$

$$|x^* - x^{**}| = 0 \Leftrightarrow x^* = x^{**},$$

hamis. □

A (3.7.2) képletből ha $p \longrightarrow \infty$ a következő hibabecslást kapjuk:

$$|x_n - x^*| \leq \frac{q^n}{1-q} |x_1 - x_0| . \tag{3.7.3}$$

Hasonlóan kimutatható, hogy:

$$|x_n - x^*| \leq \frac{q}{1-q} |x_n - x_{n-1}| , \; n \geq 1. \tag{3.7.4}$$

59. Példa. Az $x^3 + 12x - 1 = 0$ egyenletet, aminek a gyöke a $[0, 1]$ intervallumba esik, átírhatjuk többféleképpen:

(1) $x = \frac{1-x^3}{12}$,
(2) $x = \sqrt[3]{1 - 12x}$, vagy
(3) $x = \frac{1}{x^2+12}$.

Az első esetben $\varphi(x) = \frac{1-x^3}{12}$ és $|\varphi'(x)| = \frac{x^2}{4} \leq \frac{1}{4}$, $x \in [0, 1]$; a második esetben $\varphi(x) = \sqrt[3]{1 - 12x}$ függvényre nem érvényes $|\varphi'(x)| < 1$. Ha az utolsó esetet vesszük figyelembe, akkor $\varphi(x) = \frac{1}{x^2+12}$, $\varphi : [0, 1] \to [0, 1]$ és $|\varphi'(x)| < 1$. Meghatározzuk q-t:

$$q = \sup_{x \in [0,1]} |\varphi'(x)| = \sup_{x \in [0,1]} \frac{2x}{(x^2 + 12)^2} = \frac{2}{169}.$$

Kezdőértéknek legyen $x_0 = 0$

$$\Longrightarrow x_1 = \varphi(x_0) = \frac{1}{12}, \; x_2 = \varphi(x_1) = \frac{144}{1729}.$$

Ha az $\varepsilon = 10^{-4}$ hibakorlat előre meg van adva, akkor az (3.7.3)-ben kikötjük

$$|x_n - x^*| \leq \frac{q^n}{1-q} |x_1 - x_0| < \varepsilon,$$

tehát $n = 2$, vagyis x_2 megközelíti a gyököt a kívánt pontossággal.

A $|\varphi'(x)| < 1$ feltételnek eleget tevő függvény meghatározására nem létezik általános eljárás, de egyes esetekben ennek előállítására bizonyos sémát követhetünk.
Például ha $0 < m \leq f'(x) \leq M$ (vagy $M \leq f'(x) \leq m < 0$), akkor az $f(x) = 0$ egyenlet ekvivalens az

$$x = x - \frac{1}{M} f(x)$$

egyenlettel, tehát ha

$$\varphi(x) = x - \frac{1}{M} f(x), \tag{3.7.5}$$

akkor $0 \leq \varphi'(x) = 1 - \frac{f'(x)}{M} \leq 1 - \frac{m}{M} < 1$.

Az ismertet módszereknél az iteráló φ függvény a következő:

(1) Newton módszer

$$\varphi(x) = x - \frac{f(x)}{f'(x)},$$

(2) Steffensen módszer

$$\varphi(x) = x - \frac{f(x)}{\frac{f(x+f(x))-f(x)}{f(x)}},$$

(3) szelő módszer

$$\varphi(x,y) = \frac{y \cdot f(x) - x \cdot f(y)}{f(x) - f(y)}.$$

Mivel a fokozatos közelítések módszerét nem csak az $\mathbb{R}$ térben alkalmazzuk, a továbbiakban ismertetjük a fenti tétel általánosítását normált terekben.

3.8. Vektor és mátrix normák

A norma a modulust (abszolút érték) fogalom általánosítása.

60. Definíció. Az $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ függvényt vektor normának nevezzük ($\mathbb{R}^n$ térben) ha eleget tesz az alábbi tulajdonságoknak bármilyen $\boldsymbol{x}, \mathbf{y} \in \mathbb{R}^n$ és $\alpha \in \mathbb{R}$:
(i) $\|\boldsymbol{x}\| \geq 0$ és $\|\boldsymbol{x}\| = 0 \Leftrightarrow \boldsymbol{x} = 0_{\mathbb{R}^n}$;
(ii) $\|\alpha \cdot \boldsymbol{x}\| = |\alpha| \cdot \|\boldsymbol{x}\|$;
(iii) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.
Egy vektort egységvektornak nevezzünk ha $\|\boldsymbol{x}\| = 1$.

Az egyik leggyakrabban használt az úgynevezett $p-$norma:

$$\|\boldsymbol{x}\|_p = \left(|x_1|^p + ... + |x_n|^p\right)^{\frac{1}{p}}, \ p \geq 1, \ \boldsymbol{x} = (x_1, ..., x_n)^t \in \mathbb{R}^n$$

ezek között is a $p = 1,\ p = 2$ és $p = \infty$ a legismertebbek:

$$\|\boldsymbol{x}\|_1 = |x_1| + ... + |x_n|\,, \tag{3.8.1}$$

$$\|\boldsymbol{x}\|_2 = \left(x_1^2 + ... + x_n^2\right)^{\frac{1}{2}}, \tag{3.8.2}$$

$$\|\boldsymbol{x}\|_\infty = \max\left(|x_1|\,, ..., |x_n|\right). \tag{3.8.3}$$

61. Példa. Ha $\boldsymbol{x} = \begin{pmatrix} 1 \\ -5 \\ 2 \end{pmatrix} \in \mathbb{R}^3$, akkor $\|\boldsymbol{x}\|_1 = 8,\ \|\boldsymbol{x}\|_2 = \sqrt{30},\ \|\boldsymbol{x}\|_\infty = 5$.

62. Tétel. *Ha* $\frac{1}{p}+\frac{1}{q} = 1,\ p > 1, q \in \mathbb{R}$ *akkor igaz az alábbi (Hölder) egyenlőtlenség:*

$$\left|\boldsymbol{x}^t \cdot \boldsymbol{y}\right| \leq \|\boldsymbol{x}\|_p \cdot \|\boldsymbol{y}\|_q\,. \tag{3.8.4}$$

Fontos következménye az előbbi tételnek az úgynevezett Cauchy-Schwartz egyenlőtlenség:

$$\left|\boldsymbol{x}^t \cdot \boldsymbol{y}\right| \leq \|\boldsymbol{x}\|_2 \cdot \|\boldsymbol{y}\|_2\,,$$

vagy komponensekre lebontva:

$$|x_1 y_1| + ... + |x_n y_n| \leq \sqrt{x_1^2 + ... + x_n^2} \cdot \sqrt{y_1^2 + ... + y_n^2}. \tag{3.8.5}$$

63. Tétel. *A* $\mathbb{R}^n$*-n értelmezett normák ekvivalensek, vagyis ha* $\|\cdot\|_p$ *és* $\|\cdot\|_q$ *normák az* $\mathbb{R}^n$ *térben, akkor léteznek* c_1, c_2 *pozitív konstansok úgy, hogy:*

$$c_1 \|\boldsymbol{x}\|_p \leq \|\boldsymbol{x}\|_q \leq c_2 \|\boldsymbol{x}\|_p\,,\ \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Például, ha $x \in \mathbb{R}^n$ akkor:

$$\begin{aligned} \|\boldsymbol{x}\|_2 &\leq \|\boldsymbol{x}\|_1 \leq \sqrt{n}\,\|\boldsymbol{x}\|_2\,, \\ \|\boldsymbol{x}\|_\infty &\leq \|\boldsymbol{x}\|_2 \leq \sqrt{n}\,\|\boldsymbol{x}\|_\infty\,, \\ \|\boldsymbol{x}\|_\infty &\leq \|\boldsymbol{x}\|_1 \leq n\,\|\boldsymbol{x}\|_\infty\,. \end{aligned}$$

Ha $\boldsymbol{x} \in \mathbb{R}^n$ az $X \in \mathbb{R}^n$ vektor approximációja, akkor a közelítés abszolút, illetve relatív hibája:

$$\epsilon_{abs} = \|X - \boldsymbol{x}\|\,,\ \epsilon_{rel} = \frac{\|X - \boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

A mátrix norma definíciója hasonló a vektor normáéhoz.

64. DEFINÍCIÓ. Az $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ függvényt mátrix normának nevezzük ($\mathcal{M}_{mn}(\mathbb{R}) \approx \mathbb{R}^{m\times n}$ térben) ha eleget tesz az alábbi tulajdonságoknak: bármilyen $A, B \in \mathbb{R}^{m\times n}$ és $\alpha \in \mathbb{R}$:
(i) $\|A\| \geq 0$ és $\|A\| = 0 \Leftrightarrow A = 0_{\mathbb{R}^{m\times n}}$;
(ii) $\|\alpha \cdot A\| = |\alpha| \cdot \|A\|$;
(iii) $\|A + B\| \leq \|A\| + \|B\|$.
Ha egy mátrix norma teljesíti a

$$\|A \cdot B\| \leq \|A\| \cdot \|B\|, \ A, B \in \mathbb{R}^{n\times n}, \tag{3.8.6}$$

tulajdonságot, akkor (a normát) szubmultiplikatívnek nevezzük.

A leggyakrabban használt mátrix normák a Frobenius:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} = \sqrt{Tr\,(A^tA)}, \tag{3.8.7}$$

illetve egy $\boldsymbol{x} \in \mathbb{R}^n$, vektor által származtatott (indukált) $p-$norma:

$$\|A\|_p = \sup_{\boldsymbol{x}\neq 0} \frac{\|A\mathbf{x}\|_p}{\|\boldsymbol{x}\|_p}. \tag{3.8.8}$$

A (3.8.8)-ból következik, hogy:

$$\|A\|_p = \sup_{\boldsymbol{x}\neq 0} \left\| A\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_p}\right)\right\|_p = \max_{\|\boldsymbol{y}\|_p=1} \|A\boldsymbol{y}\|_p. \tag{3.8.9}$$

Mivel a mátrix p-normák kapcsolódnak a vektor p-normákhoz, következik, hogy úgyszintén a $p = 1, 2, \infty$ esetek a legfontosabbak.

Ha $A \in \mathbb{R}^{m\times n}$ akkor

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|, \tag{3.8.10}$$

$$\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|. \tag{3.8.11}$$

$p = 2$-re a $\|A\|_2^2$ norma egyenlő a $p(\lambda) = \det(A^tA - \lambda I)$ karakterisztikus polinom legnagyobb gyökével (vagyis az A^tA legnagyobb sajátértékével). Ha csak az $\|A\|_2$ norma nagyságrendje fontos, akkor használhatjuk az alábbi egyenlőtlenségeket:

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2\,, \tag{3.8.12}$$

$$\max_{i,j}|a_{ij}| \leq \|A\|_2 \leq \sqrt{mn}\max_{i,j}|a_{ij}|\,, \tag{3.8.13}$$

$$\frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{m}\,\|A\|_\infty\,, \tag{3.8.14}$$

$$\frac{1}{\sqrt{m}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\,\|A\|_1\,, \tag{3.8.15}$$

$$\|A\|_2 \leq \sqrt{\|A\|_1\,\|A\|_\infty}. \tag{3.8.16}$$

65. Példa. $A = \begin{pmatrix} 2 & -1 \\ 1.5 & 3 \\ 0 & -2 \end{pmatrix} \Rightarrow \|A\|_F = \sqrt{4+1+2.25+9+4} = 4.5$, $\|A\|_1 = \max\{3.5, 6\} = 6$, $\|A\|_\infty = \max\{3, 4.5, 2\} = 4.5, 3 \leq \|A\|_2 \leq 3\sqrt{6}$ $(\|A\|_2 \simeq 3.8388)$.

Az ismertetett normák mindegyike szubmultiplikatív tulajdonságú, viszont az $\|A\| := \max_{i,j}\{|a_{ij}|\}$ norma nem.

3.8.1. Banach-féle fixpont tétel normált terekben

66. Definíció. A $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ leképezést kontrakciónak nevezzük a $\|\cdot\|$ normára nézve, ha $\exists q \in (0,1)$ ú.h.

$$\|\Phi(X) - \Phi(Y)\| \leq q \cdot \|X - Y\|\,,\ \forall X, Y \in \mathbb{R}^n.$$

67. Tétel. *Ha $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ leképezés kontrakció, akkor Φ-nek egyetlenegy $X^* \in \mathbb{R}^n$ fixpontja van: $\Phi(X^*) = X^*$. Az $X^{(k+1)} = \Phi\left(X^{(k)}\right)$ iterációk sorozata konvergál bármilyen $X^{(0)} \in \mathbb{R}^n$ kezdőértékre. A hibára a következő becslések igazak:*

$$\left\|X^{(k)} - X^*\right\| \leq \frac{q^k}{1-q}\left\|X^{(1)} - X^{(0)}\right\|, \tag{3.8.17}$$

$$\left\|X^{(k)} - X^*\right\| \leq \frac{q}{1-q}\left\|X^{(k)} - X^{(k-1)}\right\|,\ k \geq 1. \tag{3.8.18}$$

3.9. Algebrai egyenletek numerikus megoldása

3.9.1. Polinomok alakjai Bármely n-ed fokú $p_n \in \mathbb{P}_n$ polinom felírható kanonikus alakban:

$$p_n(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0 = \sum_{i=0}^{n} a_i x^i, \tag{3.9.1}$$

ami azt jelenti hogy a $(1, x, ..., x^{n-1}, x^n)$ polinom-rendszer egy bázist alkot, vagyis a tagok lineárisan függetlenek.

Különböző okokból indokolt más alakban is felírni a polinomokat. Egy ilyen alak az úgynevezett Bézier-féle polinom alak.

68. Definíció. A Bernstein-féle alap-polinomokat a következőképpen értelmezzük:

$$b_i^n(x) = \binom{n}{i} x^i (1-x)^{n-i}, \ i = 0, ..., n; \ x \in [0,1]. \tag{3.9.2}$$

Az $[0,1]$ intervallum nem jelent megszorítást mivel az

$$x = \frac{y-a}{b-a}$$

változócserével az $[a,b] \ni y$ intervallum átalakítható $[0,1]$ intervallummá.

69. Tétel. *A b_i^n polinomokra a következő rekurrens képlet igaz:*

$$\begin{aligned} b_i^n(x) &= (1-x)\, b_i^{n-1}(x) + x b_{i-1}^{n-1}(x), \ i = 1, ..., n-1 \\ b_0^n(x) &= (1-x)\, b_0^{n-1}(x), \ b_n^n(x) = x b_{n-1}^{n-1}(x). \end{aligned} \tag{3.9.3}$$

Bizonyítás. Az értelmezést felhasználva kapjuk, hogy:

$$b_0^n(x) = \binom{n}{0} x^0 (1-x)^n = (1-x)\binom{n-1}{0} x^0 (1-x)^{n-1} = (1-x)\, b_0^{n-1}(x).$$

Hasonlóan $b_n^n(x) = xb_{n-1}^{n-1}(x)$.
Ha $i = 1, ..., n-1$ felhasználva a kombinációk rekurzív képletet kapjuk:

$$\begin{aligned} b_i^n(x) &= \binom{n}{i} x^i (1-x)^{n-i} = \left[\binom{n-1}{i} + \binom{n-1}{i-1}\right] x^i (1-x)^{n-i} = \\ &= (1-x)\binom{n-1}{i} x^i (1-x)^{n-i-1} + x\binom{n-1}{i-1} x^{i-1} (1-x)^{n-i} = \\ &= (1-x) b_i^{n-1}(x) + x b_{i-1}^{n-1}(x) \end{aligned}$$

□

70. TÉTEL. *A $b_0^n, b_1^n, ..., b_n^n$ polinomok lineárisan függetlenek.*

BIZONYÍTÁS. Legyen $c_0, c_1, ..., c_n \in \mathbb{R}$ ú.,h.:

$$c_0 b_0^n(x) + c_1 b_1^n(x) + ... + c_n b_n^n(x) = 0\,,\ x \in [0,1]\,. \tag{3.9.4}$$

Ha $n = 0$ akkor $c_0 b_0^0(x) = 0$ és mivel (3.9.2)-ből $b_0^0(x) = 1$ azt kapjuk hogy $c_0 = 0$. □

Legyen $n \geq 1$. Ha $x = 0$ (3.9.4)-ból azt kapjuk, hogy $c_0 = 0$, hasonlóan $x = 1$-re $c_n = 0$. $\Longrightarrow c_1\binom{n}{1}x^1(1-x)^{n-1} + c_2\binom{n}{2}x^2(1-x)^{n-2} ... + c_{n-1}\binom{n}{n-1}x^{n-1}(1-x) = 0\,,\ x \in [0,1]$.

Újból $x = 0$-ra $c_1 = 0$, míg $x = 1$-re $c_{n-1} = 0$ stb.$\Longrightarrow\ c_i = 0,\ \ i = 0, ..., n$ tehát b_i^n lineárisan függetlenek.

A fenti tételből következik, hogy a $(b_0^n, b_1^n, ..., b_n^n)$ rendszer egy bázist alkot a $\mathbb{P}_n$ térben, tehát bármilyen $p \in \mathbb{P}_n$ polinomot egyértelműen fel lehet írni a b_i^n polinomok lineáris kombinációjaként:

$$p = c_0 b_0^n + c_1 b_1^n + ... c_n b_n^n = \sum_{i=0}^{n} c_i b_i^n. \tag{3.9.5}$$

A fenti alakot Bézier-féle alaknak nevezzük, c_i pedig a p polinom Bézier együtthatói.

$n = 3$-ra a harmadfokú alap-polinomok: $(1-x)^3, 3x(1-x)^2, 3x^2(1-x), x^3$, egy harmadfokú polinom Bézier-féle alakja:

$$p(x) = c_0(1-x)^3 + c_1 3x(1-x)^2 + c_2 3x^2(1-x) + c_3 x^3. \tag{3.9.6}$$

71. Példa. Legyen $p \in \mathbb{P}_3$, $p(x) = 4x^3 - 3x$. $\Longrightarrow c_0 = p(0) = 0$, $c_3 = p(1) = 1$. Innen kapjuk, hogy

$$p(x) = c_1 3x(1-x)^2 + c_2 3x^2(1-x) + x^3$$

majd a c_1, c_2 tagokat tartalmazó egyenletrendszert megoldva kapjuk $c_1 = -1,\ c_2 = -2$.

A Bézier-féle alakban lévő polinom deriváltját a következő képlet adja meg:

$$p'(x) = n \sum_{i=0}^{n-1} (c_{i+1} - c_i)\, b_i^{n-1}(x).$$

Bizonyítás.

$$\left(\sum_{i=0}^{n} c_i b_i^n(x)\right)' = \sum_{i=0}^{n} c_i (b_i^n(x))' = \sum_{i=0}^{n} c_i \binom{n}{i} \left[x^i (1-x)^{n-i}\right]' =$$

$$\sum_{i=0}^{n} c_i \binom{n}{i} \left[i x^{i-1}(1-x)^{n-i} - (n-i)\, x^i (1-x)^{n-i-1}\right] =$$

$$\sum_{i=1}^{n} c_i \frac{n!}{i!\,(n-i)!} i x^{i-1}(1-x)^{n-i} - \sum_{i=0}^{n-1} c_i \frac{n!}{i!\,(n-i)!}(n-i)\, x^i (1-x)^{n-i-1} =$$

$$\sum_{i=1}^{n} c_i n \frac{(n-1)!}{(i-1)!\,(n-i)!} x^{i-1}(1-x)^{n-i} - \sum_{i=0}^{n-1} c_i n \frac{(n-1)!}{i!\,(n-i-1)!} x^i (1-x)^{n-i-1} =$$

$$n\left(\sum_{i=1}^{n} c_i \binom{n-1}{i-1} x^{i-1}(1-x)^{n-i} - \sum_{i=0}^{n-1} c_i \binom{n-1}{i} x^i (1-x)^{n-i-1}\right) =^{i-1=j}$$

$$n\left(\sum_{j=0}^{n-1} c_{j+1} b_j^{n-1}(x) - \sum_{i=0}^{n-1} c_i b_i^{n-1}(x)\right) = n \sum_{i=0}^{n-1} (c_{i+1} - c_i)\, b_i^{n-1}(x).$$

□

Igényektől függően más polinom alakról is beszélhetünk, például Lagrange, Hermite, Csebysev, Legendre stb. polinomokról.

3.9.2. A Horner, illetve de Casteljau elrendezés Adott polinomra fontos meghatározni minél pontosabban a polinom értékét bizonyos pontokban. Erre szolgál a Horner séma ha a polinom kanonikus alakban van megadva.

Legyen $p(x) = a_0 + a_1 x + ... + a_n x^n$ egy polinom, illetve egy $t \in \mathbb{R}$. A

$$p(t) = a_0 + a_1 t + ... + a_{n-1} t^{n-1} + a_n t^n$$

értéket a következő rekurzív képlettel számítjuk ki:

$$p(t) = a_0 + t(a_1 + ...t(a_{n-1} + t \cdot (a_n))). \tag{3.9.7}$$

Gyakorlatban a fenti Horner elrendezés a következőképpen valósul meg:

$$\begin{aligned} q_n &= a_n, \\ q_{n-1} &= a_{n-1} + t q_n, \\ &... \\ q_0 &= a_0 + t q_1, \end{aligned} \tag{3.9.8}$$

tehát $p(t) = q_0$.

A Horner elrendezést táblázatba is lehet írni:

	a_n	a_{n-1}	a_{n-2}	...	a_1	a_0
t	a_n	$a_n t + a_{n-1}$	$(a_n + t a_{n-1}) t + a_{n-2}$	...	$a_1 + ...$	$p(t)$

.

Ha $p(t) = 0$ akkor t gyöke a polinom függvénynek.

Felhasználva a polinomok maradékkal való osztási képletét felírhatjuk:

$$p(x) = (x - t) q(x) + r, \text{ ahol } r \text{ az osztási maradék } \Rightarrow p(t) = r.$$

72. Példa. Legyen $p(x) = 4x^3 - 3x$ és $t = 1/4$. $\Longrightarrow$

	4	0	-3	0
1/4	4	1	$-11/4$	$-11/16$

Tehát $p(1/4) = -11/16$.

A de Casteljau elrendezés ugyanazt a célt szolgálja mint a Horner elrendezés, csak a Bézier alakban felírt polinomok esetében használjuk.

Legyen

$$p(x) = c_0 b_0^n(x) + c_1 b_1^n(x) + ...c_n b_n^n(x) = \sum_{i=0}^{n} c_i b_i^n(x)$$

egy Bézier alakban felírt polinom. Akkor felhasználva a (3.9.3) rekurzív képleteket következik, hogy:

$$p(t) = c_0 b_0^n(t) + c_1 b_1^n(t) + ... + c_{n-1} b_{n-1}^n(t) + c_n b_n^n(t) =$$
$$= c_0(1-t) b_0^{n-1}(t) + c_1\left[(1-t) b_1^{n-1}(t) + t b_0^{n-1}(t)\right] + ...+$$
$$+c_{n-1}\left[(1-t) b_{n-1}^{n-1}(t) + t b_{n-1}^{n-2}(t)\right] + c_n t b_{n-1}^{n-1}(t) =$$
$$((1-t)c_0 + tc_1) b_0^{n-1}(t) + ((1-t)c_1 + tc_2) b_1^{n-1}(t) + ... + ((1-t)c_{n-1} + tc_n) b_{n-1}^{n-1}(t).$$

Jelöljük az új együtthatókat:

$$c_i^{(1)} := (1-t)c_i + tc_{i+1},\ i = 0, ..., n-1 \implies$$
$$p(t) = \sum_{i=0}^{n-1} c_i^{(1)} b_i^{n-1}(x). \tag{3.9.9}$$

Megismételve az eljárást:

$$c_i^{(2)} := (1-t)c_i^{(1)} + tc_{i+1}^{(1)},\ i = 0, ..., n-2 \implies$$
$$p(t) = \sum_{i=0}^{n-2} c_i^{(2)} b_i^{n-2}(x),$$

és az utolsó iterációban:

$$c_0^{(n)} := (1-t)c_0^{(n-1)} + tc_1^{(n-1)} \implies$$
$$p(t) = c_0^{(n)}.$$

Tehát, kezdve az előzetes együtthatókkal:

$$c_i^{(0)} := c_i,\ i = 0, ..., n$$

kiszámítjuk a lineáris kombinációkat:

$$c_i^{(k)} := (1-t)c_i^{(k-1)} + tc_{i+1}^{(k-1)},\ i = 0, ..., n-k; \tag{3.9.10}$$

az utolsó érték adja meg a $p(t)$ értékét.

Gyakorlatban az adatokat a következő táblázat szerint rendezzük:

$$\begin{array}{ccccc} c_0^{(0)} & & & & \\ c_1^{(0)} & c_0^{(1)} & & & \\ c_2^{(0)} & c_1^{(1)} & c_0^{(2)} & & \\ \dots & \dots & \dots & & \\ c_n^{(0)} & c_{n-1}^{(1)} & c_{n-2}^{(2)} & \dots & c_0^{(n)} \end{array}$$

Az első oszlop tartalmazza a p polinom eredeti együtthatóit, a második oszlop pedig a (3.9.10) képlet szerint vannak kiszámítva $k = 1$-re. A táblázat jobb-alsó sarkában lévő $c_0^{(n)}$ érték adja a $p(t)$ értéket.

Legyen p az előbbi példából $p \in \mathbb{P}_3$, és a Bézier együtthatók: $c_0 = 0$, $c_1 = -1$, $c_2 = -2$, $c_3 = 1$. A polinom értéke $t = 1/4$-ben a következő:

$$\begin{array}{cccc} 0 & & & \\ -1 & -\frac{1}{4} & & \\ -2 & -\frac{5}{4} & -\frac{1}{2} & \\ 1 & -\frac{5}{4} & -\frac{5}{4} & -\frac{11}{16} \end{array}$$

Tehát $p(1/4) = -11/16$.

3.9.3. Algebrai egyenletek numerikus megoldása Tekintsük a következő algebrai egyenletet:

$$p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 = 0. \tag{3.9.11}$$

3.9.3.1. Newton-Horner módszer A Newton-Horner módszer egy algebrai egyenlet valós gyökeinek a meghatározására szolgál. Ennek érdekében alkalmazzuk a Newton-féle módszert kihasználva a polinomok tulajdonságait.

Felhasználjuk a Newton módszerből ismert iteráló sorozatot:

$$x_{i+1} = x_i - \frac{p_n(x_i)}{p_n'(x_i)},\ i \geq 0$$

ahol kezdeti értéknek fel lehet használni például az $x_0 = 0$, vagy $x_0 = -\frac{a_0}{a_1}$. A $p_n(x_i)$ értéknek a meghatározására a Horner sémát fogjuk használni:

$$p_n(x_i) = a_0 + x_i(a_1 + \dots x_i(a_{n-1} + x_i(a_n))).$$

A polinomokra vonatkozó maradékos osztás tétele szerint:

$$p_n(x) = (x - x_i)\, p_{n-1}(x) + R, \tag{3.9.12}$$

ahol $p_{n-1} \in \mathbb{P}_{n-1}$és $R = p_n(x_i)$ az $(x - x_i)$-val osztási maradék. Jelöljük

$$p_{n-1}(x) := b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + ... + b_1 x + b_0.$$

Mint ismeretes a b_j, $j = 0, ..., n-1$ együtthatók kiszámíthatóak a (3.9.8) Horner sémából :

$$b_{n-1} = a_n$$
$$b_{n-j} = a_{n-j+1} + x_i b_{n-j+1},\ j = 2, ..., n.$$

Deriválva a (3.9.12) képletet kapjuk hogy:

$$p_n'(x) = (x - x_i)\, p_{n-1}'(x) + p_{n-1}(x)$$

ahonnan:

$$p_n'(x_i) = p_{n-1}(x_i). \tag{3.9.13}$$

	a_n	a_{n-1}	a_{n-2}	...	a_1	a_0
x_i	a_n	$a_n x_i + a_{n-1}$			$(\ldots)\, x_i + a_1$	$(\ldots)\, x_i + a_0$
	b_{n-1}	b_{n-2}			b_0	

A $p_{n-1}(x_i)$ érték kiszámításához újból felhasználjuk a Horner sémát a b_j együtthatókra:

$$p_{n-1}(x_i) = b_0 + x_i\,(b_1 + ... + x_i\,(b_{n-2} + x_i b_{n-1}))\,.$$

Az eljárást abbahagyjuk ha a kapott gyök elég pontos. Ha meghatároztuk az egyik gyököt (α) elosztjuk a p_n polinomot $(x - \alpha)$-val es a kapott $n-1$-ed fokú polinomra újból alkalmazzuk a Newton-Horner módszert. Legvégül másodfokú polinomot kapunk aminek meghatározzuk a gyökeit.

73. Példa. Határozzuk meg az alábbi egyenlet valós gyökeit ($x_0 = 0$):

$$x^4 - 10x^3 + 35x^2 - 50x + 24 = 0.$$

$$x_1 = x_0 - \frac{p_4(0)}{p_4'(0)} = 0 + \frac{24}{50} = 0.48,\ |x_1 - x_0| = 0.48$$

	1	−10	35	−50	24
0	1	−10	35	−50	24
0	1	−10	35	−50	

$$x_2 = x_1 - \frac{p_4(0.48)}{p_4'(0.48)} = 0.48 - \frac{7.011}{-22.869} = 0.78,\ |x_2 - x_1| = 0.3\ ...$$

	1	−10	35	−50	24
0.48	1	−9.52	30.43	−35.39	7.011
0.48	1	−9.04	26.09	−22.869	

.

3.9.3.2. Bairstow módszer A P polinomot $x^2 + px + q$ polinommal osztjuk, majd a másodfokú polinom gyökeit meghatározzuk.

3.10. Szélsőérték számítás

Egy f függvény helyi szélsőértékeit (min vagy max) a stacionárius pontjai között keressük: $f'\{x\} = 0$. Tehát, ha a függvény deriváltja ismert, akkor a szélsőérték meghatározása visszavezetődik egyenletek megoldására. A továbbiakban olyan módszereket tárgyalunk amelyeknél nem szükséges a derivált ismerete.

3.10.1. Harmadoló módszer A harmadoló módszer, a felező módszerhez hasonlóan, a f függvény minimumát tartalmazó $x_{min} \in [a, b]$ intervallumok folyamatos szűkítéséből áll. Az eljárás hasonló a felező módszerhez, viszont ebben az esetben két részintervallum nem vezet eredményre.

Az $[a, b]$ intervallum harmadolásához kiszámítjuk a h lépést (az $[a, b]$ intervallum harmadát):

$$h = \frac{1}{3}(b - a), \tag{3.10.1}$$

majd előállítjuk a végektől $\frac{1}{3}$ -ra lévő értékeket:

$$u = a + h,\ v = b - h. \tag{3.10.2}$$

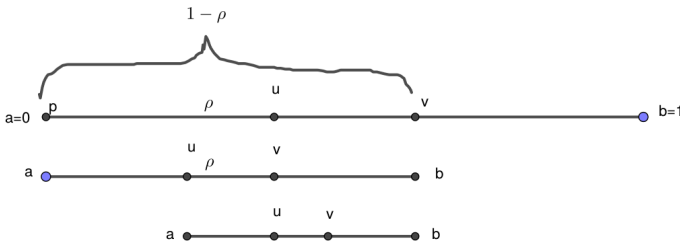
Az f függvény x_{min} szélsőértéke vagy az $[a, v]$, vagy $[u, b]$ intervallumban található, attól függően, hogy $f(u) \leq f(v)$ vagy $f(u) \geq f(v)$. Az eljárást folytatjuk az új intervallummal, vagyis az első esetben $b := v$, a másodikban pedig $a := u$.

74. PÉLDA. $f(x) = x + \frac{1}{x} \longrightarrow min,\ x \in [0, 3]$.

Ezzel az eljárással minden lépésben az intervallum az előbbinek a $\frac{2}{3}$-ra zsugorodik, tehát az algoritmus lineáris. Kilépési kritériumként felhasználható, hogy az $[a, b]$ intervallum hossza kisebb mint egy adott pontosság:

$$|b - a| < \epsilon.$$

3.10.2. Aranymetszés módszer A harmadoló módszer esetében minden iterációban az f függvényt kétszer kellett kiértékelni ($f(u), f(v)$). Ha a harmadolás helyett egy másik, ρ arányt használunk lehetségessé válik, hogy az u/v és vele együtt az $f(u)/f(v)$ értéket újból felhasználjuk.



A keresett arány kiszámításához figyelembe vesszük hogyan aránylik a nagy szakasz a kicsi szakaszhoz.

$$\frac{1}{1-\rho} = \frac{1-\rho}{\rho},$$

vagyis

$$\rho^2 - 3\rho + 1 = 0,$$

aminek az egység alatti megoldása

$$\rho = \frac{3 - \sqrt{5}}{2}.$$

A ρ arány kifejezhető a $\phi = \frac{1+\sqrt{5}}{2} \approx 1.618$ aranymetszet szám segítségével:

$$\rho = 2 - \phi \approx 0.381966. \tag{3.10.3}$$

4. FEJEZET

Egyenletrendszerek numerikus megoldása

4.1. Lineáris egyenletrendszerek

Egy

$$\left\{\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n &=& b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n &=& b_2 \\ & \dots & \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n &=& b_m \end{array}\right. ,$$

általános $m \times n$ lineáris egyenletrendszer

$$Ax = b, \tag{4.1.1}$$

mátrix alakba írható, ahol

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & ... & a_{mn} \end{pmatrix}, \; b = \begin{pmatrix} b_1 \\ b_2 \\ \\ \vdots \\ b_m \end{pmatrix}, \; x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Először a négyzetes $m = n$ egyenletrendszerek megoldásával foglalkozunk, majd tárgyaljuk az $m \neq n$ esetet is.

Egy $A \in \mathcal{M}_n(\mathbb{R})$ $n \times n$ egyenletrendszer esetében nem alkalmazhatjuk a Cramer szabályt, mert egy $n \times n$ típusú determináns kiszámításának műveletigénye $n!$ Tehát, egy 100×100 determináns (klasszikus) kiszámításához kb. $100! = 10^{157.9}$ művelet szükséges.

A számítások alatt a hibák összegeződnek ezért - mérettől függően - különböző technikákat alkalmazunk. Ha az ismeretlenek száma kisebb mint 10^3, akkor pontos (direkt) módszereket használunk például Gauss, illetve faktorizációs módszereket LU, LL^t, QR, stb. Ha az ismeretlenek száma 10^3 és 10^6 között van, akkor iteratív módszereket

használunk (Jacobi, Gauss-Seidel, stb). Ha az ismeretlenek száma nagyobb mint 10^6, akkor valószínűségi módszereket alkalmazunk (Monte Carlo módszer).

4.1.1. Direkt módszerek A direkt módszerek esetében az egyenletrendszer mátrixát transzformációk segítségével sajátos alakra hozzuk (általában háromszög alakra), amit könnyebben lehet kezelni. A direkt módszereknél pontosan ismerjük hány iteráció szükséges ennek a kivitelezésére.

4.1.1.1. Háromszög alakú egyenletrendszerek Az

$$\left\{\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n & = & b_1 \\ a_{22}x_2 + ... + a_{2n}x_n & = & b_2 \\ \ddots & & \\ a_{nn}x_n & = & b_n \end{array}\right. \tag{4.1.2}$$

sajátos lineáris egyenletrendszert felső-háromszögűnek nevezzük ha az egyenletrendszer A mátrixában a főátló alatt csak nullák vannak:

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ 0 & a_{22} & ... & a_{2n} \\ \vdots & & & \\ 0 & 0 & ... & a_{nn} \end{pmatrix}.$$

Ha $a_{ii} \neq 0,\ i = \overline{1,n}$, akkor a (4.1.2) megoldását visszahelyettesítéssel kapjuk meg:

$$\begin{array}{rcl} x_n & = & \dfrac{b_n}{a_{nn}}, \\ x_k & = & \dfrac{b_k - \sum_{i=k+1}^{n} a_{ki}x_i}{a_{kk}},\ k = \overline{n-1,1}. \end{array} \tag{4.1.3}$$

Az

$$\text{(4.1.4)} \qquad \begin{cases} a_{11}x_1 & = & b_1 \\ a_{21}x_1 + a_{22}x_2 & = & b_2 \\ \qquad\qquad \ddots & & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n & = & b_n \end{cases}$$

sajátos lineáris egyenletrendszert alsó-háromszögűnek nevezzük ha az egyenletrendszer A mátrixában a főátló fölött csak nullák vannak:

$$A = \begin{pmatrix} a_{11} & 0 & \ldots & 0 \\ a_{21} & a_{22} & \ldots & 0 \\ \vdots & & & \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}.$$

Ha $a_{ii} \neq 0,\ i = \overline{1,n}$, akkor a (4.1.4) megoldása:

$$\text{(4.1.5)} \qquad \begin{aligned} x_1 &= \frac{b_1}{a_{11}}, \\ x_k &= \frac{b_k - \sum_{i=1}^{k-1} a_{ki}x_i}{a_{kk}},\ k = \overline{2,n}. \end{aligned}$$

A visszahelyettesítés műveletigénye:

- az x_1 kiszámítása 0 összeadást és 1 szorzást igényel
- az x_2 kiszámítása 1 összeadást és 2 szorzást igényel
- az x_n kiszámítása $(n-1)$ összeadást és n szorzást igényel, tehát összesen

$$(0+1) + (1+2) + \ldots ((n-1)+n) = \sum_{k=1}^{n} ((k-1)+k)$$

$$\text{(4.1.6)} \qquad = -\sum_{k=1}^{n} 1 + 2\sum_{k=1}^{n} k = -n + 2\frac{n\,(n+1)}{2} = n^2.$$

4.1.1.2. A Gauss kiküszöbölési módszer A Gauss módszer a lineáris egyenletrendszerek numerikus megoldásának egyik legrégibb sémája. Lineáris transzformációkat alkalmazva a sorok között, a (4.1.1) egyenletrendszert visszavezetjük a (4.1.2) háromszög alakú egyenletrendszerre.

Az egyszerűség kedvéért az eljárást egy 4×4-es $Ax = b$ egyenletrendszeren mutatjuk be:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}. \tag{4.1.7}$$

Ha $a_{11} \neq 0$, akkor az első sort rendre

$$-\frac{a_{21}}{a_{11}} =: \widetilde{l}_{21},\ -\frac{a_{31}}{a_{11}} =: \widetilde{l}_{31},\ -\frac{a_{41}}{a_{11}} =: \widetilde{l}_{41} \tag{4.1.8}$$

értékekkel szorozzuk (a szabadtagokat is beleértve), majd (rendre) hozzáadjuk a második, harmadik, illetve negyedik sorhoz:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & a_{42}^{(2)} & a_{43}^{(2)} & a_{44}^{(2)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2^{(2)} \\ b_3^{(2)} \\ b_4^{(2)} \end{pmatrix} \tag{4.1.9}$$

ahol

$$a_{ij}^{(2)} = a_{1j}\left(-\frac{a_{i1}}{a_{11}}\right) + a_{ij}, i, j \geq 2, \acute{e}s b_i^{(2)} = b_1\left(-\frac{a_{i1}}{a_{11}}\right) + b_i,\ i \geq 2.$$

Ha $a_{22}^{(2)} \neq 0$, akkor a második lépésben a második sort szorozzuk be a

$$-\frac{a_{32}^{(2)}}{a_{22}^{(2)}} =: \widetilde{l}_{32},\ -\frac{a_{42}^{(2)}}{a_{22}^{(2)}} =: \widetilde{l}_{42} \tag{4.1.10}$$

értékekkel, majd hozzáadjuk a harmadik, illetve negyedik sorhoz:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & a_{34}^{(3)} \\ 0 & 0 & a_{43}^{(3)} & a_{44}^{(3)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2^{(2)} \\ b_3^{(3)} \\ b_4^{(3)} \end{pmatrix}. \tag{4.1.11}$$

Az eljárást mindaddig folytatjuk míg egy felső-háromszög lineáris egyenletrendszert kapunk:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & a_{34}^{(3)} \\ 0 & 0 & 0 & a_{44}^{(4)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2^{(2)} \\ b_3^{(3)} \\ b_4^{(4)} \end{pmatrix} \tag{4.1.12}$$

amit a (4.1.3) képletnek megfelelően visszahelyettesítéssel oldunk meg.

Az általános $n \times n$-es egyenletrendszer $k+1$-ik lépésben a kiküszöbölési séma a következő:

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} + a_{kj}^{(k)} \left(-\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \right), \ i,j \geq k+1, \tag{4.1.13}$$

$$b_{i}^{(k+1)} = b_{i}^{(k)} + b_{k}^{(k)} \left(-\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \right), \ i \geq k+1. \tag{4.1.14}$$

Egy $n \times n$-es lineáris egyenletrendszer Gauss-kiküszöbölési algoritmus műveletigénye a következő:

- az első iterációban $(n-1)$ nulla előállításához 1(szorzó előállítása)$+$ n(szorzás) $+\, n$(összeadás) műveletre van szükség (a főelemmel való szorzást/összeadást nem számítjuk hiszen az eredmény nulla)
- a második iterációban $(n-2)$ nulla előállításához $1+(n-1)+$ $(n-1)$ műveletre van szükség
- az utolsó ($(n-1)$-ik) iterációban egy nulla előállításához $1+$ $2+2$ műveletre van szükség, tehát összesen:

$$\begin{aligned} &(n-1)\cdot(1+2n)+(n-2)\cdot(1+2\,(n-1))+1\cdot(1+2\cdot 2)= \\ &\qquad = \sum_{k=1}^{n-1} k\cdot(1+2\,(k+1)) = 3\sum_{k=1}^{n-1} k + 2\sum_{k=1}^{n-1} k^2 = \\ &= 3\frac{(n-1)\,n}{2} + 2\frac{(n-1)\,n\,(2n-1)}{6} = \frac{(n-1)\,n\,(4n+7)}{6} = \\ &\qquad = \frac{2}{3}n^3 + \frac{1}{2}n^2 + \frac{7}{6}n = \frac{2}{3}n^3 + O\left(n^2\right) = O\left(n^3\right). \end{aligned} \tag{4.1.15}$$

Összehasonlítva a visszahelyettesítési műveletigennyél (n^2) azt a következtetést lehet levonni, hogy az egyenletrendszer megoldásában a kiküszöbölés jóval több műveletet igényel. Például, ha $n = 100$, akkor a kiküszöbölés 10^6 nagyságrendű, míg a visszahelyettesítés 10^4 műveletet igényel.

75. DEFINÍCIÓ. Az $a_{11}^{(1)}$, $a_{22}^{(2)}$, $a_{33}^{(3)}$, $a_{44}^{(4)}$ nem-nulla tagokat főelemeknek nevezzük.

Ha valamely $a_{ii}^{(i)}$ nulla - például a (4.1.9) egyenletrendszerben ha $a_{22}^{(2)} = 0$ - a séma nem alkalmazható, ugyanis az $\widetilde{l}_{32}, \widetilde{l}_{42}$ szorzókat nem képezhetjük. Tételezzük fel, hogy $a_{22}^{(2)}$ oszlopában ($a_{22}^{(2)}$ alatt) vannak nem-nulla tagok. Ebben az esetben felcserélhetjük az $a_{22}^{(2)}$ sorát a nem-nulla tag sorával mert az egyenletrendszer megoldása nem módosul. Ezt az eljárást (részleges) főelemkiválasztásnak nevezzük. Az új főelem kiválasztásánál az első nem-nulla tag helyett ajánlatos az abszolút értékben legnagyobb tagot választani ($\max\left\{\left|a_{32}^{(2)}\right|, \left|a_{42}^{(2)}\right|\right\}$ a mi esetünkben), ugyanis így elkerülhető az értékes számjegyek vesztése az osztás miatt.

Ha $a_{22}^{(2)} = 0$ oszlopában ($a_{22}^{(2)}$ alatt) minden elem nulla, akkor az A mátrix szinguláris, ugyanis

$$\det A = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & 0 & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} \end{vmatrix} = a_{11} \begin{vmatrix} 0 & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & a_{43}^{(2)} & a_{44}^{(2)} \end{vmatrix} = 0,$$

vagyis a lineáris egyenletrendszer b-től függően összeférhetetlen vagy határozatlan.

Ha a főelemet a $\begin{pmatrix} a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ a_{32}^{(2)} & a_{33}^{(2)} & a_{34}^{(2)} \\ a_{42}^{(2)} & a_{43}^{(2)} & a_{44}^{(2)} \end{pmatrix}$ részmátrixból választjuk ki, akkor az eljárást teljes főelemkiválasztásnak nevezzük. Ebben az esetben egy sor és egy oszlop cserét kell végrehajtani. Az oszlop cserét figyelembe kell venni az algoritmus végén a megoldások sorrendjének a leolvasásánál.

Ha a kiküszöbölést nem csak az átló alatt hanem fölötte is elvégezzük, akkor egy diagonális lineáris egyenletrendszerhez jutunk. Ez a séma Gauss-Jordan néven ismert. Habár a diagonális lineáris egyenletrendszert egyszerűbb megoldani mint egy háromszög egyenletrendszert ($O(n)$ összehasonlítva $O(n^2)$ művelettel), a Gauss-Jordan kiküszöbölési műveletigénye kétszer nagyobb a Gauss algoritmusénál, ezért összességében a Gauss algoritmus gazdaságosabb mint a Gauss-Jordan.

A Gauss algoritmus alkalmas determináns kiszámítására is ugyanis

$$\det A = \det A^{(1)} = ... = \det A^{(n)} = a_{11}a_{22}^{(2)}...a_{nn}^{(n)}, \tag{4.1.16}$$

ahol $A^{(k)}$ a k iterációban, Gauss módszerrel generált mátrix ($A^{(1)} = A$).

76. Példa. Az

$$\begin{pmatrix} 10 & 3 & -1 \\ 7 & 2 & 0 \\ -5 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \\ -1 \end{pmatrix} \tag{4.1.17}$$

egyenletrendszer első oszlopának a kiküszöbölésére az első sorát beszorozzuk $\widetilde{l}_{21} = -\frac{7}{10}$, illetve $\widetilde{l}_{31} = \frac{5}{10}$-el, majd hozzáadjuk a második, illetve harmadik sorhoz

$$\Longrightarrow \begin{pmatrix} 10 & 3 & -1 \\ 0 & \frac{-1}{10} & \frac{7}{10} \\ 0 & \frac{35}{10} & \frac{25}{10} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ \frac{15}{10} \\ \frac{15}{10} \end{pmatrix}.$$

A következő iterációban a második sort szorozzuk be $\widetilde{l}_{23} = 35$-el és hozzáadjuk a harmadik sorhoz

$$\Longrightarrow \begin{pmatrix} 10 & 3 & -1 \\ 0 & \frac{-1}{10} & \frac{7}{10} \\ 0 & 0 & 27 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ \frac{15}{10} \\ 54 \end{pmatrix},$$

majd a felső-háromszög egyenletrendszert visszahelyettesítéssel megoldjuk: $x_3 = 2,\ x_2 = -1,\ x_1 = 1$.

Az A mátrix determinánsát a (4.1.16) képlet szerint az $A^{(3)} = \begin{pmatrix} 10 & 3 & -1 \\ 0 & \frac{-1}{10} & \frac{7}{10} \\ 0 & 0 & 27 \end{pmatrix}$

mátrix főátlóján lévő tagok szorzata adja:

$$\det A = \det A^{(3)} = 10 \cdot \left(-\tfrac{1}{10}\right) \cdot 27 = -27.$$

A Gauss módszert alkalmazhatjuk egy adott A mátrixnak az inverzének a meghatározására. Ennek érdekében figyelembe vesszük, hogy

$$A \cdot A^{-1} = I_n,$$

tehát

$$A \cdot (X_1 \,|X_2|\, ...\, |X_n) = \begin{pmatrix} 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ & & \ddots & \\ 0 & 0 & & 1 \end{pmatrix} \tag{4.1.18}$$

ahol X_k az A^{-1} inverz mátrixnak a k-ik oszlopa $X_k = \begin{pmatrix} x_k^1 \\ x_k^2 \\ \vdots \\ x_k^n \end{pmatrix}$.

Oszlopokra bontva a (4.1.18)-ből következik, hogy:

$$A \cdot X_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},\ A \cdot X_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, ...,\ A \cdot X_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

vagyis az inverz mátrixnak az $X_k,\ k = \overline{1,n}$ oszlopainak a meghatározására n-szer alkalmazzuk a Gauss módszert egyforma A mátrixszal és különböző szabad-tag vektorral.

77. Példa. Az $A = \begin{pmatrix} 10 & 3 & -1 \\ 7 & 2 & 0 \\ -5 & 2 & 3 \end{pmatrix}$ mátrix inverzét $A^{-1} = (X\,|Y|\,Z)$ három lineáris egyenletrendszer megoldásából kapjuk:

$$\begin{pmatrix} 10 & 3 & -1 \\ 7 & 2 & 0 \\ -5 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -\frac{2}{9} \\ \frac{7}{9} \\ -\frac{8}{9} \end{pmatrix},$$

$$\begin{pmatrix} 10 & 3 & -1 \\ 7 & 2 & 0 \\ -5 & 2 & 3 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \Longrightarrow Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \frac{11}{27} \\ -\frac{25}{27} \\ \frac{35}{27} \end{pmatrix},$$

$$\begin{pmatrix} 10 & 3 & -1 \\ 7 & 2 & 0 \\ -5 & 2 & 3 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \Longrightarrow Z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} -\frac{2}{27} \\ \frac{7}{27} \\ \frac{1}{27} \end{pmatrix},$$

tehát $A^{-1} = \begin{pmatrix} -\frac{2}{9} & \frac{11}{27} & -\frac{2}{27} \\ \frac{7}{9} & -\frac{25}{27} & \frac{7}{27} \\ -\frac{8}{9} & \frac{35}{27} & \frac{1}{27} \end{pmatrix}$.

4.1.1.3. LU faktorizáció LU faktorizáción egy $A = (a_{i,j})_{i,j=1,n}$ mátrix

$$A = L \cdot U \tag{4.1.19}$$

szorzatra való felbontását értjük, ahol L egy alsó- U pedig egy felső-háromszög mátrix:

$$L = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ \vdots & & & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix}, \; U = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \vdots & & & \\ 0 & 0 & \dots & u_{nn} \end{pmatrix}.$$

A két mátrix meghatározására $n^2 + n$ elem szükséges, ugyanakkor a (4.1.19)-ban csak n^2 egyenlettel rendelkezünk. Ha az L vagy az U mátrix főátlóját 1-el tesszük egyenlővé a két mátrix egyértelműen meghatározható. Az első eset $l_{ii} = 1,\ i = \overline{1,n}$ Doolittle, míg a második $u_{ii} = 1,\ i = \overline{1,n}$ Crout faktorizáció néven ismert.

A faktorizációs eljárásban a kiszámítási sorrendet úgy állítjuk fel, hogy az éppen sorra kerülő ismeretlen kiszámításánál minden adat rendelkezésünkre álljon. A Doolittle faktorizáció esetében alkalmazhatjuk például a sor-szerinti rendezést, vagyis a k-ik lépésben kiszámítjuk a

$$L = \begin{pmatrix} 1 & 0 & ... & 0 \\ l_{21} & 1 & ... & 0 \\ \vdots & & & \\ l_{n1} & l_{n2} & ... & 1 \end{pmatrix}, \; U = \begin{pmatrix} u_{11} & u_{12} & ... & u_{1n} \\ 0 & u_{22} & ... & u_{2n} \\ \vdots & & & \\ 0 & 0 & ... & u_{nn} \end{pmatrix} \tag{4.1.20}$$

mátrixok k-ik sorát $l_{k1}, l_{k2}, ..., l_{k,k-1}, u_{kk}, u_{k.k+1}, ..., u_{k,n}$ (ebben a sorrendben)

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}}{u_{jj}}, \; i > j, \tag{4.1.21}$$

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}, \; i \leq j. \tag{4.1.22}$$

78. Példa. Az $A = \begin{pmatrix} 10 & 3 & -1 \\ 7 & 2 & 0 \\ -5 & 2 & 3 \end{pmatrix}$ mátrix esetében az L és U mátrixok tagjait a következő sorrendben számítjuk ki:

$$u_{11}, u_{12}, u_{13}; \; l_{21}, u_{22}, u_{23}; \; l_{31}, l_{32}, u_{33}$$

és a következő mátrixokat kapjuk:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{7}{10} & 1 & 0 \\ -\frac{5}{10} & -35 & 1 \end{pmatrix}, U = \begin{pmatrix} 10 & 3 & -1 \\ 0 & -\frac{1}{10} & \frac{7}{10} \\ 0 & 0 & 27 \end{pmatrix}.$$

Ha az A mátrixot felbontottuk, akkor az

$$Ax = b,$$

egyenletrendszert átírhatjuk

$$LUx = b, \tag{4.1.23}$$

alakra, ami ekvivalens az alábbi két háromszög lineáris egyenletrendszerrel:

$$\begin{aligned} Ly &= b, \\ Ux &= y. \end{aligned} \tag{4.1.24}$$

E két lineáris egyenletrendszer (alsó-felső) megoldásához használjuk a már ismert (4.1.5), (4.1.3) sémákat.

Az LU felbontás alkalmas a determináns kiszámítására is

$$\det A = \det(LU) = \det L \cdot \det U = \det U = \prod_{i=1}^{n} u_{ii}. \tag{4.1.25}$$

79. Példa. Ismerve az $A = LU$ felbontását az (4.1.17) példában szereplő A mátrixnak, a (4.1.24)-ból kapjuk, hogy

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{7}{10} & 1 & 0 \\ -\frac{5}{10} & -35 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \\ -1 \end{pmatrix},$$

aminek megoldása $y = \begin{pmatrix} 5 \\ \frac{3}{2} \\ 54 \end{pmatrix}$, majd az

$$\begin{pmatrix} 10 & 3 & -1 \\ 0 & -\frac{1}{10} & \frac{7}{10} \\ 0 & 0 & 27 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ \frac{3}{2} \\ 54 \end{pmatrix}$$

egyenletrendszerből $x = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$.

Ugyanakkor a (4.1.25) képletből kiszámítható A determinánsa:

$$\det A = \det U = -27.$$

4.1.1.4. A Gauss és az LU módszer kapcsolata Az LU faktorizációs módszer levezethető a Gauss módszerből, ugyanis ha a Gauss módszerben ismert szorzók $\widetilde{l}_{ik} = -\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ segítségével megszerkesztjük az $L^{(k)}$

elemi alsó-háromszög mátrixot:

$$L^{(k)} = \begin{pmatrix} 1 & & & & & \\ \vdots & \ddots & & & & \\ 0 & & 1 & & & \\ 0 & & \widetilde{l}_{k+1,k} & \ddots & & \\ \vdots & & \vdots & & 1 & \\ 0 & & \widetilde{l}_{nk} & & 0 & 1 \end{pmatrix}, \tag{4.1.26}$$

akkor

$$A^{(k+1)} = L^{(k)\cdot} A^{(k)}, \tag{4.1.27}$$

ahol $A^{(k)}$ a k iterációban, Gauss módszerrel generált mátrix ($A^{(1)} = A$).

Tehát

$$U = A^{(n)} = L^{(n-1)\cdot} L^{(n-2)} \cdot \ldots \cdot L^{(1)} A \tag{4.1.28}$$

ahonnan

$$A = \left(L^{(1)}\right)^{-1} \ldots \left(L^{(n-2)}\right)^{-1} \left(L^{(n-1)}\right)^{-1} U. \tag{4.1.29}$$

Az $L^{(k)}$ mátrix egyszerű szerkezetének köszönhetően könnyen kiszámítható az inverze:

$$\left(L^{(k)}\right)^{-1} = \begin{pmatrix} 1 & & & & & \\ \vdots & \ddots & & & & \\ 0 & & 1 & & & \\ 0 & & -\widetilde{l}_{k+1,k} & \ddots & & \\ \vdots & & \vdots & & 1 & \\ 0 & & -\widetilde{l}_{nk} & & 0 & 1 \end{pmatrix} \tag{4.1.30}$$

és igazolható, hogy
(4.1.31)

$$\left(L^{(1)}\right)^{-1}\ldots\left(L^{(n-2)}\right)^{-1}\left(L^{(n-1)}\right)^{-1}=\begin{pmatrix}1&&&&\\-\widetilde{l}_{21}&1&&&\\-\widetilde{l}_{31}&-\widetilde{l}_{32}&\ddots&&\\-\widetilde{l}_{41}&-\widetilde{l}_{42}&&&\\\vdots&&&1&\\-\widetilde{l}_{n1}&-\widetilde{l}_{n2}&\cdots&-\widetilde{l}_{n,n-1}&1\end{pmatrix}=L$$

vagyis (4.1.29)-ből $A=LU$.

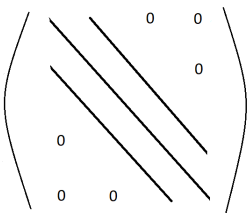
80. Példa. Az előbbi példában $A=A^{(1)}=\begin{pmatrix}10&3&-1\\7&2&0\\-5&2&3\end{pmatrix}$ a Gauss a szorzók rendre $\widetilde{l}_{21}=\left(-\frac{7}{10}\right),\widetilde{l}_{31}=\left(\frac{5}{10}\right)$, majd $\widetilde{l}_{32}=35$ voltak
$\Longrightarrow$

$$L^{(1)}=\begin{pmatrix}1&0&0\\-\frac{7}{10}&1&0\\\frac{5}{10}&0&1\end{pmatrix}\Longrightarrow A^{(2)}=L^{(1)}A^{(1)}=\begin{pmatrix}10&3&-1\\0&-\frac{1}{10}&\frac{7}{10}\\0&\frac{7}{2}&\frac{5}{2}\end{pmatrix},$$

$$L^{(2)}=\begin{pmatrix}1&0&0\\0&1&0\\0&35&1\end{pmatrix}\Longrightarrow A^{(3)}=L^{(2)}A^{(2)}=\begin{pmatrix}10&3&-1\\0&-\frac{1}{10}&\frac{7}{10}\\0&0&27\end{pmatrix}=U,$$

$$L=\left(L^{(1)}\right)^{-1}\left(L^{(2)}\right)^{-1}=\begin{pmatrix}1&0&0\\\frac{7}{10}&1&0\\-\frac{5}{10}&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\0&-35&1\end{pmatrix}=\begin{pmatrix}1&0&0\\\frac{7}{10}&1&0\\-\frac{5}{10}&-35&1\end{pmatrix}.$$

4.1.1.5. Tridiagonális egyenletrendszerek A gyakorlatban gyakran találkozunk ritka mátrixokkal, vagyis olyan mátrixokkal amelyeknek legtöbb elemük nulla. Ilyen például a diagonális, illetve tridiagonális le. Tridiagonálisnak nevezzük azt a négyzetes mátrixot amelynek csak a főátlóján, illetve az alatta és fölötte lévő átlón van nem-nulla eleme:



4.1.1. ábra. Tridiagonális mátrix

vagyis

$$\begin{pmatrix} a_1 & c_1 & 0 & 0 & 0 \\ b_1 & a_2 & c_2 & 0 & 0 \\ 0 & b_2 & \ddots & & 0 \\ 0 & 0 & & a_{n-1} & c_{n-1} \\ 0 & 0 & 0 & b_{n-1} & a_n \end{pmatrix}$$

81. Példa. Oldjuk meg Gauss kiküszöböléssel az alábbi tridiagonális egyenletrendszert:

$$\begin{cases} 2x_1 & +x_2 & & & = & -1 \\ 2x_1 & +3x_2 & +x_3 & & = & 3 \\ & x_2 & +4x_3 & +2x_4 & = & 5 \\ & & x_3 & +3x_4 & = & -4 \end{cases}.$$

Egy tridiag le megoldásánál a () képletben tárgyalt műveletnél jóval kevesebb műveletre van szükség. Pontosabban, minden iterációban 1(szorzó előállítása) $+ 2$(szorzás) $+ 2$(összeadás) műveltre van szükség, összesen $5\,(n-1)$. A visszahelyettesítéshez minden x_i kiszámításához 1(összeadás) $+ 1$(szorzás) műveltre van szükség, összesen $2n$. Tehát, összességében egy tridiag le műveletigénye lineáris $O\,(n)$.

4.1.2. Lineáris egyenletrendszerek iteratív megoldása Az iteratív eljárás a numerikus számítások egyik alapvető módszerének számít.

Az $(X_n)_n$ megoldások sorozatát fokozatosan közelítjük a:

$$X^{(k+1)} = \Phi\left(X^{(k)}\right) \tag{4.1.32}$$

iterációkat használva, vagy $X^{(k+1)} = \Phi\left(X^{(k)}, X^{(k-1)}, ...\right)$ ha az eljárás többlépéses.

A direkt módszerekkel ellentétben az iteratív módszerekben az iterációk száma nem a mátrix méretétől függ, hanem egy adott $\epsilon > 0$ pontosságtól. ?Míg a direkt módszerekben minden iterációval a hiba kumulálodott, addig az iteratív eljárásban minden iterációval közelebb kerülünk a keresett megoldáshoz.?

4.1.2.1. Jacobi-féle módszer Ahhoz, hogy az $AX = b$ lineáris egyenletrendszert (4.1.32) alakra hozzuk kifejezzük az egyenletrendszer minden i-edik sorában az x_i ismeretlent:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 \ + ... + a_{2n}x_n &= b_2 \\ \ \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n &= b_n \end{cases} \Rightarrow$$

$$\begin{cases} x_1 = & -\frac{a_{12}}{a_{11}}x_2 & \cdots & -\frac{a_{1n}}{a_{11}}x_n & +\frac{b_1}{a_{11}} \\ x_2 = -\frac{a_{21}}{a_{22}}x_1 & & \cdots & -\frac{a_{2n}}{a_{22}}x_n & +\frac{b_2}{a_{22}} \\ \ \vdots & & & & \\ x_n = -\frac{a_{n1}}{a_{nn}}x_1 & -\frac{a_{n2}}{a_{nn}}x_2 & \cdots & & +\frac{b_n}{a_{nn}} \end{cases},$$

vagy mátrix alakban:

$$X = BX + C, \tag{4.1.33}$$

ahol

(4.1.34)
$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \; B = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \dots & -\frac{a_{2n}}{a_{22}} \\ \vdots & & & \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \dots & 0 \end{pmatrix}, \; C = \begin{pmatrix} \frac{b_1}{a_{11}} \\ \frac{b_2}{a_{22}} \\ \vdots \\ \frac{b_n}{a_{nn}} \end{pmatrix}.$$

Φ-vel jelölve a (4.1.33) képlet jobboldalát

$$\Phi(X) = BX + C, \tag{4.1.35}$$

az $X = \Phi(X)$ iteratív eljáráshoz jutunk, vagyis

$$X^{(k+1)} = \Phi\left(X^{(k)}\right), \; k \in \mathbb{N} \tag{4.1.36}$$

ahol $X^{(0)}$ egy kezdeti megoldás.

A Banach-féle fixpont tételnek megfelelően ha Φ =kontrakció az $\left(X^{(k)}\right)_k$ megoldások sorozata konvergál tetszőleges $X^{(0)}$ kezdeti megoldásra. Felhasználva a (4.1.35) képletet következik, hogy

$$\|\Phi(X) - \Phi(Y)\| = \|BX + C - BY - C\| = \|B(X - Y)\| \leq \|B\| \|X - Y\|,$$

tehát ha $\|B\| < 1$, akkor Φ kontrakció és az eljárás konvergál.

82. Tétel. *Ha $\|B\| < 1$ akkor a (4.1.36) fokozatos közelítések sorozata konvergens.*

A tételben szereplő elégséges feltétel átírható az eredeti A mátrixra. $\|\cdot\|_\infty$ típusú normát használva következik, hogy:

$$\|B\|_\infty = \max \sum_{i,j=1, j\neq i}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1, \; i = \overline{1,n},$$

ami ekvivalens az alábbi feltétellel:

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| < |a_{ii}|, \; i = \overline{1,n}. \tag{4.1.37}$$

83. Definíció. Egy A mátrixok átlósan dominánsnak nevezünk ha eleget tesz az (4.1.37) kikötésnek.

84. Példa. Az $A = \begin{pmatrix} -7 & 2 & -2 \\ 3 & 5 & -1 \\ 0 & 2 & -3 \end{pmatrix}$ mátrix átlósan domináns.

85. Következmény. *Ha az $Ax = b$ egyenletrendszer A mátrixa átlósan domináns, akkor az (4.1.36)-ban megadott iterációk konvergálnak a lineáris egyenletrendszer megoldásához.*

86. Példa. $\begin{cases} -7x_1 + 2x_2 - 2x_3 &= -1 \\ 3x_1 + 5x_2 - x_3 &= 14 \\ 2x_2 - 3x_3 &= 7 \end{cases}$, $AX = b \Leftrightarrow X = BX + C$ ahol $B = \begin{pmatrix} 0 & \frac{2}{7} & \frac{-2}{7} \\ \frac{-3}{5} & 0 & \frac{1}{5} \\ 0 & \frac{2}{3} & 0 \end{pmatrix}$, $C = \begin{pmatrix} \frac{1}{7} \\ \frac{14}{5} \\ \frac{-7}{3} \end{pmatrix}$. Ha $X^{(0)} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$ egy tetszőleges kezdeti megoldás, akkor

$$X^{(1)} = BX^{(0)} + C = \begin{pmatrix} \frac{1}{7} \\ 3 \\ -\frac{5}{3} \end{pmatrix} \simeq \begin{pmatrix} 0.14 \\ 3 \\ -1.67 \end{pmatrix}$$

és az abszolút eltérés a két megoldás között

$$\left\| X^{(1)} - X^{(0)} \right\|_\infty = \frac{8}{3} \simeq 2.67.$$

Hasonlóan

$$X^{(2)} = BX^{(1)} + C = \begin{pmatrix} \frac{31}{21} \\ \frac{50}{21} \\ -\frac{1}{3} \end{pmatrix} \simeq \begin{pmatrix} 1.48 \\ 2.38 \\ -0.33 \end{pmatrix},$$

$$X^{(3)} = BX^{(2)} + C = \begin{pmatrix} \frac{45}{49} \\ \frac{194}{105} \\ -\frac{47}{63} \end{pmatrix} \simeq \begin{pmatrix} 0.92 \\ 1.85 \\ -0.75 \end{pmatrix},$$

és a hibák

$$\left\| X^{(2)} - X^{(1)} \right\|_\infty = \frac{4}{3} \simeq 1.33,$$

$$\left\| X^{(3)} - X^{(2)} \right\|_\infty = \frac{82}{147} \simeq 0.55.$$

Mivel az A mátrix átlósan domináns, az $\left(X^{(k)}\right)_k$ megoldások sorozata konvergál a pontos megoldáshoz:

$$\lim_{k\to\infty} X^{(k)} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}.$$

A $(k+1)$-ik iterációban az

$$X^{(k+1)} = \begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \vdots \\ x_n^{(k+1)} \end{pmatrix},$$

megoldás komponenseit az alábbi képlettel számítjuk ki:

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - a_{i1}x_1^{(k)} - a_{i2}x_2^{(k)} - \ldots / \ldots - a_{in}x_n^{(k)}\right) = \tag{4.1.38}$$

$$= \frac{1}{a_{ii}}\left(b_i - \sum_{j=1, j\neq i} a_{ij}x_j^{(k)}\right),\ i = \overline{1,n}.$$

A (4.1.37) egy elégséges de nem szükséges feltétel, vagyis átlósan nem-domináns lineáris egyenletrendszer megoldása konvergálhat a Jacobi módszerrel. Egy szükséges és elégséges feltétel a konvergenciára a spektrál sugár segítségével fejezhető ki.

87. Definíció. Egy $A \in \mathbb{R}^{n\times n}$ mátrixnak a sajátértékeinek a halmazát spektrumnak nevezzük

$$\sigma(A) = \left\{\lambda_i : \lambda_i = \text{sajátérték},\ i = \overline{1,n}\right\}.$$

Spektrálsugárnak nevezzük az A mátrix legnagyobb sajátértékének az abszolút értékét

$$\rho(A) = |\lambda_1|,$$

ahol $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$.

Ha $(\lambda, \boldsymbol{x})$ páros sajátértéke, sajátvektora az A mátrixnak, akkor az $\left(\lambda^k, \boldsymbol{x}\right)$ páros sajátértéke, sajátvektora az A^k mátrixnak, vagyis

$$\rho\left(A^k\right) = \left(\rho(A)\right)^k,\ k = 1, 2, ...$$

88. TÉTEL. (*i*) *Bármilyen* $\|\cdot\|$ *szubmultiplikatív mátrix norma esetében igaz az alábbi egyenlőtlenség:*

$$\rho(A) \leq \|A\|.$$

(*ii*) *Bármilyen* $\epsilon > 0$*-ra létezik egy* $\|\cdot\|$ *szubmultiplikatív mátrix norma amelyre*

$$\rho(A) \leq \|A\| \leq \rho(A) + \epsilon.$$

BIZONYÍTÁS. (*i*) Ha $(\lambda, \boldsymbol{x})$ páros sajátérték, sajátvektora az A mátrixnak, akkor

$$|\lambda| \cdot \|\mathbf{x}\| = \|\lambda \cdot \boldsymbol{x}\| = \|A \cdot \boldsymbol{x}\| \leq \|A\| \cdot \|\boldsymbol{x}\|,$$

és mivel $\boldsymbol{x} \neq 0$ következik, hogy $|\lambda| \leq \|A\|$. □

89. TÉTEL. *A (4.1.36) sorozat konvergál (bármilyen* $X^{(0)}$ *kezdeti értékre), akkor és csak akkor ha*

$$\rho(B) < 1. \tag{4.1.39}$$

BIZONYÍTÁS. Feltételezzük, hogy $\rho(B) < 1$. Akkor létezik $\epsilon > 0$ ú.h. $\rho(B) + \epsilon < 1$. Az előbbi tétel szerint létezik egy norma amelyre $\|B\| \leq \rho(B) + \epsilon < 1$, tehát Φ kontrakció és a sorozat konvergens. Feltételezzük, hogy

$$X^{(k+1)} = \Phi\left(X^{(k)}\right) = BX^{(k)} + C,$$

sorozat konvergál X^*-hez bármilyen kezdeti megoldásra. Akkor legyen $X^{(0)}$ ú.h. $X^{(0)} - X^*$ sajátvektora B-nek. Következik, hogy

$$\begin{aligned} X^{(k+1)} - X^* = \Phi\left(X^{(k)}\right) - \Phi\left(X^*\right) = B\left(X^{(k)} - X^*\right) = \ldots \\ = B^{k+1}\left(X^{(0)} - X^*\right) = \lambda^{k+1}\left(X^{(0)} - X^*\right). \end{aligned}$$

A baloldal $k \to \infty$ esetén zéróhoz közelít, ugyanez igaz a jobboldalra, tehát

$$\lambda^{k+1} \to 0,$$

ami akkor igaz ha $|\lambda| < 1$. Mivel tetszőleges λ-ra igaz, következik, hogy

$$\rho(B) < 1.$$

□

90. PÉLDA. $\left\{\begin{array}{rcl} x_1-2x_2 & = & -1 \\ -9x_1+32x_2 & = & 23 \end{array}\right.$, $AX=b \Leftrightarrow X=BX+C$ ahol $B=\begin{pmatrix} 0 & 2 \\ \frac{9}{32} & 0 \end{pmatrix}$, $C=\begin{pmatrix} -1 \\ \frac{23}{32} \end{pmatrix}$. Az A mátrix nem átlósan dominános és a B mátrix normái mind nagyobbak 1-nél, tehát a konvergenciára vonatkozó elégséges feltétel nem használható. A B mátrix sajátértékei $\det(B-\lambda I)=0 \Leftrightarrow \begin{vmatrix} -\lambda & 2 \\ \frac{9}{32} & -\lambda \end{vmatrix}=0 \Leftrightarrow \lambda^2-\frac{9}{16}=0$, ahonnan $|\lambda_{1,2}|=\frac{3}{4}<1$, tehát $\rho(B)<1$, a Jacobi módszer konvergens. Ha $X^{(0)}=\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ egy tetszőleges kezdeti megoldás, akkor $X^{(1)}=BX^{(0)}+C=\begin{pmatrix} -1 \\ \frac{23}{32} \end{pmatrix}$, $\left\|X^{(1)}-X^{(0)}\right\|_2=1.23$, $X^{(2)}=BX^{(1)}+C=\begin{pmatrix} \frac{7}{16} \\ \frac{7}{16} \end{pmatrix}$, $\left\|X^{(2)}-X^{(1)}\right\|_2=1.46...$, $X^{(k)}\xrightarrow[k\to\infty]{}X^*=\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Az alábbi ábrán az első 20 iteráció hibaváltozása látható. ÁBRA 'Lejacobihiba.eps'

4.1.2.2. Gauss-Seidel és SOR módszerek A Jacobi módszer (4.1.38) képlet szerint a $(k+1)$-ik iterációban a megoldás csak az előbbi, (k)-ik komponensektől függnek, viszont a $x_i^{(k+1)}$ komponens kiszámításánál a $x_1^{(k+1)},...,x_{i-1}^{(k+1)}$ értékek már ismertek és jobb közelítést adnak a pontos megoldásnak. Tehát a

$$x_i^{(k+1)}=\frac{1}{a_{ii}}\Big(b_i-a_{i1}x_1^{(k+1)}-a_{i2}x_2^{(k+1)}\ldots- \\ -a_{i,i-1}x_{i-1}^{(k+1)} \;/\; -a_{i,i+1}x_{i+1}^{(k)}\ldots-a_{in}x_n^{(k)}\Big),\ i=\overline{1,n},$$

vagy

$$x_i^{(k+1)}=\frac{1}{a_{ii}}\left(b_i-\sum_{j<i}a_{ij}x_j^{(k+1)}-\sum_{j>i}a_{ij}x_j^{(k)}\right),\ i=\overline{1,n}, \tag{4.1.40}$$

iteratív eljárás gyorsabb mint a (4.1.38). A (4.1.40) algoritmus Gauss-Seidel néven ismert.

A Jacobi és a Gauss-Seidel módszerek mátrix alakját levezethetjük ha az A mátrixot felbontjuk az alábbi módon:

$$A = L + D + U, \tag{4.1.41}$$

ahol L, illetve U szigorúan alsó, illetve felső mátrixok

$$L = \begin{pmatrix} 0 & \cdots & 0 & 0 \\ a_{21} & 0 & & 0 \\ \vdots & & \ddots & \\ a_{n1} & \dots & a_{n,n-1} & 0 \end{pmatrix}, U = \begin{pmatrix} 0 & a_{12} & \dots & a_{1n} \\ 0 & 0 & & a_{2n} \\ \vdots & & \ddots & \\ 0 & \dots & 0 & 0 \end{pmatrix}$$

és D átlós mátrix

$$D = \begin{pmatrix} a_{11} & \cdots & 0 & 0 \\ 0 & a_{22} & & 0 \\ \vdots & & \ddots & \\ 0 & \dots & 0 & a_{nn} \end{pmatrix}.$$

Az $Ax = b$ egyenlet átírható

$$(L + D + U)\, x = b,$$

alakba vagyis

$$Dx = b - (L + U)\, x,$$

ahonnan

$$x = D^{-1}\left(b - (L + U)\, x\right),$$

és az iteratív eljárás:

$$x^{(k+1)} = D^{-1}\left(b - (L + U)\, x^{(k)}\right). \tag{4.1.42}$$

A D egyszerű alakjának köszönhetően a D^{-1} mátrixot könnyű kiszámítani:

$$D^{-1} = \begin{pmatrix} \frac{1}{a_{11}} & \cdots & 0 & 0 \\ 0 & \frac{1}{a_{22}} & & 0 \\ \vdots & & \ddots & \\ 0 & \dots & 0 & \frac{1}{a_{nn}} \end{pmatrix}.$$

A (4.1.42) képlet azonos a Jacobi (4.1.38) algoritmussal.

A Gauss-Seidel módszerhez a csoportosítást a következőképpen végezzük el:

$$Ax = b \Leftrightarrow ((L + D) + U)\, x = b \Leftrightarrow (L + D)\, x = b - Ux,$$

$$x = (L + D)^{-1} (b - Ux)\,,$$

és az iteratív eljárás

$$x^{(k+1)} = (L + D)^{-1} \left(b - Ux^{(k)}\right), \tag{4.1.43}$$

azonos a (4.1.40) eljárással.

SOR (Succesive over relaxation) módszer: Ha $\omega \in (0, 2)$, akkor

$$\omega Ax = \omega b \Leftrightarrow \omega\, (L + D + U)\, x = \omega b \Leftrightarrow$$

$$(\omega L + D + (1 - \omega)\, D + \omega U)\, x = \omega b,$$

$$x = (\omega L + D)^{-1} (\omega b - (\omega U + (1 - \omega)\, D)\, x)\,,$$

ahonnan az iteratív eljárás:

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

vagyis komponensekre lebontva:

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

4.1.3. Hibabecslés, kondicionálás Egy

$$Ax = b,$$

lineáris egyenletrendszer megoldásának a pontosságát az

$$r = Ax - b, \tag{4.1.44}$$

maradék vektorral lehet megbecsülni. Pontos megoldás esetén az r nullvektor aminek hossza nulla, ellenkező esetben a vektor hossza szigorúan pozitív.

Az $Ax = b$ lineáris egyenletrendszer, úgy A mint a b, tagjai megfigyelések vagy számítások eredménye, tehát mindkét esetben az eredeti értékeknek csak a közelítése áll rendelkezésünkre. Felmerül a kérdés mennyire befolyásolja az x megoldást az A, illetve b közelítő ismerete.

Feltételezzük, hogy a b vektor $\triangle b$ perturbációja, az x megoldás egy $\triangle x$ változását eredményezi. Ekkor:

$$A \cdot (x + \triangle x) = b + \triangle b, \Rightarrow A \cdot x + A \cdot \triangle x = b + \triangle b,$$

majd egyszerűsítés ($Ax = b$) után

$$A \cdot \triangle x = \triangle b,$$

vagyis

$$\triangle x = A^{-1} \cdot \triangle b.$$

Tehát

$$\|\triangle x\| = \left\|A^{-1} \cdot \triangle b\right\| \leq \left\|A^{-1}\right\| \cdot \|\triangle b\| = \frac{\|A\| \cdot \|A^{-1}\| \cdot \|\triangle b\|}{\|A\|}, \tag{4.1.45}$$

ahonnan a megoldás relatív eltérése

$$\frac{\|\triangle x\|}{\|x\|} = \|A\| \cdot \left\|A^{-1}\right\| \frac{\|\triangle b\|}{\|A\| \cdot \|x\|}.$$

Felhasználva a

$$\|b\| = \|Ax\| \leq \|A\| \cdot \|x\|,$$

egyenlőtlenséget következik, hogy

$$\frac{\|\triangle x\|}{\|x\|} \leq \|A\| \cdot \left\|A^{-1}\right\| \frac{\|\triangle b\|}{\|b\|}, \tag{4.1.46}$$

vagyis, az x relatív hibáját a b vektor relatív hiba többszöröse majorálja.

91. Definíció. A

$$\mathcal{K}(A) = \|A\| \cdot \left\|A^{-1}\right\|, \tag{4.1.47}$$

számot az A mátrix kondíciószámának nevezzük.

Hasonló a helyzet ha az A mátrix helyett egy perturbált $A+\triangle A$ mátrixot használunk:

$$\begin{aligned}
(A+\triangle A)(x+\triangle x) &= b, \Rightarrow Ax + A\cdot\triangle x + \triangle A(x+\triangle x) = b \Rightarrow \\
A\cdot\triangle x &= -\triangle A(x+\triangle x) \Longrightarrow \triangle x = -A^{-1}\cdot\triangle A(x+\triangle x) \Rightarrow \\
\|\triangle x\| &= \left\|A^{-1}\triangle A(x+\triangle x)\right\| \leq \left\|A^{-1}\right\|\cdot\|\triangle A\|\cdot\|x+\triangle x\| \Rightarrow \\
\frac{\|\triangle x\|}{\|x+\triangle x\|} &\leq \left\|A^{-1}\right\|\cdot\|\triangle A\| = \left\|A^{-1}\right\|\cdot\|A\|\frac{\|\triangle A\|}{\|A\|},
\end{aligned}$$

ahonnan

$$\frac{\|\triangle x\|}{\|x+\triangle x\|} \leq \mathcal{K}(A)\frac{\|\triangle A\|}{\|A\|}. \tag{4.1.48}$$

Habár a (4.1.48) képletben csak az x relatív hibának a közelítése szerepel, a kondíciószám jelentősége jól szemléltethető.

Tehát a lineáris egyenletrendszer megoldásának relatív hibáját majorálja a kondíciószám szorzata A vagy b relatív hibájával amelyek rendszerint kicsik. Ha $\mathcal{K}(A)$ is kis érték akkor a szorzat is kicsi lesz. Ebben az esetben stabil vagy jól kondicionált rendszerekről beszélünk. Viszont, ha $\mathcal{K}(A)$ nagy, akkor a szorzat is az (lehet), ami nagy relatív hibához vezethet.

A kondíciószám legkisebb értéke 1:

$$1 = \|I\| = \left\|A\cdot A^{-1}\right\| \leq \|A\|\cdot\left\|A^{-1}\right\| = \mathcal{K}(A).$$

Habár a kondíciószám függ a normától, ennek nagyságrendje a fontos ezért mindegy milyen normával dolgozunk.

92. Példa. $H = hilb(3) = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix} = \left(\frac{1}{i+j-1}\right)_{i,j}$, $\text{cond}_2(\text{H})\approx 524$.

93. Példa. $V = vander([1\,2\,3]) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 4 & 9 \end{pmatrix}$, $cond_2(V) \approx 70$.

1. Golub, van Loan- pg.107

Az $Ax = b$ lineáris egyenletrendszerek pontosságában szerepet játszik úgy az A mátrix, mint a b szabad tag.

PÉLDA. $A = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-6} \end{pmatrix}$ mátrix kondíciószáma $\mathcal{K}(A) = 10^6$, míg a determinánsa $\det(A) = 10^{-6}$.

4.1.4. Túlhatározott lineáris egyenletrendszerek A gyakorlatban előfordul, hogy olyan lineáris egyenletrendszereket kell megoldani aminek több (vagy kevesebb) egyenlete van mint ismeretlenje. Például ha egy bizonyos n számú paraméter meghatározására n-nél több kísérletet végzünk a keletkező egyenletrendszer nem lesz négyzetes. Gyakran válik ilyen esetben az egyenletrendszer összeférhetetlené, annak ellenére hogy a paraméterek léteznek. Ennek egyik oka lehet például a kísérleteknek hibás végzése, leolvasása.

94. DEFINÍCIÓ. Egy

$$Ax = b, \tag{4.1.49}$$

$$A = (a_{ij})_{\substack{i=1,m \\ j=1,n}}\ ,\ x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},\ b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}, \tag{4.1.50}$$

egyenletrendszerről azt mondjuk hogy túlhatározott ha $n > m$.

95. PÉLDA. Az $\begin{cases} x_1 + x_2 = 2 \\ x_1 + 2x_2 = 3 \\ 2x_1 + x_2 = 4 \end{cases}$ egyenletrendszer túlhatározott és összeférhetetlen.

A feladat legjobb megoldásának kiszámítása érdekében vezessük be a következő $r \in \mathbb{R}^m$ (maradék) vektort:

$$r = Ax - b. \tag{4.1.51}$$

Ekkor a (4.1.49) egyenletrendszernek akkor van (klasszikus értelemben vett) megoldása ha r a nullvektor $r = \theta_{\mathbb{R}^m}$. Ha $r \neq \theta_{\mathbb{R}^m}$, akkor az adott egyenletrendszer összeférhetetlen és ebben az esetben keressük a megoldást a legkisebb négyzetek módszere szerint vagyis, az x vektort

úgy határozzuk meg hogy az r vektor normája minimális legyen:

$$\|r\| = \|Ax - b\| \longrightarrow \min . \tag{4.1.52}$$

Természetesen, ha $\min\|r\| = 0$, akkor visszakapjuk a klasszikus értelemben vett megoldást.

Ha $\|\cdot\|_2$ típusú normát használunk, akkor a (4.1.52) kikötés ekvivalens a következővel:

$$(\|Ax-b\|_2)^2 = \left(\left\|\sum_{j=1}^{n} a_{ij}x_j - b_i\right\|_2\right)^2 = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} a_{ij}x_j - b_i\right)^2 \longrightarrow \min .$$

96. Definíció. Azt (4.1.52) kikötésnek eleget tevő x vektort az egyenletrendszer legkisebb négyzetek szerinti megoldásának nevezzük.

Jelöljük f-el a következő függvényt $f : \mathbb{R}^n \to \mathbb{R}$

$$f(x) = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} a_{ij}x_j - b_i\right)^2 .$$

Ahhoz, hogy az f-nek a minimumát meghatározzuk kiszámítjuk a stacionárius pontjait:

$$\frac{\partial f}{\partial x_k} = \sum_{i=1}^{m} 2\left(\sum_{j=1}^{n} a_{ij}x_j - b_i\right) a_{ik} = 0,\ \ k = 1, ..., n$$

$\Leftrightarrow$

$$\sum_{i=1}^{m} a_{ik}\left(\sum_{j=1}^{n} a_{ij}x_j - b_i\right) = 0,\ \ k = 1, ..., n$$

ami átírva mátrix alakra:

$$A^t(Ax - b) = \theta_{\mathbb{R}^m},$$

vagy:

$$A^tA \cdot x = A^tb. \tag{4.1.53}$$

Tehát, az (4.1.49) egyenletrendszer legkisebb négyzetek szerinti megoldása megegyezik a (4.1.53) egyenletrendszer klasszikus értelemben vett megoldásával.

97. TÉTEL. *Az (4.1.53) lineáris egyenletrendszer megoldására lesz az*

$$\|r\|_2 = \|Ax - b\|_2 ,$$

kifejezés minimális.

BIZONYÍTÁS. Jelöljük az x, illetve y vektoroknak megfelelő maradékvektorokat r_x, r_y-el:

$$r_x = Ax - b$$

$$r_y = Ay - b$$

és igazoljuk, hogy

$$\|r_y\| \geq \|r_x\| .$$

Az értelmezésből következik, hogy

$$\begin{aligned} r_y &= Ay - b = Ay - Ax + Ax - b = A(y - x) + r_x = r_x + A(y - x) , \\ r_y^t &= r_x^t + (y - x)^t A^t , \end{aligned}$$

$$\begin{aligned} (4.1.54) \quad r_y^t r_y &= \left(r_x^t + (y - x)^t A^t\right)(r_x + A(y - x)) = \\ &= r_x^t r_x + (y - x)^t A^t r_x + r_x^t A(y - x) + (y - x)^t A^t A(y - x) = \\ &= r_x^t r_x + (y - x)^t A^t A(y - x) = r_x^t r_x + (A(y - x))^t A(y - x) , \end{aligned}$$

mert

$$A^t r_x = A^t (Ax - b) = \theta ,$$

és

$$r_x^t A = \left(A^t r_x\right)^t = \theta .$$

Figyelembe véve, hogy bármilyen $\mathbf{v}$ vektorra igaz az alábbi állítás:

$$\mathbf{v}^t \cdot \mathbf{v} = \|\mathbf{v}\|_2^2 ,$$

a (4.1.54)-ból következik hogy:

$$\|r_y\|_2^2 = \|r_x\|_2^2 + \|A(y - x)\|_2^2 \geq \|r_x\|_2^2 ,$$

vagyis $\|r_y\| \geq \|r_x\|$. □

98. Példa. Az előbbi példát felhasználva, keressük az (4.1.53) egyenletrendszer megoldását:

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

$$\Leftrightarrow$$

$$\begin{pmatrix} 6 & 5 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 13 \\ 12 \end{pmatrix}$$

$$\Leftrightarrow$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{18}{11} \\ \frac{7}{11} \end{pmatrix}$$

Tehát $x_1 = 18/11,\ x_2 = 7/11$ az egyenletrendszer legkisebb négyzetek módszerének megfelelő megoldása.

4.2. Nemlineáris egyenletrendszerek.

4.2.1. A Newton-Raphson-féle módszer A Newton-Raphson módszer egy iteratív eljárás a nemlineáris egyenletrendszerek megoldására, vagyis egy X^0 kezdeti közelítő megoldást felhasználva egy olyan $X^1, X^2, ..., X^k$ vektor sorozatot hozunk létre ami konvergál az egyenletrendszer megoldásához.

Tekintsük az alábbi egyenletrendszert:

$$\begin{cases} f(x,y) = 0 \\ g(x,y) = 0 \end{cases}, \tag{4.2.1}$$

ahol $f, g : D \subset \mathbb{R}^2 \to \mathbb{R}$, deriválható függvények. Legyen $X^0 = (x_0, y_0)$ az egyenletrendszer egy közelítő megoldása.

Az f, g függvényeket Taylor sorba fejtve (a lineáris tagokkal bezárólag)

$$\begin{aligned} f(x,y) &= f(x_0,y_0) + (x-x_0)\frac{\partial f}{\partial x}(x_0,y_0) + (y-y_0)\frac{\partial f}{\partial y}(x_0,y_0) \\ g(x,y) &= g(x_0,y_0) + (x-x_0)\frac{\partial g}{\partial x}(x_0,y_0) + (y-y_0)\frac{\partial g}{\partial y}(x_0,y_0), \end{aligned}$$

és figyelembe véve a (4.2.1), az egyenletrendszer megoldása a következő összefüggéshez vezet:

$$x - x_0 = -\frac{\left|\begin{array}{cc} f & \frac{\partial f}{\partial y} \\ g & \frac{\partial g}{\partial y} \end{array}\right|_{(x_0,y_0)}}{\det J(x_0,y_0)}, \qquad y - y_0 = -\frac{\left|\begin{array}{cc} \frac{\partial f}{\partial x} & f \\ \frac{\partial g}{\partial x} & g \end{array}\right|_{(x_0,y_0)}}{\det J(x_0,y_0)}, \tag{4.2.2}$$

ahol J a Jacobi mátrix és

$$\det J(x_0,y_0) = \left(\frac{\partial f}{\partial x}\cdot\frac{\partial g}{\partial y} - \frac{\partial g}{\partial x}\cdot\frac{\partial f}{\partial y}\right)(x_0,y_0) = \left|\begin{array}{cc} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{array}\right|_{(x_0,y_0)} \neq 0.$$

A (4.2.1) egyenletrendszer megoldásának következő $X^1 = (x_1, y_1)$ közelítése:

$$x_1 = x_0 - \frac{\left|\begin{array}{cc} f & \frac{\partial f}{\partial y} \\ g & \frac{\partial g}{\partial y} \end{array}\right|_{(x_0,y_0)}}{\det J(x_0,y_0)}, \qquad y_1 = y_0 - \frac{\left|\begin{array}{cc} \frac{\partial f}{\partial x} & f \\ \frac{\partial g}{\partial x} & g \end{array}\right|_{(x_0,y_0)}}{\det J(x_0,y_0)}. \tag{4.2.3}$$

Hasonlóan a k-ik lépésben ha $\det J(x_k,y_k) \neq 0$ akkor:

$$x_{k+1} = x_k - \frac{\left|\begin{array}{cc} f & \frac{\partial f}{\partial y} \\ g & \frac{\partial g}{\partial y} \end{array}\right|_{(x_k,y_k)}}{\det J(x_k,y_k)}, \qquad y_{k+1} = y_k - \frac{\left|\begin{array}{cc} \frac{\partial f}{\partial x} & f \\ \frac{\partial g}{\partial x} & g \end{array}\right|_{(x_k,y_k)}}{\det J(x_k,y_k)}. \tag{4.2.4}$$

A (4.2.1) feladat általánosítható

$$F(X) = \mathbf{0}, \tag{4.2.5}$$

alakba, ahol

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n, \ \mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

és

$$F : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}, \ F = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}$$

egy vektorváltozós, vektor függvény ($f_i : D_i \subset \mathbb{R} \to \mathbb{R}$).

Az

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

egyváltozós Newton módszer általánosítható

$$X^{k+1} = X^k - \left(J_F\left(X^k\right)\right)^{-1} \cdot F\left(X^k\right), \tag{4.2.6}$$

alakra, ahol

$$J_F(X) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(X) & \dots & \frac{\partial f_1}{\partial x_n}(X) \\ \frac{\partial f_2}{\partial x_1}(X) & & \frac{\partial f_2}{\partial x_n}(X) \\ \vdots & & \\ \frac{\partial f_n}{\partial x_1}(X) & & \frac{\partial f_n}{\partial x_n}(X) \end{pmatrix} \tag{4.2.7}$$

az F függvény Jacobi mátrixa.

Az eljárást folytatjuk míg két X^k, illetve X^{k+1} egymásutáni vektor különbségének a normája kisebb mint egy adott $\epsilon > 0$ pontosság:

$$\left\|X^{k+1} - X^k\right\| \leq \epsilon, \tag{4.2.8}$$

$$\frac{\left\|X^{k+1} - X^k\right\|}{\|X^k\|} \leq \epsilon, \tag{4.2.9}$$

vagy, figyelembe véve, hogy határértékben $F\left(X^k\right) \underset{k\to\infty}{\longrightarrow} \mathbf{0}$, az iterációkból kiléphetünk ha:

$$\left\|F\left(X^k\right)\right\| \leq \epsilon. \tag{4.2.10}$$

99. Példa. A Newton-Raphson módszert használva oldjuk meg a következő egyenletrendszert:

$$\begin{cases} f(x,y) = x^3 + y^3 - 6x + 3 = 0 \\ g(x,y) = x^3 - y^3 - 6y + 2 = 0 \end{cases}. \tag{4.2.11}$$

Megoldás. Legyen $X^0 = (x_0, y_0) = (0,0)$ egy kezdeti megoldás.$\Rightarrow$ $f(0,0) = 3$, $g(0,0) = 2$,

$$J(x_0, y_0) = \begin{pmatrix} 3x^2 - 6 & 3y^2 \\ 3x^2 & -3y^2 - 6 \end{pmatrix}(0,0) = \begin{pmatrix} -6 & 0 \\ 0 & -6 \end{pmatrix}$$

$\Longrightarrow \det J(0,0) = 36$. Akkor a (4.2.3) megfelelően kapjuk a következő megközelítést:

$$\begin{aligned} x_1 &= 0 - \frac{\begin{vmatrix} 3 & 0 \\ 2 & -6 \end{vmatrix}}{36} = \frac{18}{36} = 0.5,\ y_1 = 0 - \frac{\begin{vmatrix} -6 & 3 \\ 0 & 2 \end{vmatrix}}{36} = \frac{12}{36} = 0.33 \\ &\Rightarrow X^1 = (0.5, 0.33), \left\| X^1 - X^0 \right\|_2 = 0.6, ... \end{aligned}$$

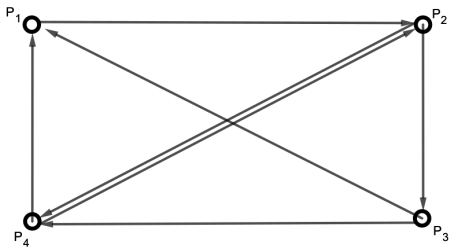
A (4.2.11) egyenletrendszernek a mértani jelentése az alábbi ábrán látható:

Ha az $X^0 = (0,0)$ egy kezdeti megoldás, akkor az $\left(X^k\right)_k$ megoldás sorozat a $(0.53, 0.35)$ pont felé konvergál.

5. FEJEZET

Sajátérték, sajátvektor numerikus kiszámítása

Tekintsük az alábbi irányított gráfot ami például internet linkek közötti forgalmat modellezheti:



5.0.1. ábra. Gráf

A gráf szomszédsági mátrixa:

$$szomszédsági\ mátrix = csomópontba\ \mathbf{be} \rightarrow \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \end{array} \begin{array}{c} \begin{array}{cccc} csomópontból\ \mathbf{ki} \downarrow & & & \\ P_1 & P_2 & P_3 & P_4 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

és a megfelelő valószínűségi (sztochasztikus) mátrix:

$$A = \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

Szeretnénk megtudni mennyire látogatottak a linkek, vagyis egy tetszőleges séta alkalmával milyen gyakran vagyunk egyik-vagy másik csomópontban?

Válasz: $\mathbf{v} \simeq \begin{pmatrix} 0.21 \\ 0.34 \\ 0.17 \\ 0.26 \end{pmatrix}$ domináns sajátvektor.

Legyen $A \in \mathcal{M}_n(\mathbb{R})$ egy négyzetes mátrix.

100. Definíció. Egy $X \in \mathbb{C}^n,\ X \neq 0$ vektort az A mátrix sajátvektorának nevezzük ha

$$AX = \lambda X. \tag{5.0.1}$$

A λ skalárt az A mátrix, X vektornak megfelelő sajátértéknek nevezzük.

101. Példa. Ha $A = \begin{pmatrix} 3 & 1 \\ 2 & 2 \end{pmatrix}$, akkor $A \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 4 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, tehát $X = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ a mátrix sajátvektora és $\lambda = 4$ a megfelelő sajátértéke.

A (5.0.1) képletből következik, hogy

$$AX - \lambda X = 0$$

$$(A - \lambda I_n) X = 0. \tag{5.0.2}$$

A (5.0.2) homogén egyenletrendszernek akkor van nullától különböző megoldása ha

$$\det(A - \lambda I_n) = 0. \tag{5.0.3}$$

102. Definíció. A $\det(A - \lambda I_n)$ karakterisztikus determinánsnak, az (5.0.3) egyenletet pedig karakterisztikus egyenletnek nevezzük.

A (5.0.3) karakterisztikus determináns egy n-ed fokú polinom λ-ba

$$P_A(\lambda) := \det(A - \lambda I_n) = (-1)^n \left(\lambda^n - S_1 \lambda^{n-1} + S_2 \lambda^{n-2} - ... + (-1)^n S_n\right)$$

ahol S_i a főátlón elhelyezkedő $i \times i$ minorok összege: $S_1 = Tr(A), ..., S_n = \det(A)$.

A Viéte összefüggésekből következik, hogy

$$\prod_{i=1}^{n} \lambda_i = \lambda_1 ... \lambda_n = \det(A),$$

vagyis ha A szinguláris $\det(A) = 0$, akkor (legalább) egy sajátérték zéró: $\exists k \in \overline{1,n}$, ú.h. $\lambda_k = 0$.

103. TÉTEL. *Az A mátrix kielégíti a saját karakterisztikus egyenletét*

$$A^n - S_1 A^{n-1} + S_2 A^{n-2} - ... + (-1)^n S_n I_n = 0. \tag{5.0.4}$$

5.1. Krylov módszer

104. DEFINÍCIÓ. Az $A \in M_n$ és $y = (y_1, y_2, ..., y_n)^T \in \mathbb{R}^n$ oszlopmátrix által generált

$$\mathcal{K}_n(A, y) = span\left\{y, Ay, ..., A^{n-1}y\right\} \tag{5.1.1}$$

Krylov vektorrendszernek nevezzük.

A karakterisztikus polinom S_i, $i = \overline{1,n}$ együtthatói kiszámítása érdekében átírjuk a (5.0.4) egyenletet

$$A^n + S_1 A^{n-1} + S_2 A^{n-2} + ... + S_n I_n = 0, \tag{5.1.2}$$

alakba, majd jobboldalról beszorozzuk egy $y^{(0)} \in \mathbb{R}^n$ vektorral $\Rightarrow$

$$A^n y^{(0)} + S_1 A^{n-1} y^{(0)} + ... + S_{n-1} A y^{(0)} + S_n I_n y^{(0)} = 0. \tag{5.1.3}$$

Jelöljük

(5.1.4)

$$y^{(1)} := Ay^{(0)},\ y^{(2)} := A^2 y^{(0)} = Ay^{(1)}, ...,\ y^{(n)} := A^n y^{(0)} = Ay^{(n-1)},$$

amit behelyettesítve a (5.1.3) a következő lineáris egyenletrendszert eredményezi:

$$y^{(n)} + S_1 y^{(n-1)} + ... + S_{n-1} y^{(1)} + S_n y^{(0)} = 0,$$

vagyis

$$S_1 y^{(n-1)} + ... + S_{n-1} y^{(1)} + S_n y^{(0)} = -y^{(n)}. \tag{5.1.5}$$

Ha a (5.1.1) vektorrendszer lineárisan független, akkor a (5.1.5) négyzetes lineáris egyenletrendszer összeférhető, és a megoldását behelyettesítjük a

$$\lambda^n + S_1\lambda^{n-1} + S_2\lambda^{n-2} - ... + S_n = 0$$

karakterisztikus egyenletbe, majd megoldjuk.

105. Példa. Az $A = \begin{pmatrix} 3 & 1 \\ 2 & 2 \end{pmatrix}$ mátrix karakterisztikus egyenletét felírjuk $\lambda^2 + S_1\lambda + S_2 = 0$ alakba, ahol S_1, S_2 együtthatók az $S_1y^{(1)} + S_2y^{(0)} = -y^{(2)}$ egyenletrendszer megoldásai. Legyen $y^{(0)} \in \mathbb{R}^2$ egy tetszőleges oszlopvektor, például $y^{(0)} = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$. A (5.1.4)-ből következik, hogy $y^{(1)} = \begin{pmatrix} 7 \\ 2 \end{pmatrix}, y^{(2)} = \begin{pmatrix} 23 \\ 18 \end{pmatrix}$, tehát a (5.1.5) lineáris egyenletrendszer $\begin{cases} 7S_1 + 3S_2 = -23 \\ 2S_1 - 2S_2 = -18 \end{cases}$, melynek megoldása $S_1 = -5, S_2 = 4$. Innen a karakterisztikus egyenlet $\lambda^2 - 5\lambda + 4 = 0$ melynek megoldása $\lambda_1 = 4, \lambda_2 = 1$.

Ha $y^{(0)} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ akkor $y^{(1)} = y^{(2)} = y^{(0)} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ tehát az (5.1.1) vektorrendszer lineárisan összefüggő és az egyenletrendszernek nincs megoldása. Ebben az esetben más $y^{(0)}$ választunk.

5.2. Hatvány módszer

Feltételezzük, hogy az $A \in M_n(\mathbb{R})$ mátrix $\lambda_i,\ i = \overline{1,n}$ sajátértékei valósak és különbözőek:

$$|\lambda_1| > |\lambda_2| > ... > |\lambda_n|.$$

A hatvány módszer segítségével meghatározható a legnagyobb (domináns) sajátérték λ_1, illetve a megfelelő (domináns) sajátvektor X_1.

Mivel $\lambda_i \neq \lambda_j$ következik, hogy $X_i,\ i = 1, \ldots, n$ sajátvektorok lineárisan függetlenek, tehát az $\mathbb{R}^n$ térben egy bázist alkotnak. Ebben a

bázisban bármilyen $\forall Y \in \mathbb{R}^n$ vektor egyértelműen előállítható a bázisvektorok segítségével

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n.$$

Innen hatványozással következik, hogy:,

$$\begin{aligned}
AY &= \alpha_1 AX_1 + \alpha_2 AX_2 + \ldots + \alpha_n AX_n = \alpha_1\lambda_1 X_1 + \alpha_2\lambda_2 X_2 + \ldots + \alpha_n\lambda_n X_n,\\
A^2Y &= A\left(AY\right) = A\left(\alpha_1\lambda_1 X_1 + \alpha_2\lambda_2 X_2 + \ldots + \alpha_n\lambda_n X_n\right) =\\
&\quad = \alpha_1\lambda_1^2 X_1 + \alpha_2\lambda_2^2 X_2 + \ldots + \alpha_n\lambda_n^2 X_n,\\
&\vdots\\
A^kY &= \alpha_1\lambda_1^k X_1 + \alpha_2\lambda_2^k X_2 + \ldots + \alpha_n\lambda_n^k X_n,
\end{aligned}$$

és mivel $\frac{\lambda_i}{\lambda_1} \in (-1,1)\,, i > 1$ következik, hogy

$$A^kY = \lambda_1^k\left(\alpha_1 X_1 + \alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^k X_2 + \ldots + \alpha_n\left(\frac{\lambda_n}{\lambda_1}\right)^k X_n\right) \underset{k\to\infty}{\longrightarrow} \lambda_1^k\alpha_1 X_1.$$

Tehát, nagy k értékre

$$A^kY \approx \lambda_1^k\alpha_1 X_1$$

és hasonlóan

$$A^{k+1}Y \approx \lambda_1^{k+1}\alpha_1 X_1 = \lambda_1\left(\lambda_1^k\alpha_1 X_1\right) = \lambda_1\left(A^kY\right),$$

vagyis A^kY és $A^{k+1}Y$ vektorok komponensei arányosak, az arány pedig λ_1 a domináns sajátérték. Tehát

$$A\left(A^kY\right) \approx \lambda_1\left(A^kY\right), \tag{5.2.1}$$

vagyis $\left(A^kY\right)$ a λ_1-hez hozzárendelt (domináns) sajátvektor.

Egy adott sajátvektor esetében a megfelelő sajátértéket a Rayleigh tétellel is kiszámíthatjuk:

106. Tétel. *Ha X az A mátrix sajátvektora, akkor a hozzárendelt sajátérték:*

$$\lambda = \frac{AX \cdot X}{X \cdot X} \tag{5.2.2}$$

ahol $\cdot$ a vektorok skaláris szorzatát jelöli.

BIZONYÍTÁS. Ha X az A mátrix sajátvektora, akkor $AX = \lambda X$ tehát

$$\frac{AX \cdot X}{X \cdot X} = \frac{\lambda X \cdot X}{X \cdot X} = \frac{\lambda (X \cdot X)}{X \cdot X} = \lambda.$$

□

107. PÉLDA. Az $A = \begin{pmatrix} 3 & 1 \\ 2 & 2 \end{pmatrix}$ mátrix esetén (aminek domináns sajátértéke/sajávektora $\left(4, \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)$) legyen $Y = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$. A hatványozás követően:

$$AY = \begin{pmatrix} 5 \\ 2 \end{pmatrix},\ A^2Y = \begin{pmatrix} 17 \\ 14 \end{pmatrix},\ A^3Y = \begin{pmatrix} 65 \\ 62 \end{pmatrix},\ A^4Y = \begin{pmatrix} 257 \\ 254 \end{pmatrix},$$

tehát a domináns sajátvektor közelíthető az $X_1 = \begin{pmatrix} 257 \\ 254 \end{pmatrix} = \frac{1}{257} \begin{pmatrix} 1 \\ \frac{254}{257} \end{pmatrix}$ vektorral. A két utolsó vektor komponenseinek az aránya

$$\frac{257}{65} \approx 3.9538,\ \frac{254}{62} \approx 4.0968,$$

és határértékben az arány a pontos értékhez (4 -hez) közelít.

Ha a Rayleigh eljárást használjuk, akkor

$$\frac{AX \cdot X}{X \cdot X} = \frac{\begin{pmatrix} 3 & 1 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ \frac{254}{257} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \frac{254}{257} \end{pmatrix}}{\begin{pmatrix} 1 \\ \frac{254}{257} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \frac{254}{257} \end{pmatrix}} \approx 4.0058$$

Annak érdekében, hogy a vektorok komponensei ne nőjenek túlságosan normalizálhatjuk őket. Ha a vektorok *sup* normáját használjuk,

akkor

$$AY = \begin{pmatrix} 5 \\ 2 \end{pmatrix} = 5 \begin{pmatrix} 1 \\ \frac{2}{5} \end{pmatrix} \to \begin{pmatrix} 1 \\ \frac{2}{5} \end{pmatrix} := Y^{(2)},$$

$$AY^{(2)} = \begin{pmatrix} \frac{17}{5} \\ \frac{14}{5} \end{pmatrix} = \frac{17}{5} \begin{pmatrix} 1 \\ \frac{14}{17} \end{pmatrix} \to \begin{pmatrix} 1 \\ \frac{14}{17} \end{pmatrix} := Y^{(3)}$$

$$AY^{(3)} = \begin{pmatrix} \frac{65}{17} \\ \frac{62}{17} \end{pmatrix} = \frac{65}{17} \begin{pmatrix} 1 \\ \frac{62}{65} \end{pmatrix} \to \begin{pmatrix} 1 \\ \frac{62}{65} \end{pmatrix} := Y^{(4)}$$

$$AY^{(4)} = \begin{pmatrix} \frac{257}{65} \\ \frac{254}{65} \end{pmatrix} = \frac{257}{65} \begin{pmatrix} 1 \\ \frac{254}{257} \end{pmatrix} \to \begin{pmatrix} 1 \\ \frac{254}{257} \end{pmatrix} := Y^{(5)}$$

Az $AY^{(k)}$ vektorok *sup* normájának a sorozata $5, \frac{17}{5}, \frac{65}{17}, \frac{257}{65}, \ldots$ konvergál a domináns sajátértékhez (jelen esetben 4-hez), míg a megfelelő $Y^{(k)}$ a domináns sajátvektorhoz konvergál (jelen esetben $X_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ vagy euklideszi normával $X_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$).

Az eljárás konvergenciáját a $\frac{|\lambda_2|}{|\lambda_1|}$ arány határozza meg: minél közelebb van az egységhez annál lassabban konvergálnak az iterációk.

A többi sajátérték az ún. deflációs eljárással számíthatók ki.

6. FEJEZET

Függvény approximáció, interpoláció

6.1. Analitikus függvények közelítése Taylor sorral

Legyen $f : [a,b] \to R$ egy végtelenszer deriválható függvény, és $x_0 \in (a,b)$.

108. Definíció. Az f függvényt analitikusnak nevezzük ha hatványsorként előállítható:

$$f(x) = f(x_0) + \frac{1}{1!}f'(x_0)(x-x_0) + \frac{1}{2!}f''(x_0)(x-x_0)^2 + ... \tag{6.1.1}$$

Tehát az f függvény felírható mint T_n egy n-ed fokú polinom (Taylor), és a megfelelő R_n maradék összegeként:

$$f(x) = T_n(x) + R_n(x), \tag{6.1.2}$$

ahol

(6.1.3)

$$T_n(x) = f(x_0) + \frac{1}{1!}f'(x_0)(x-x_0) + \frac{1}{2!}f''(x_0)(x-x_0)^2 + ... + \frac{1}{n!}f^{(n)}(x_0)(x-x_0)^n$$

és a maradék tag :

$$R_n(x) = \left| \frac{1}{(n+1)!}f^{(n+1)}(x_0)(x-x_0)^{n+1} + ... \right|. \tag{6.1.4}$$

Numerikusan az f függvényt az első $(n+1)$ taggal, vagyis az n-ed fokú Taylor polinommal közelítjük meg:

$$f(x) \simeq T_n(x) = \sum_{i=0}^{n} \frac{1}{i!}f^{(i)}(x_0)(x-x_0)^i. \tag{6.1.5}$$

109. Tétel. *Az R_n maradék tagot a következő ún. Lagrange -féle alakban lehet kifejezni:*

$$R_n(x) = \left| \frac{1}{(n+1)!}f^{(n+1)}(\theta)(x-x_0)^{n+1} \right|, \ \theta \in (x, x_0). \tag{6.1.6}$$

Az $x_0 = 0$ pont körüli sorfejtés egy sajátos esete a Taylor kifejtésnek és a MacLaurin nevet viseli.

110. PÉLDA. Az e^x függvény sorfejtése a következő:

$$e^x = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + ... + \frac{1}{n!}x^n + ...$$

amit egy n-ed fokú T_n polinommal közelíthetünk:

$$e^x \approx T_n(x) = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + ... + \frac{1}{n!}x^n.$$

A közelítésből eredő hiba:

$$R_n(x) = \left| \frac{e^\theta}{(n+1)!}x^{n+1} \right|,\ \theta \in (0,x)\ \ (\text{vagy } \theta \in (x,0)\,).$$

Innen, ha $x = \frac{1}{2}$ következik, hogy:

$$\sqrt{e} = e^{\frac{1}{2}} = 1 + \frac{1}{1!}\left(\frac{1}{2}\right)^1 + \frac{1}{2!}\left(\frac{1}{2}\right)^2 + ... + \frac{1}{n!}\left(\frac{1}{2}\right)^n$$

és n-et a következő feltételből számítjuk ki:

$$R_n(x) = \left| \frac{e^\theta}{(n+1)!}x^{n+1} \right| < \epsilon,\ \theta \in \left(0, \frac{1}{2}\right),$$

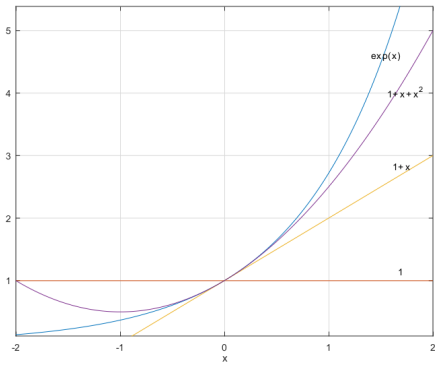
ahol $\epsilon > 0$ egy előre megadott pontosság. Felhasználva, hogy $e \in (2,3)$ a következő becsléshez jutunk:

$$\left| \frac{3^\theta}{(n+1)!}\left(\frac{1}{2}\right)^{n+1} \right| < \frac{3}{(n+1)!}\left(\frac{1}{2}\right)^{n+1} < \epsilon.$$

Konkrétan, ha a pontosság $\epsilon = 10^{-2}$, akkor $n = 3$-ra a fenti egyenlőtlenség teljesül, tehát 4 tag összegzésével garantált a megadott pontosság:

$$\sqrt{e} \simeq 1 + \frac{1}{1!}\left(\frac{1}{2}\right)^1 + \frac{1}{2!}\left(\frac{1}{2}\right)^2 + \frac{1}{3!}\left(\frac{1}{2}\right)^3 = \frac{79}{48}$$

A Taylor polinom csak lokálisan, az x_0 pont környékén alkalmas a közelítésre.



6.1.1. ábra. Taylor közelítés

111. Példa. Számítsuk ki $\sqrt{0.99}$ közelítő értékét!

Az $f(x) = \sqrt{1+x} = (1+x)^{\frac{1}{2}}$ függvényre alkalmazzuk a Maclaurin sorfejtést:

$$f(x) = 1 + \frac{1}{1!}\left(\frac{1}{2}\right)x + \frac{1}{2!}\left(-\frac{1}{4}\right)x^2 + \ldots,$$

tehát $x = -0.01 = -10^{-2}$ -re a következő közelítéseket kapjuk:

$$\sqrt{0.99} \approx 1,$$

vagy

$$\sqrt{0.99} \approx 1 + \frac{1}{2}(-0.01) = 0.995,$$

illetve

$$\sqrt{0.99} \approx 1 - \frac{10^{-2}}{2} - \frac{10^{-4}}{8} = 0.9949875.$$

6.2. Interpoláció

Gyakran történik meg a gyakorlatban hogy bizonyos folyamatokat (csak) diszkrét mérésekkel lehet elemezni. Például különböző

$$x_0 < x_1 < \ldots < x_n$$

időpontokban (nevezzük interpolációs alappontoknak vagy osztópontok) a mérések eredményei

$$y_0, y_1, \ldots, y_n$$

amik nem másak mint a folyamatot leíró f (ismeretlen) függvény értékei az alappontokban:

$$y_0 = f(x_0),\ y_1 = f(x_1),\ ...,\ y_n = f(x_n).$$

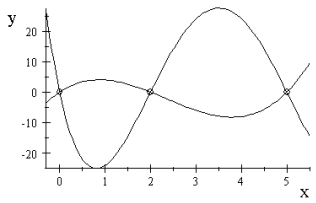
Interpolációnak nevezzük azt az eljárást amellyel egy olyan

$$F : [x_0, x_n] \to \mathbb{R}$$

függvényt értelmezünk amely megegyezik az alappontokban f-el:

$$F(x_i) = f(x_i) = y_i,\ i = 0, ..., n. \tag{6.2.1}$$

Mértanilag ez azt jelenti hogy olyan meghatározott típusú $y = F(x)$ görbét kell keresnünk, amely az adott $(x_0, y_0), ..., (x_n, y_n)$ pontokra illeszkedik.



6.2.1. ábra. Pontokra illeszkedő függvények

Kézenfekvő az interpoláló F függvényt polinom (algebrai vagy trigonometrikus), racionális függvény, stb. alakban keresni.

Extrapoláción azt az eljárást jelenti amellyel az $[x_0, x_n]$ intervallumon kívül próbáljuk az f-et megközelíteni.

6.2.1. Polinomiális interpoláció Feltételezzük hogy az

$$x_0 < x_1 < ... < x_n$$

alappontok esetében ismertek a következő adatok:

$$y_i = f(x_i),\ i = 0, ..., n.$$

Az interpoláló P függvényt (algebrai) polinom alakban keressük:

$$P(x) = a_0 x^n + ... + a_{n-1} x + a_n. \tag{6.2.2}$$

Mivel az a_i ismeretlenek száma $(n+1)$ a polinom fokszámát kisebb vagy egyenlőnek tekintjük mint n, különben megeshet hogy nem létezik a keresett polinom.

112. TÉTEL. *Ha a P interpoláló polinom fokszáma egyenlő n-el, akkor a polinom egyértelműen meghatározott.*

BIZONYÍTÁS. A $P(x_i) = y_i,\ i = 0, ..., n$ interpolációs feltételekből következik, hogy:

$$\text{(6.2.3)} \qquad \begin{cases} a_0 x_0^n + ... + a_{n-1} x_0 + a_n = y_0 \\ a_0 x_1^n + ... + a_{n-1} x_1 + a_n = y_1 \\ \qquad\qquad \vdots \\ a_0 x_n^n + ... + a_{n-1} x_n + a_n = y_n \end{cases},$$

vagy mátrix alakban:

$$\begin{pmatrix} x_0^n & x_0^{n-1} & ... & x_0 & 1 \\ x_1^n & x_1^{n-1} & ... & x_1 & 1 \\ ... & & & & \\ x_n^n & x_n^{n-1} & ... & x_n & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \\ y_n \end{pmatrix} \Leftrightarrow V \cdot a = y.$$

A lineáris egyenletrendszer mátrixa, illetve determinánsa sajátos, ú.n. Vandermonde típusú:

(6.2.4)

$$\det(V) = \begin{vmatrix} x_0^n & x_0^{n-1} & ... & x_0 & 1 \\ x_1^n & x_1^{n-1} & ... & x_1 & 1 \\ ... & & & & \\ x_n^n & x_n^{n-1} & ... & x_n & 1 \end{vmatrix} = (-1)^{n+1} \prod_{0 \le x_i < x_j \le n} (x_j - x_i) \neq 0,$$

és mivel $x_i \neq x_j,\ \forall i \neq j$, a determináns különbözik nullától, tehát az egyenletrendszer egyértelműen meghatározott. □

A bemutatott módszernek két hátránya is van: egyfelől az együtthatókat csak egy másik eljárás eredményeként kapjuk meg, másrészt a keletkezett egyenletrendszer gyakran lesz rosszul kondicionált.

113. PÉLDA. Az

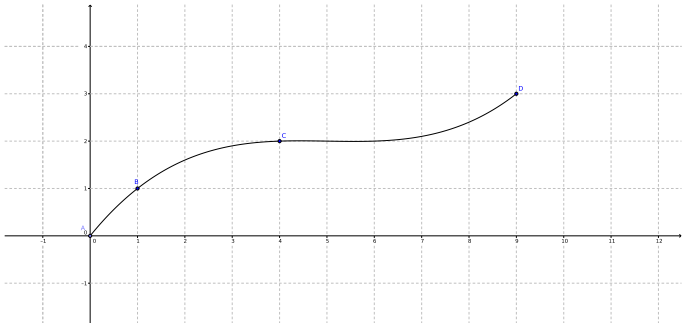
x_i	0	1	4	9
y_i	0	1	2	3

interpoláló pontok esetében a Vandermonde mátrix

$$V = \begin{pmatrix} 0^3 & 0^2 & 0 & 1 \\ 1^3 & 1^2 & 1 & 1 \\ 4^3 & 4^2 & 4 & 1 \\ 9^3 & 9^2 & 9 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 64 & 16 & 4 & 1 \\ 729 & 81 & 9 & 1 \end{pmatrix}.$$

A $V \cdot a = y$ egyenletrendszerből megkapjuk az $a = \begin{pmatrix} 0 \\ \frac{37}{30} \\ -\frac{1}{4} \\ \frac{1}{60} \end{pmatrix}$ együtthatókat, tehát az interpoláló polinom:

$$P(x) = 0 + \frac{37}{30}x - \frac{1}{4}x^2 + \frac{1}{60}x^3,\ x \in [0, 9].$$



6.2.2. ábra. Köbös polinommal interpolált pontok

A V mátrix kondíciószáma $cond(V) = 1.66 \cdot 10^3$.

6.2.1.1. Lagrange-féle interpolációs polinom Jelöljük $L_n f$-el az

$$(x_0, y_0), ..., (x_n, y_n)$$

pontokat összekötő n-ed fokú polinomot amit a következő alakban keresünk:

$$L_n f(x) = y_0 l_0(x) + y_1 l_1(x) + ... + y_n l_n(x). \tag{6.2.5}$$

Az $l_i(x)$, $i=0,...,n$ az úgynevezett n-ed fokú Lagrange -féle alappolinomok. Az

$$L_n f(x_i) = f(x_i) = y_i,\ i=0,...,n, \tag{6.2.6}$$

interpoláló feltételeket felhasználva a következő egyenletrendszerhez jutunk:

$$\begin{cases} y_0 l_0(x_0) + y_1 l_1(x_0) + ... + y_n l_n(x_0) = y_0 \\ y_0 l_0(x_1) + y_1 l_1(x_1) + ... + y_n l_n(x_1) = y_1 \\ \vdots \\ y_0 l_0(x_n) + y_1 l_1(x_n) + ... + y_n l_n(x_n) = y_n \end{cases} \tag{6.2.7}$$

aminek egyetlen megoldása:

$$l_0(x_0) = 1,\ l_1(x_0) = 0, ...,\ l_n(x_0) = 0 \tag{6.2.8}$$

$$l_0(x_1) = 0,\ l_1(x_1) = 1, ...,\ l_n(x_1) = 0 \tag{6.2.9}$$

$$... \tag{6.2.10}$$

$$l_0(x_n) = 0,\ l_1(x_n) = 0, ...,\ l_n(x_n) = 1 \tag{6.2.11}$$

vagy a

$$\delta_{ij} = \begin{cases} 1, \text{ ha } i=j \\ 0, \text{ ha } i \neq j \end{cases}$$

Kronecker szimbólumot felhasználva:

$$l_i(x_j) = \delta_{ij},\ i,j=0,...,n. \tag{6.2.12}$$

A (6.2.8)...(6.2.11)-ból következik hogy minden $l_i(x)$ polinomnak n különböző gyöke van, nevezetesen: $x_0,...,x_{i-1},x_{i+1},...,x_n$, tehát:

$$l_i(x) = c_i(x-x_0)...(x-x_{i-1})(x-x_{i+1})...(x-x_n).$$

A c_i együttható meghatározható a $l_i(x_i)=1$ egyenletből:

$$c_i = \frac{1}{(x_i-x_0)...(x_i-x_{i-1})(x_i-x_{i+1})...(x_i-x_n)},\ i=0,...,n$$

vagyis:

$$c_i = \frac{1}{\prod_{k\neq i}(x_i-x_k)},\ i=0,...,n. \tag{6.2.13}$$

Ennek segítségével felírjuk az alappolinomokat:

$$l_i(x) = \frac{\prod_{k\neq i}(x - x_k)}{\prod_{k\neq i}(x_i - x_k)}, \tag{6.2.14}$$

vagy bevezetve a következő polinomot:

$$\varpi(x) = (x - x_0)\dots(x - x_n),$$

aminek a deriváltja x_i-ben:

$$\varpi'(x_i) = (x_i - x_0)\dots(x_i - x_{i-1})(x_i - x_{i+1})\dots(x_i - x_n), \tag{6.2.15}$$

a (6.2.14) képlet a következő lesz:

$$l_i(x) = \frac{\varpi(x)}{(x - x_i)\varpi'(x_i)}.$$

Tehát a Lagrange interpoláló polinom a következő alakot veszi fel:

$$L_n f(x) = \sum_{i=0}^{n} y_i \cdot l_i(x) = \sum_{i=0}^{n} y_i \cdot \frac{\varpi(x)}{(x - x_i)\varpi'(x_i)}. \tag{6.2.16}$$

114. Tétel. *A (6.2.16) Lagrange interpolációs polinommal való közelítésből eredő hiba a következő:*

$$|f(x) - L_n f(x)| = \left|\frac{f^{(n+1)}(c)}{(n+1)!}\varpi(x)\right|, \ c \in (a,b). \tag{6.2.17}$$

Sajátos esetként megemlíthető az $n = 1$ és $n = 2$. Az első esetben $L_1 f(x)$ polinom nem más mind a $(x_0, y_0), (x_1, y_1)$ pontokat összekötő egyenes egyenlete:

$$L_1 f(x) = y_0\frac{x - x_1}{x_0 - x_1} + y_1\frac{x - x_0}{x_1 - x_0}, \ x \in [x_0, x_1],$$

míg $n = 2$-re az $(x_0, y_0), (x_1, y_1), (x_2, y_2)$ pontokat összekötő parabola egyenlete:

$$L_2 f(x) = y_0\frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + y_1\frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + \\ + y_2\frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}, \quad x \in [x_0, x_2].$$

115. Példa. Határozzuk meg a Lagrange -féle interpoláló polinomot a következő adatok esetében:

x_i	0	1	4	9
y_i	0	1	2	3

MEGOLDÁS. A harmadfokú Lagrange interpoláló polinom

$$L_3 f(x) = y_0 l_0(x) + y_1 l_1(x) + y_2 l_2(x) + y_3 l_3(x),$$

ahol

$$\begin{aligned} l_0(x) &= \frac{(x-1)(x-4)(x-9)}{(0-1)(0-4)(0-9)}, \; l_1(x) = \frac{(x-0)(x-4)(x-9)}{(1-0)(1-4)(1-9)} \\ l_2(x) &= \frac{(x-0)(x-1)(x-9)}{(4-0)(4-1)(4-9)}, \; l_3(x) = \frac{(x-0)(x-1)(x-4)}{(9-0)(9-1)(9-4)} \end{aligned}$$

tehát

$$L_3 f(x) = 0 \cdot l_0(x) + 1 \cdot \frac{x(x-4)(x-9)}{24} + 2 \cdot \frac{x(x-1)(x-9)}{-60} + \\ + 3 \cdot \frac{x(x-1)(x-4)}{360}, \; x \in [0,9].$$

A polinom és az ábrája azonos az előbbi példában bemutatott polinommal, csak a reprezentációja más.

□

116. PÉLDA. Interpoláljuk az adatokat

x_i	0	1	2	3	4
y_i	1	1	2	6	24

Adjunk becslést $P(k)$-ra, ha $k = 3.1$, majd hasonlítsuk össze a $\Gamma(k+1)$ értékkel, ahol

$$\Gamma(s) = \int_0^\infty t^{s-1} e^t \mathrm{dt},$$

(gamma függvény)!

117. PÉLDA. Egy testet 100 m magasból szabadon elengedünk. Ha 1 mp, 2 mp, 3 mp után 95, 80, 55.9, méter magasan van a test mennyi idő után csapódik a talajba?

118. PÉLDA. Interpoláljuk az adatokat

x_i	0	1	2	3	π
$y_i = \sin(x_i)$	0	$\sin(1)$	$\sin(2)$	$\sin(3)$	0

Adjunk becslést $P(k)$-ra ha $k = \frac{\pi}{2}$, majd hasonlítsuk össze a $\sin(x)$ függvénnyel!

Ha n túl nagy a magas fokú interpolációs polinom miatt a kilengések is nagyok lesznek, ami a valóságnak a torzításához vezethet. Ezért célszerű relatív alacsony fokú polinommal dolgozni, például úgy, hogy csak egy pár jellegzetes alappontot választunk ki és ezek segítségével építjük fel a Lagrange -féle polinomot.

6.2.1.2. Véges differenciák Feltételezzük hogy adottak a következő egyenközű (ekvidisztáns) alappontok:

$$x_0 < x_1 < ... < x_n,$$

illetve egy f függvény értékei a fenti pontokban:

$$f(x_i) = y_i,\ i = 0, ..., n.$$

Mivel az alappontok ekvidisztánsak két egymásutáni különbsége konstans $h = x_{i+1} - x_i$. A h-t lépéstávolságnak vagy növekménynek nevezzük. x_0 és h függvényében kifejezhető az összes többi alappont:

$$x_1 = x_0 + h,\ x_2 = x_0 + 2h, ..., x_i = x_0 + ih, ...$$

119. Definíció. A

$$\begin{aligned} \Delta f(x_i) &= f(x_i + h) - f(x_i) = f(x_{i+1}) - f(x_i) \text{ vagy} \\ \Delta y_i &= y_{i+1} - y_i,\ \ i = 0, ..., n-1 \end{aligned}$$

kifejezést az f függvény első véges differenciájának nevezzük.

A fenti Δ operátor lineáris, vagyis:

$$\begin{aligned} \text{-adítiv:} &\qquad \Delta(f+g) = \Delta f + \Delta g \\ \text{-homogén :} &\qquad \Delta(cf) = c\Delta f\ . \end{aligned}$$

A magasabb rendű differenciákat a következőképpen értelmezzük:

$$\begin{aligned}\Delta^2 f(x_i) &= \Delta(\Delta f(x_i)) = \Delta(f(x_{i+1}) - f(x_i)) = \\ &= \Delta f(x_{i+1}) - \Delta f(x_i) = f(x_{i+2}) - 2f(x_{i+1}) + f(x_i) \\ &\dots \\ \Delta^n f(x_i) &= \Delta\left(\Delta^{n-1} f(x_i)\right).\end{aligned}$$

A gyakorlatban a véges differenciák kiszámításához táblázatokat használunk:

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$
x_0	y_0			
		$\Delta y_0 = y_1 - y_0$		
x_1	y_1		$\Delta^2 y_0 = \Delta y_1 - \Delta y_0$	
		$\Delta y_1 = y_2 - y_1$		$\Delta^3 y_0 = \Delta^2 y_1 - \Delta^2 y_0$
x_2	y_2		$\Delta^2 y_1 = \Delta y_2 - \Delta y_1$	
		$\Delta y_2 = y_3 - y_2$		
x_3	y_3			
$\vdots$	$\vdots$			

120. Példa. $x_i = i,\ i = 0, ..., 4,\ f(x) = x^2$.

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$
0	0			
		$\Delta y_0 = 1$		
1	1		$\Delta y_0^2 = 2$	
		$\Delta y_1 = 3$		$\Delta y_0^3 = 0$
2	4		$\Delta y_1^2 = 2$	
		$\Delta y_2 = 5$		$\Delta y_1^3 = 0$
3	9		$\Delta y_2^2 = 2$	
		$\Delta y_3 = 7$		
4	16			

A fenti példából kitűnik hogy az n-ed fokú polinomnak a $n+1$-ed rendű véges differenciája egyenlő zéróval:

$$f \in P_n \ \Rightarrow\ \Delta^{n+1} f(x_i) = 0.$$

Ugyanakkor $\Delta^n f(x_i)$ =konstans.

A fent értelmezett differencia:

$$\Delta f(x_i) = f(x_i + h) - f(x_i) = f(x_{i+1}) - f(x_i)$$

haladó-differenciának nevezzük. Ugyanakkor értelmezhető retrográd-:

$$\nabla f(x_i) = f(x_i) - f(x_i - h) = f(x_i) - f(x_{i-1}),$$

illetve centrális-differencia:

$$cf(x_i) = f\left(x_i + \frac{h}{2}\right) - f\left(x_i - \frac{h}{2}\right).$$

Ezek esetében hasonló képletek vezethetők le.

6.2.1.3. Newton-féle interpoláló polinom Tekintsük az $x_0, x_1, ..., x_n$ egyenközű alappontokat: $x_i = x_0 + ih$, $i = 0, ..., n$, illetve egy függvény értékeit az adott pontokban: $f(x_i) = y_i$, $i = 0, ..., n$. A feladat egy legfeljebb n-ed fokú polinom megszerkesztése ami interpolálja az (x_i, y_i), $i = 0, ..., n$ pontokat. A polinomot a következő alakban keressük:

(6.2.18)

$$N_n f(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + ... + a_n(x - x_0)..(x - x_{n-1}).$$

Az a_i együtthatókat az interpolációs feltételekből határozzuk meg:

$$\begin{aligned}
x &= x_0 \Rightarrow a_0 = y_0, \\
x &= x_1 \Rightarrow a_0 + a_1(x_1 - x_0) = y_1, \\
x &= x_2 \Rightarrow\ a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) = y_2, \\
&\ \ ... \\
x &= x_n \Rightarrow a_0 + a_1(x_n - x_0) + ... + a_n(x_n - x_0)..(x_n - x_{n-1}) = y_n
\end{aligned}$$

Mivel az alappontok egyenközűek $\Rightarrow$

$$\begin{cases}
a_0 & & & & & = y_0 \\
a_0 & +a_1 h & & & & = y_1 \\
a_0 & +a_1 2h & +a_2 2h^2 & & & = y_2 \\
\vdots & & & & & \\
a_0 & +a_1 nh & +a_2 n(n-1)h^2 & \dots & +a_n n(n-1)...2 \cdot 1 h^n & = y_n
\end{cases}.$$

A fenti háromszög alakú egyenletrendszer megoldása:

$$\begin{aligned} a_0 &= y_0 \\ a_1 &= \frac{y_1 - y_0}{h} = \frac{\Delta y_0}{h} \\ a_2 &= \frac{\Delta^2 y_0}{2!\, h^2} \\ &\vdots \\ a_n &= \frac{\Delta^n y_0}{n! h^n} \implies \end{aligned}$$

$$\begin{aligned} (N_n f)(x) &= y_0 + \frac{1}{1!}\frac{\Delta y_0}{h}(x - x_0) + \frac{1}{2!}\frac{\Delta^2 y_0}{h^2}(x - x_0)(x - x_1) + ... \\ &+ \frac{1}{n!}\frac{\Delta^n y_0}{h^n}(x - x_0)\dots(x - x_{n-1}) \end{aligned} \tag{6.2.20}$$

Az $|N_n f(x) - f(x)|$ abszolút hibát a (6.2.20) képletből mint a rákövetkező tag kapjuk meg:

$$|N_n f(x) - f(x)| = \frac{1}{(n+1)!}\frac{\Delta^{n+1} y_0}{h^{n+1}}(x - x_0)\,..\,(x - x_n) = \frac{1}{(n+1)!}\frac{\Delta^{n+1} y_0}{h^{n+1}}\varpi(x).$$

Ha egy adott x értékre kell a $N_n f$ polinomot kiszámítani, akkor jelöljük: $\alpha = \frac{x - x_0}{h}$, tehát $\frac{x - x_1}{h} = \frac{x - x_0 - (x_1 - x_0)}{h} = \alpha - 1$, stb. és a (6.2.20)-os képlet a következőképpen alakul:

(6.2.21)

$$(N_n f)(x) = y_0 + \frac{1}{1!}\alpha\,\Delta y_0 + \frac{1}{2!}\alpha(\alpha - 1)\,\Delta^2 y_0 + ... + \frac{1}{n!}\alpha(\alpha - 1)\,..\,(\alpha - n + 1)\,\Delta^n y_0.$$

A (6.2.21) képlet előnye, hogy az interpolációs csomópontok nem szerepelnek explicit módon.

121. Példa. Felhasználva a harmadfokú Newton interpoláló polinomot számítsuk ki $\sin(6^o)$ ha ismertek $\sin(x)$ függvény alábbi értékei:

x_i	5^o	7^o	9^o	11^o	13^o	15^o
y_i	0.087156	0.121869	0.156434	0.190809	0.224951	0.258819.

Mivel 6^o az 5^o és 7^o értékek közé esik az első négy értéket használjuk fel a harmadfokú Newton polinom meghatározására. Ehhez előbb a

véges differencia táblázatot szerkesztjük meg:

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$
5	<u>0.087156</u>			
		<u>0.034713</u>		
7	0.121869		<u>−0.000148</u>	
		0.034565		<u>−0.000042</u>
9	0.156434		-0.000190	
		0.034375		
11	0.190809			

A táblázatban szereplő aláhúzott véges differenciák segítségével megszerkesztjük a (6.2.20) képletnek megfelelő harmadfokú interpolációs polinomot:

$$N_3 f(x) = 0.087156 + \frac{0.034713}{2}(x-5) + \frac{(-0.000148)}{2!\cdot 2^2}(x-5)(x-7) + \\ + \frac{(-0000042)}{3!\cdot 2^3}(x-5)(x-7)(x-9).$$

A polinom kiértékeléséhez $x = 6$ -ban használhatjuk a (6.2.21) képletet ahol $h = 2,\ x = 6,\ x_0 = 5,\ \alpha = \frac{6-5}{2} \Longrightarrow$

$$\sin 6^o = 0.087156 + 0.5\ 0.034713 + \frac{0.5\,(0.5-1)\,(-0.000148)}{2} + \\ \frac{0.5\,(0.5-1)\,(0.5-2)}{3!}(-0.000042) = 0.104528.$$

Habár a Lagrange, illetve Newton-féle polinom egyforma adott pontokra, mégis, numerikus szempontból előnyösebb Newton alakú polinommal dolgozni mert újabb alappontok bevezetése eseteben nem kell az összes számítást újból elvégezni (a háromszög alakú egyenletrendszernek köszönhetően).

Nem egyenközű alappontokra a Newton polinom meghatározására az ú.n. *osztott differenciákat* használjuk.

Habár a képletek aránylag egyszerűek ha az alappontok egyenközű, approximációs szempontból mégsem a legmegfelelőbbek. Ennek oka az, hogy magas fokú polinomok esetében a kilengések is nagyok. Ez jól szemléltethető az ú.n. Runge-féle példán.

122. Példa. Legyen $f(x) = \frac{1}{1+25x^2},\ x \in [-1,1]$. Ha felbontjuk a $[-1,1]$ intervallumot $(n+1)$ különböző egyenközű alappontra, majd interpoláljuk azt kapjuk eredményként, hogy a központi részen az interpoláló polinom elég jól közelíti meg az f függvényt, viszont a peremen minél magasabb a polinom fokszáma annál rosszabb a megközelítés.

Alapvetően két módszerrel lehet a polinomiális interpoláció hátrányait kiküszöbölni: az egyik módszer abban áll, hogy egyenközű alappontok helyett a Csebisev-féle polinom gyökeit használjuk. Ezt a módszert akkor alkalmazhatjuk ha az alappontokat szabadon választhatjuk meg. A másik módszer abban áll hogy globális interpoláció helyett szakaszos interpolációt használunk.

6.2.1.4. A Csebisev-féle elsőfajú polinomok

123. Definíció. n-ed fokú (elsőfajú) Csebisev-féle polinomnak nevezzük a következő függvényt:

$$T_n : [-1,1] \to \mathbb{R},\ T_n(x) = \cos(n \arccos x).$$

Különböző n-re a polinomok a következőképpen néznek ki:

$$\begin{aligned} T_0(x) &= 1,\ T_1(x) = x, \\ T_2(x) &= \cos(2\arccos x) = 2\cos^2(\arccos x) - 1 = 2x^2 - 1, ... \end{aligned}$$

Felhasználva a $t = \arccos x$ változócserét meghatározhatunk egy rekurzív képletet a polinomok kiszámítására:

$$\cos((n+1)t) + \cos((n-1)t) = 2\cos(nt)\cos t \Leftrightarrow$$

$$T_{n+1}(x) + T_{n-1}(x) = 2T_n(x)\ x \ \Leftrightarrow$$

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0,\ x \in [-1,1]. \tag{6.2.22}$$

Tehát felhasználva hogy $T_0(x) = 1,\ T_1(x) = x \Rightarrow T_2(x) = 2x^2 - 1$ amit az előbb is megkaptunk.

Kiszámítjuk a T_n polinom gyökeit:

$$T_n(x) = 0 \Leftrightarrow \cos(n \arccos x) = 0 \Leftrightarrow n \arccos x = \frac{2k+1}{2}\pi \tag{6.2.23}$$

$$\arccos x = \frac{2k+1}{n}\frac{\pi}{2},\ k = 0, ..., n-1 \Leftrightarrow \tag{6.2.24}$$

$$x_k = \cos\left(\frac{2k+1}{n}\frac{\pi}{2}\right),\ k = 0, ..., n-1. \tag{6.2.25}$$

Mértanilag a Csebisev gyököket a következőképpen szerkesztjük meg: a felső félkört n egyenlő részre osztjuk majd ezeket a pontokat levetítjük az Ox tengelyre. Ezek a vetületek szolgálják a Csebisev polinom gyökeit. Észrevehető hogy ezeknek a sűrűsége nagyobb a széleken mint középen. Ennek köszönhetik a jobb interpolációs tulajdonságot.

Az összes n-ed fokú interpoláló polinom közül az (6.2.25) alappontokra felépített interpoláló polinomnak lesz a legjobb megközelítése.

124. Tétel. *Az összes* $[-1, 1]$ *intervallumon értelmezett* n*-ed fokú interpolációs polinom közül annak van a legkisebb hibája amelynek az alappontjai megegyeznek az Csebisev-féle* n*-ed fokú polinom gyökeivel.*

Ha a polinomok egy tetszőleges $x \in [a, b]$ intervallumon vannak értelmezve akkor egy

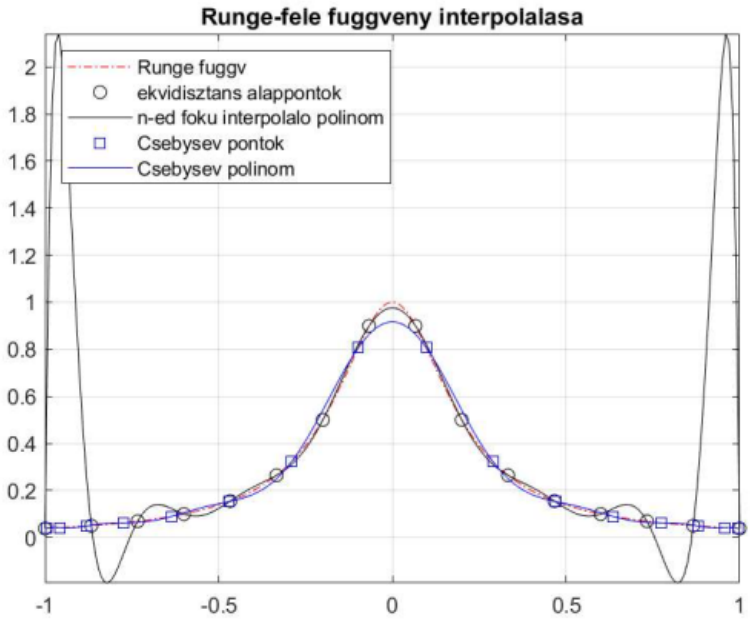
$$t = \frac{x - a + x - b}{b - a}$$

változócserével visszavezethetjük az $[-1, 1]$ intervallumra. Hasonlóan

$$x = \frac{b-a}{2} + \frac{b+a}{2}t$$

változócserével kivetíthetjük a Csebisev gyököket egy tetszőleges $[a, b]$ intervallumra:

$$x_k = \frac{b+a}{2} + \frac{b-a}{2}\cos\left(\frac{2k+1}{n}\frac{\pi}{2}\right),\ k = 0, ..., n-1.$$



6.2.3. ábra. Runge interpoláció

Egy másik módszer a polinomiális interpoláció hátrányainak a kiküszöbölésére abban áll hogy globális interpolálás helyett szakaszos interpolálást alkalmazzunk.

6.2.2. Szakaszos interpoláció A különbség a polinomiális és a szakaszos interpolálás között abban áll hogy ez utóbbiban nem egy (globális) polinomot szerkesztünk hanem minden részintervallumon egy meghatározott fokszámú polinomot. Ezzel aránylag alacsonyan tudjuk tartani a polinom fokszámat tekintet nélkül az interpoláló alappontok számára.

6.2.2.1. Lineáris interpoláció A legegyszerűbb függvény ami interpolálja az:

$$(x_A, y_A)\,,\ (x_B, y_B)$$

pontokat a következő első fokú polinom:

$$F(x) = y_A \frac{x_B - x}{h} + y_B \frac{x - x_A}{h},\ \text{ahol } h = x_B - x_A,$$

vagy:

$$F(x) = y_A + (x - x_A)\frac{y_B - y_A}{h}.$$

Ezt megkaphatjuk mint Lagrange -féle első fokú interpoláló polinom vagy egyszerű ellenőrzéssel.

Általánosításként tételezzük fel, hogy adottak az $(x_i, y_i)_{i=0}^n$ interpoláló pontok:

$$f(x_i) = y_i,\ i = 0, ..., n.$$

Ebben az esetben két egymásutáni alappont között, vagyis az $[x_i, x_{i+1}]$ intervallumon, a lineáris interpoláló függvény a következő lesz ($h_i = x_{i+1} - x_i$):

$$F(x) = y_i\frac{x_{i+1} - x}{h_i} + y_{i+1}\frac{x - x_i}{h_i},\ \text{ha } x \in [x_i, x_{i+1}] \tag{6.2.26}$$

vagy:

$$F(x) = y_i + (x - x_i)\frac{y_{i+1} - y_i}{h_i},\ x \in [x_i, x_{i+1}]\ . \tag{6.2.27}$$

Valóban az $[x_i, x_{i+1}]$ intervallumon F lineáris és teljesíti az interpoláló feltételeket:

$$F(x_i) = y_i,\ F(x_{i+1}) = y_{i+1}.$$

Összekapcsolva a lineáris szakaszokat az alappontokban egy tört vonalat kapunk, vagyis egy C^0 osztályú szakaszosan lineáris függvényt vagyis, az alappontokban az F függvény folytonos de a deriváltja már nem.

125. Példa. Szerkesszük meg a lineáris interpoláló függvényt az alábbi adatok esetében:

x_i	0	1	4	9
y_i	0	1	2	3

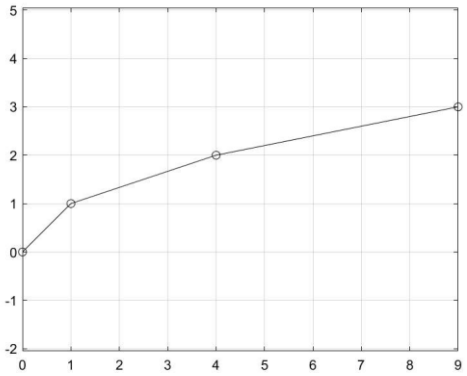
.

Az $[0, 1]$ intervallumon az interpoláló függvény:

$$F(x) = 0 \cdot \frac{1 - x}{1 - 0} + 1 \cdot \frac{x - 0}{1 - 0} = x,\ x \in [0, 1].$$

Hasonlóan

$$\begin{aligned} F(x) &= 1 \cdot \frac{4-x}{4-1} + 2 \cdot \frac{x-1}{4-1} = \frac{2}{3} + \frac{1}{3}x, \ x \in [1,4] \\ F(x) &= 2 \cdot \frac{9-x}{9-4} + 3 \cdot \frac{x-4}{9-4} = \frac{6}{5} + \frac{1}{5}x, \ x \in [4,9]. \end{aligned}$$



6.2.4. ábra. Lineáris interpoláció

6.2.2.2. Harmadfokú Hermite szakaszos interpoláció A lineáris interpoláció segítségével egy gyors képet alkothatunk magunknak az interpolációs folyamatról, a természetben viszont kevés olyan folyamat van ami modellezhető lenne a lineáris interpolációval. Ehelyett "simább" függvényekre lesz szükségünk amit magasabb fokú polinomokkal érhetünk el. Ehhez viszont több adatra lesz szükségünk.

Tételezzük fel hogy az $x_A,\ x_B$ alappontokban ismerjük nem csak az $y_A,\ y_B$ ordinátákat hanem a $d_A,\ d_B$ irányítényezőket is. Keressük azt a harmadfokú H polinomot ami eleget tesz a következő feltételeknek:

$$H(x_A) = y_A,\ H(x_B) = y_B, \tag{6.2.28}$$

$$H'(x_A) = d_A,\ H'(x_B) = d_B. \tag{6.2.29}$$

126. Definíció. A (6.2.29) feltételeknek eleget tevő polinomot harmadfokú Hermite -féle interpoláló polinomnak nevezzük.

Hasonlóan a Lagrange interpolációs polinommal, itt sem fogjuk használni a kanonikus bázist hanem az ú.n. Hermite -féle bázis függvényeket.

127. Tétel. *A harmadfokú Hermite -féle interpoláló polinomot a következőképpen lehet megadni:*

$$H(x) = y_A h_0(x) + y_B h_1(x) + d_A h_2(x) + d_B h_3(x) \tag{6.2.30}$$

ahol $h_i,\ i = \overline{0,3}$ *a Hermite -féle alappolinomokat jelöli:*

$$h_0(x) = \Phi\left(\frac{b-x}{h}\right),\ h_1(x) = \Phi\left(\frac{x-a}{h}\right) \tag{6.2.31}$$

$$h_2(x) = -h\ \Psi\left(\frac{b-x}{h}\right),\ h_3(x) = h\ \Psi\left(\frac{x-a}{h}\right) \tag{6.2.32}$$

és

$$\Phi(t) = 3t^2 - 2t^3,\ \Psi(t) = t^3 - t^2,\ h = b - a. \tag{6.2.33}$$

Bizonyítás. Bizonyításként a $h_i,\ i = \overline{0,3}$ alappolinomok bázis jellegére szorítkozunk.

$$h_0(a) = \Phi\left(\frac{b-a}{h}\right) = \Phi(1) = 1,\ h_1(a) = h_2(a) = h_3(a) = 0$$

Hasonlóan

$$h_1(b) = 1,\ h_0(b) = h_2(b) = h_3(b) = 0.$$

$$h_2'(a) = -h\ \Psi'\left(\frac{b-x}{h}\right)|_{x=a} = -h\left(-\frac{1}{h}\right)\left(3\left(\frac{b-x}{h}\right)^2 - 2\left(\frac{b-x}{h}\right)\right)|_{x=a} = 1,$$

$$h_0'(a) = h_1'(a) = h_3'(a) = 0,$$

□

$$h_3'(b) = h\ \Psi'\left(\frac{x-a}{h}\right)|_{x=b} = h\frac{1}{h}\left(3\left(\frac{x-a}{h}\right)^2 - 2\left(\frac{x-a}{h}\right)\right)|_{x=b} = 1,$$

$$h_0'(b) = h_1'(b) = h_2'(b) = 0.$$

Az előbbi eset általánosításaként tételezzük fel hogy az $x_0 < x_1 < ... < x_n$ alappontokban ismertek a következő adatok:

$$f(x_i) = y_i,\ f'(x_i) = d_i,\ i = 0, ..., n. \tag{6.2.34}$$

Ebben az esetben minden $[x_i, x_{i+1}]$ intervallumon külön megszerkesztjük a harmadfokú Hermite -féle interpoláló polinomot.

128. Tétel. *Az $[x_i, x_{i+1}]$ intervallumon a harmadfokú Hermite -féle interpoláló polinomot a következőképpen lehet megadni:*

$$H(x) = y_i h_0(x) + y_{i+1} h_1(x) + d_i h_2(x) + d_{i+1} h_3(x) \tag{6.2.35}$$

ahol a $h_i,\ i = \overline{0,3}$ Hermite alappolinomokat a következőképpen értelmezzük:

$$h_0(x) = \Phi\left(\frac{x_{i+1} - x}{h_i}\right),\ h_1(x) = \Phi\left(\frac{x - x_i}{h_i}\right) \tag{6.2.36}$$

$$h_2(x) = -h\ \Psi\left(\frac{x_{i+1} - x}{h_i}\right),\ h_3(x) = h\ \Psi\left(\frac{x - x_i}{h_i}\right) \tag{6.2.37}$$

ahol a lépés $h_i = x_{i+1} - x_i$, illetve a $\Phi,\ \Psi$ függvények (6.2.33) képletben vannak értelmezve.

Természetesen, ha a d_i deriváltak ismeretlenek ezeknek az értékeit meg lehet közelíteni az x_i, y_i adatokból kiindulva. Például, hármasával csoportosítva, minden alappontban a derivált egyenlő a három pontra épített parabola irányténye zőjével (Bessel -féle interpoláció).

Mivel az alappontokban a deriváltak megegyeznek az összetett Hermite -féle interpoláló polinom C^1 osztályú lesz.

129. Példa. Szerkesszünk egy Hermite -féle interpoláló függvényt a következő adatokra:

x_i	0	1	4	9
y_i	0	1	2	3

. Elsősorban minden ponthoz hozzá kell rendelni egy irányténye zőt. Legyen az (x_1, y_1) pontban az irányténye ző egyenlő a szomszédos pontok által meghatározott egyenes irányténye zőjével:

$$d_1 = \frac{y_2 - y_0}{x_2 - x_0} = \frac{2 - 0}{4 - 0} = \frac{1}{2}.$$

Hasonlóan:

$$d_2 = \frac{y_3 - y_1}{x_3 - x_1} = \frac{3-1}{9-1} = \frac{1}{4}.$$

A d_0-t egyenlővé tesszük az első szakasz irányténye zőjével:

$$d_0 = \frac{y_1 - y_0}{x_1 - x_0} = 1.$$

Hasonlóan

$$d_3 = \frac{y_3 - y_2}{x_3 - x_2} = \frac{1}{5}.$$

A (6.2.35) képletből a $[0,1]$ intervallumon

$$H(x) = 0 \cdot h_0(x) + 1 \cdot h_1(x) + 1 \cdot h_2(x) + \frac{1}{2} \cdot h_3(x)$$

ahol

$$\begin{aligned} h_1(x) &= \Phi(x) = 3x^2 - 2x^3, \\ h_2(x) &= -\Psi(1-x) = -(1-x)^3 + (1-x)^2, \; h_3(x) = \Psi(x) = x^3 - x^2. \end{aligned}$$

Tehát

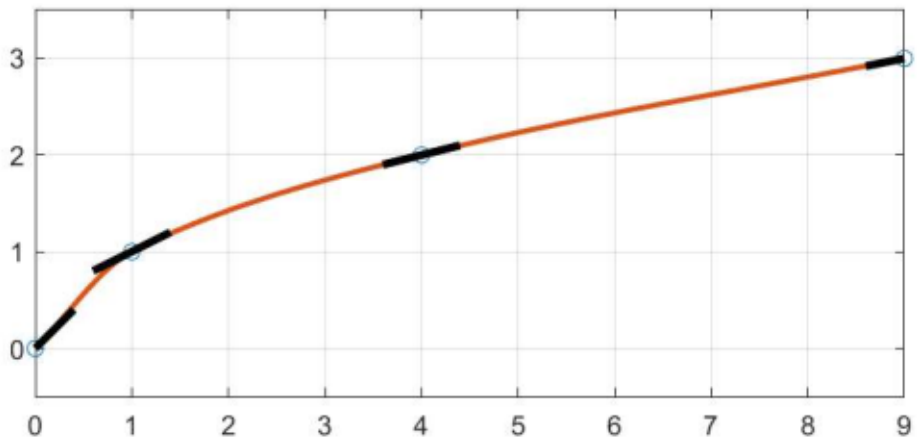
$$H(x) = 3x^2 - 2x^3 - (1-x)^3 + (1-x)^2 + \frac{x^3 - x^2}{2}, \; x \in [0,1].$$

Hasonlóan, az $[1,4]$ intervallumon:

$$\begin{aligned} H(x) = \frac{(4-x)^2}{3} - \frac{2(4-x)^3}{27} + \frac{2(x-1)^2}{3} - \frac{4(x-1)^3}{27} - \\ - \frac{(4-x)^3}{18} + \frac{(4-x)^2}{6} + \frac{(x-1)^3}{36} - \frac{(x-1)^2}{12}, \end{aligned}$$

illetve a $[4,9]$ intervallumon:

$$\begin{aligned} H(x) = \frac{6(9-x)^2}{25} - \frac{4(9-x)^3}{125} + \frac{9(x-4)^2}{25} - \frac{4(x-4)^3}{125} - \\ - \frac{(9-x)^3}{100} + \frac{(9-x)^2}{20} + \frac{(x-4)^3}{125} - \frac{(x-4)^2}{25}. \end{aligned}$$



6.2.5. ábra. Hermite interpoláció

6.2.2.3. Harmadfokú spline interpoláció Az előbbi interpolációs eljárásnál simább függvényt kapunk ha a másodrendű derivált folytonosságát is megköveteljük. Az így szerkesztett interpolációs polinom neve spline.

Ha az előbbiekben a deriváltakat úgy határozzuk meg hogy a kapott függvény kétszer folytonosan deriválható legyen, akkor a harmadfokú (cubic) spline interpolációs módszert kapjuk.

Legyenek tehát $(x_i, y_i)_{i=0}^n$ az interpoláló pontok. Minden $[x_i, x_{i+1}]$ szakaszon egy s_i, $i = 0, ..., n-1$ ($s_i : [x_i, x_{i+1}] \to \mathbb{R}$) harmadfokú polinomot szerkesztünk meg úgy, hogy az alappontokban az első fokú és a másodfokú derivált folytonos legyen, tehát $4n$ ismeretlenünk lesz. Az ismeretlenekhez hozzárendelt egyenleteket a következő feltételekből határozzuk meg:

- $(n+1)$ egyenlet az interpoláló feltételekből:

$$s_i(x_i) = y_i, \ i = 0, \ldots, n,$$

- $(n-1)$ egyenlet a folytonossági feltételekből a belső alappontokban:

$$s_i(x_i) = s_{i+1}(x_i), \ i = 1, \ldots, n-1,$$

- $(n-1)$ egyenlet az első derivált folytonossági feltételekből a belső alappontokban:

$$s_i'(x_i) = s_{i+1}'(x_i), \ i = 1, \ldots, n-1,$$

- $(n-1)$ egyenlet a második derivált folytonossági feltételből a belsőalappontokban:

$$s_i''(x_i) = s_{i+1}'' x_i, \ i = 1, \ldots, n-1,$$

összesen $(4n-2)$ egyenlet. Hogy egyértelműen meghatározott egyenletrendszert kapjunk szükség van meg 2 feltételre. Ezeket rendszerint a széleken levő alappontokhoz kötjük (perem feltételek). A $4n \times 4n$-es lineáris egyenletrendszer megoldása szolgáltatja a szakaszokra értelmezett polinomok együtthatóit.

A továbbiakban ismertetjük a leggyakrabban használt peremfeltételeket:

(1) Ha ismertek a d_0, d_n iránytényezők a széleken, akkor a következő kikötéseket használjuk:

$$s_0'(x_0) = d_0, \ s_{n-1}'(x_n) = d_n.$$

Az így felépített függvényt teljes ("complete") spline"-nak nevezzük.

(2) Ha ismertek a másodrendű deriváltak értékei D_0, D_n a széleken, akkor a következő kikötéseket használjuk:

$$s_0''(x_0) = D_0, \ s_{n-1}''(x_n) = D_n.$$

(3) Az egyik legrégebbi perem feltétel az ú.n. "természetes" feltétel:

$$s_0''(x_0) = 0, \ s_{n-1}''(x_n) = 0.$$

A "natural spline" név a rajzoló szerkezetre (spline, splain) vezethető vissza, ugyanis a végpontokban a fémrúd görbülete nulla (lásd a gyakorlati kivitelezését).

(4) Ha a deriváltak ismeretlenek egy másik peremfeltétel az ún. "not-a-knot", vagyis az első és utolsó belső interpolációs pont

inaktív. Ehhez kikötjük hogy a harmadrendű derivált s''' folytonos legyen x_1, és x_{n-1}-ben. Eredményként $s_0 = s_1$illetve $s_{n-1} = s_{n-2}$.

Az összes szakaszonként értelmezett harmadfokú polinomok közül a köbös (cubic) spline függvénynek van maximális simasági foka, nevezetesen C^2. Ennél magasabb simasági fokkal csak a globális harmadfokú polinom rendelkezik.

A harmadfokú spline függvény általánosítása a következő:

130. Definíció. Az $(x_i)_{i=0}^n$ alappontokhoz hozzárendelt $(2n+1)$-ed fokú spline-nak azt a függvényt nevezzük, amely minden egyes $[x_i, x_{i+1}]$ szakaszon $2n+1$-ed fokú polinom, továbbá az $x_0, ..., x_n$ pontokban a deriváltak $2n$-ed rendig bezárólag eleget tesznek a folytonossági feltételeknek.

Az egyik legegyszerűbb spline a már ismertet lineáris interpoláló függvény (a tört vonal) aminek neve lineáris spline. Minden szakaszon elsőfokú polinom a folytonossági rendje pedig C^0.

Bármilyen fokú sline függvény értelmezhető, de a páratlan fokú spline-ok esetében az az előny hogy a peremfeltételeket szimmetrikusan lehet kezelni.

131. Példa. Szerkesszük meg a köbös spline interpolációs függvényt az alábbi adatok esetében különböző peremfeltételek esetében:

x_i	0	1	4	9
y_i	0	1	2	3

.

$$s(x) = \begin{cases} s_1(x), & x \in [0,1) \\ s_2(x), & x \in [1,4) \\ s_3(x), & x \in [4,9] \end{cases}$$

ahol

$$s_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i, \ i = 0,1,2$$

A 12 együtthatót a következő lineáris egyenletrendszerből számítjuk ki:

$$\begin{cases} 0 \cdot a_0 + 0 \cdot b_0 + 0 \cdot c_0 + d_0 = 0 \\ a_1 + b_1 + c_1 + d_1 = 1 \\ 64a_2 + 16b_2 + 4c_2 + d_2 = 2 \\ 279a_2 + 81b_2 + 9c_2 + d_2 = 3 \\ a_0 + b_0 + c_0 + d_0 = a_1 + b_1 + c_1 + d_1 \\ 64a_1 + 16b_1 + 4c_1 + d_1 = 64a_2 + 16b_2 + 4c_2 + d_2 \\ 3a_0 + 2b_0 + c_0 = 3a_1 + 2b_1 + c_1 \\ 48a_1 + 8b_1 + c_1 = 48a_2 + 8b_2 + c_2 \\ 6a_0 + 2b_0 = 6a_1 + 2b_1 \\ 24a_1 + 2b_1 = 24a_2 + b_2 \\ 2b_0 = 0 \\ 54a_2 + 2b_2 = 0 \end{cases}$$

ahol az utolsó két egyenletben a "természetes" peremfeltételeket használtuk.

A spline függvények esetében is a kanonikus bázis felíráson kívül van más alkalmasabb bázis az ú.n. B-spline függvények.

6.3. Kétváltozós lineáris (bilineáris) interpoláció

Hasonlóan az egyváltozós esettel, a kétváltozós függvények esetében felmerül a probléma hogy bizonyos diszkrét pontokból kiindulva egy interpoláló felületet kell szerkeszteni. A legegyszerűbb módszer, a bilineáris interpoláció, akkor alkalmazható ha az adott diszkrét pontok rácsszerűen helyezkednek el.

Legyen $f(x, y)$ egy kétváltozós függvény amelynek csak $(x_i, y_j)_{i=0,j=0}^{m,\,n}$ pontokban ismerjük az értékeit:

$$z_{ij} = f(x_i, y_j), \ i = 0, ..., m, \ j = 0, ..., n.$$

Az (x_i, y_j) párokról azt mondjuk hogy egy $m \times n$-es háló (rács) csomópontjait alkotják.

$x \setminus y$	y_0	...	y_j	...	y_n
x_0	z_{00}		z_{0j}		z_{0n}
$\vdots$					
x_i	z_{i0}		z_{ij}		z_{in}
$\vdots$					
x_m	z_{m0}		z_{mj}		z_{mn}.

A feladat egy kétváltozós $F(x, y)$ függvény megszerkesztése amely megegyezik a háló pontjaiban f-el és megközelíti egy tetszőleges $(\overline{x}, \overline{y})$ pontban. A legegyszerűbb módszer természetesen a (bi)lineáris interpoláció amit úgy valósítunk meg hogy mind a két tengely irányában alkalmazzuk a lineáris interpolációt i.e., egyszer az egyik, majd a másik változót tekintjük konstansnak.

Az első lépés meghatározni az $(\overline{x}, \overline{y})$ szomszédságában lévő pontokat, vagyis közrefogni a $\overline{x}$ pontot két egymásutáni Ox tengelyen levő alapponttal:

$$x_i < \overline{x} < x_{i+1},$$

majd hasonlóan:

$$y_j < \overline{y} < y_{j+1}.$$

Az így meghatározott pontok illetve az f függvény értékei ezekben a pontokban az alábbi táblázatban szerepelnek:

$x \diagdown y$	y_j	$\overline{y}$	y_{j+1}
x_i	$f(x_i, y_j) = z_{ij}$		$f(x_i, y_{j+1}) = z_{i,j+1}$
$\overline{x}$		$f(\overline{x}, \overline{y}) = ?$	
x_{i+1}	$f(x_{i+1}, y_j) = z_{i+1,j}$		$f(x_{i+1}, y_{j+1}) = z_{i+1,j+1}$.

Rögzítjük például az $y = y_j$ változót, ekkor az F függvény egyváltozós válik (x változó szerinti): $F(x, y_j)$. Az így kapott egyváltozós függvényre alkalmazzuk a (6.2.27) lineáris interpolációs képletet:

$$F(x, y_j) = f(x_i, y_j) + (x - x_i) \frac{f(x_{i+1}, y_j) - f(x_i, y_j)}{x_{i+1} - x_i}, \; x \in [x_i, x_{i+1}],$$

ahonnan

$$F(\overline{x}, y_j) = f(x_i, y_j) + (\overline{x} - x_i) \frac{f(x_{i+1}, y_j) - f(x_i, y_j)}{x_{i+1} - x_i}. \tag{6.3.1}$$

Hasonlóan rögzített $y = y_{j+1}$ azt kapjuk hogy:

(6.3.2)

$$F(\overline{x}, y_{j+1}) = f(x_i, y_{j+1}) + (\overline{x} - x_i) \frac{f(x_{i+1}, y_{j+1}) - f(x_i, y_{j+1})}{x_{i+1} - x_i}.$$

A már ismert $F(\overline{x}, y_j), F(\overline{x}, y_{j+1})$ adatokra újból alkalmazzuk a (6.2.27) képletet és rögzített $x = \overline{x}$-re azt kapjuk hogy:

(6.3.3)

$$F(\overline{x}, y) = F(\overline{x}, y_j) + (y - y_j) \frac{F(\overline{x}, y_{j+1}) - F(\overline{x}, y_j)}{y_{j+1} - y_j}, \ y \in [y_j, y_{j+1}]$$

vagyis:

$$F(\overline{x}, \overline{y}) = F(\overline{x}, y_j) + (\overline{y} - y_j) \frac{F(\overline{x}, y_{j+1}) - F(\overline{x}, y_j)}{y_{j+1} - y_j}. \tag{6.3.4}$$

A (6.3.4) értéknél jobb megközelítést kapunk ha a bilineáris interpoláció helyett a bicubic sémát alkalmazzuk, vagyis mindkét tengely mentén harmadfokú polinommal interpolálunk.

132. Példa. Az $f(\theta, \varphi) = \int_0^{\varphi} \frac{d\varphi}{\sqrt{1 - \sin^2 \theta \sin^2 \varphi}}$ elliptikus integrálra ismert a következő táblázatban szereplő értékek:

$\theta \diagdown \varphi$	50°	55°	60°	65°	70°
50°	0.9401	1.05	1.1643	1.2833	1.4068
60°	0.9647	1.0848	1.2125	1.3489	1.4944
70°	0.9876	1.1186	1.2619	1.4199	1.5959
80°	1.0044	1.1444	1.3014	1.4810	1.6918

Számítsuk ki bilineáris interpolációval $f(55, 53)$ közelítő értékét.

Bizonyítás. Az (55,53) pont szomszédságában lévő adatok:

$x \diagdown y$	50	53	55
50	0.9401		1.05
55	0.9524		1.0675
60	0.9647		1.0848

Alkalmazzuk a (6.3.2) ill. (6.3.3) képletet:

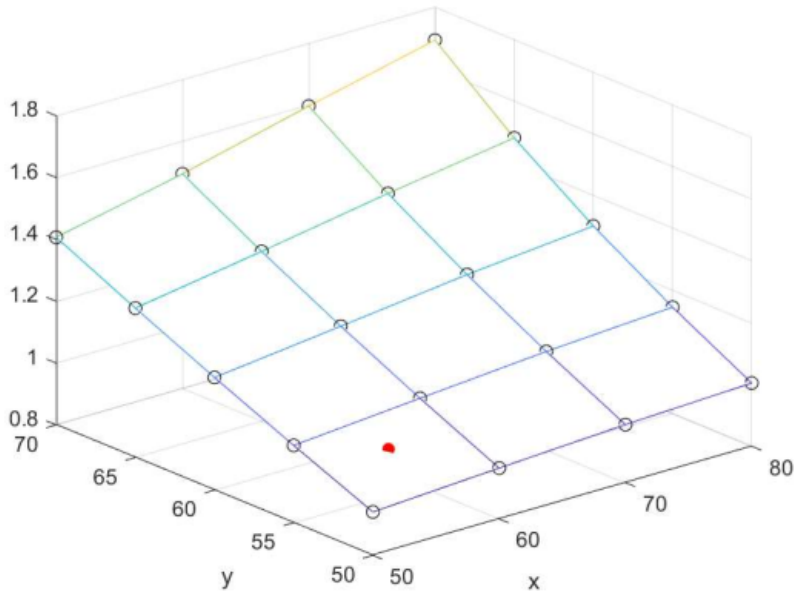
$$F(55,50) = F(50,50) + (55-50)\frac{F(60,50)-F(50,50)}{60-50} =$$
$$0,9401 + 5 * (0,9647 - 0,9401)/10 = 0.9524$$

$$F(55,55) = f(50,55) + (55-60)\frac{F(60,55)-F(50,55)}{60-50} = 1.0675$$

majd rögzítjük a $x = 55$ változót és alkalmazzuk a (6.3.4) képletet:

$$F(55,53) = F(55,50) + (53-50)\frac{F(55,55)-F(55,50)}{55-50} =$$
$$0.9524 + 3 * (1.0675 - 0.9524)/5 = 1.0214.$$

□



6.3.1. ábra. Bilineáris interpoláció

6.4. Függvény approximáció a legkisebb négyzetek módszerével

Amint az interpolációnál láttuk, ha ismert egy f függvény értékei az $x_i,\ i = 0, ..., n$ csomópontokban akkor felépíthető egy F approximáció úgy, hogy $F(x_i) = f(x_i) = y_i$. Ha viszont az $f(x_i)$ értékek nem pontosak (például hozzávetőleges mérések alapján számított értékek) akkor a pontok interpolációja nem

Az adott pontok elhelyezkedésétől függően a közelítés különböző fokszámú polinommal történhet. A lineáris függvénnyel való közelítés: $F(x) = ax + b$. Az interpolációs felt következik, hogy

$$\begin{cases} ax_0 + b = & y_0 \\ ax_1 + b = & y_1 \\ \quad \vdots & \\ ax_n + b = & y_n \end{cases}$$

vagy mátrix alakban

$$\begin{pmatrix} x_0 & 1 \\ x_1 & 1 \\ \vdots & \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \Leftrightarrow X \begin{pmatrix} a \\ b \end{pmatrix} = Y,$$

ahol $X = \begin{pmatrix} x_0 & 1 \\ x_1 & 1 \\ \vdots & \\ x_n & 1 \end{pmatrix}$, $Y = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$. A túlhatározott le megoldására használhatjuk a 4.5 alfejezetben tárgyalt módszert .....

$$X^t X \begin{pmatrix} a \\ b \end{pmatrix} = X^t Y$$

vagyis

$$\begin{cases} \sum_{i=1}^n x_i^2 \cdot a + \sum_{i=1}^n x_i \cdot b & = & \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i \cdot a + \sum_{i=1}^n 1 \cdot b & = & \sum_{i=1}^n y_i \end{cases} . \tag{6.4.1}$$

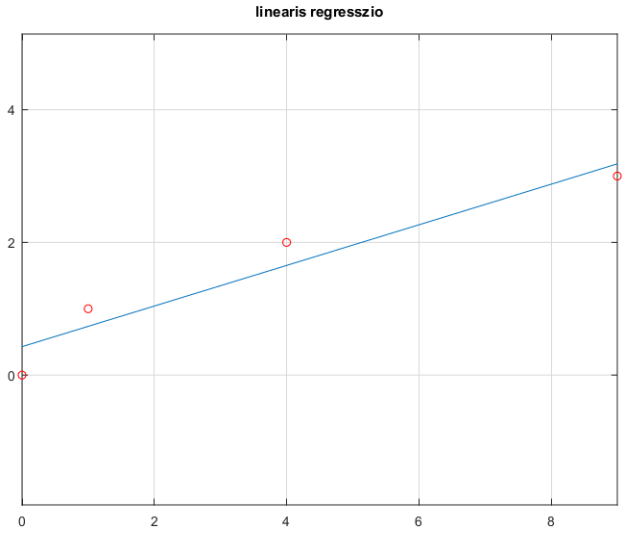
133. Példa. A

x_i	0	1	4	9
y_i	0	1	2	3

pontok esetében az (6.4.1) egyenletrendszer

$$\begin{cases} 98a + 14b &= 36 \\ 14a + 4b &= 6 \end{cases},$$

és a megoldása: $a = \frac{15}{49}$, $b = \frac{3}{7}$, tehát a pontokra legjobban illeszkedő egyenes egyenlete: $y = \frac{15}{49}x + \frac{3}{7}$.



6.4.1. ábra. Lineáris regresszió

Lineáris közelítés helyett használhatunk egy tetszőleges a $(x, y)_i$ pontokat közelíthetjük egy tetszőleges $k \leq n$ fokú polinommal A () egyenletrendszer általánosítható egy $k-$ad fokú közelítő polinomra

$$F(x) = a_0 x^k + \ldots + a_{k-1}x + a_k.$$

$$\begin{cases} M_{2k} \cdot a_0 + M_{2k-1}a_1 + \ldots + M_k \cdot a_k &= V_k \\ M_{2k-1} \cdot a_0 + M_{2k-2}a_1 + \ldots + M_{k-1} \cdot a_k &= V_{k-1} \\ \vdots & \\ M_k \cdot a_0 + M_{k-1}a_1 + \ldots + M_0 \cdot a_k &= V_0 \end{cases}$$

ahol

$$M_j = \sum_{i=1}^{n} x_i^j,\ j = 0, \ldots, 2k,$$

$$V_j = \sum_{i=1}^{n} x_i^j \cdot y_i,\ j = 0, \ldots, k.$$

134. Példa. Az előbbi adatokkal az

$$F(x) = a_2x^2 + a_1x + a_0$$

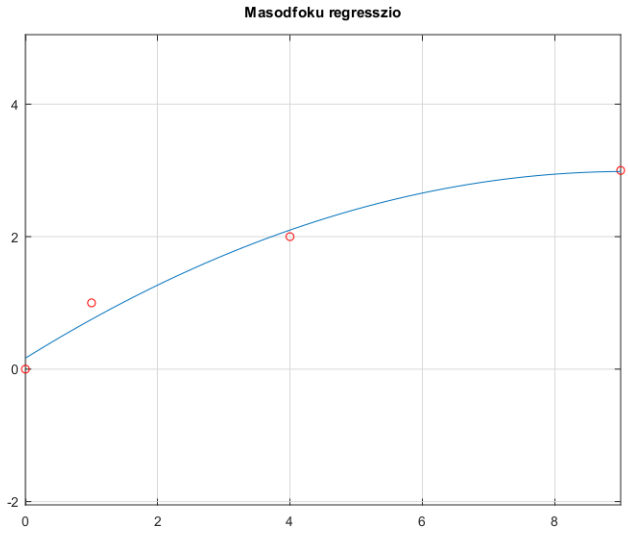
másodfokú regressziós közelítő polinom együtthatóit az alábbi egyenletrendszerből határozzuk meg:

$$\left\{\begin{array}{rcl} \sum_{i=1}^n x_i^4 \cdot a_0 + \sum_{i=1}^n x_i^3 \cdot a_1 + \sum_{i=1}^n x_i^2 \cdot a_2 & = & \sum_{i=1}^n x_i^2 y_i \\ \sum_{i=1}^n x_i^3 \cdot a_0 + \sum_{i=1}^n x_i^2 \cdot a_1 + \sum_{i=1}^n x_i \cdot a_2 & = & \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i^2 \cdot a_0 + \sum_{i=1}^n x_i \cdot a_1 + \sum_{i=1}^n 1 \cdot a_2 & = & \sum_{i=1}^n y_i \end{array}\right.$$

$\Leftrightarrow$

$$\left\{\begin{array}{rcl} 6818a_0 + 794a_1 + 98a_2 & = & 276 \\ 794a_0 + 98a_1 + 14a_2 & = & 36 \\ 98a_0 + 14a_1 + 4a_2 & = & 6 \end{array}\right.$$

a megoldás $a_0 = \frac{-37}{1086}$, $a_1 = \frac{673}{1086}$, $a_2 = \frac{30}{181}$.



6.4.2. ábra. Másodfokú regresszió

7. FEJEZET

Bézier görbék, Bézier felületek

A Bézier görbék a CAGD tantárgy alapjait képezik. Habár elméleti szinten már a 30-as években jelentek meg dolgozatok e téren, az igazi fejlődése a 60-as években kezdődött P. de Casteljau (Citroen), P. Bézier (Renault), S. Coons (Ford), W. Gordon (General Motors), J. Ferguson (Boeing) munkásságával.

Az interpolációtól eltérően, a Bézier görbék szerkesztése a "free form" kategóriába sorolható vagyis, a formatervező szabadon választ meg bizonyos pontokat amelyeket összekötve egy megközelítő képet adnak a görbe alakjáról.

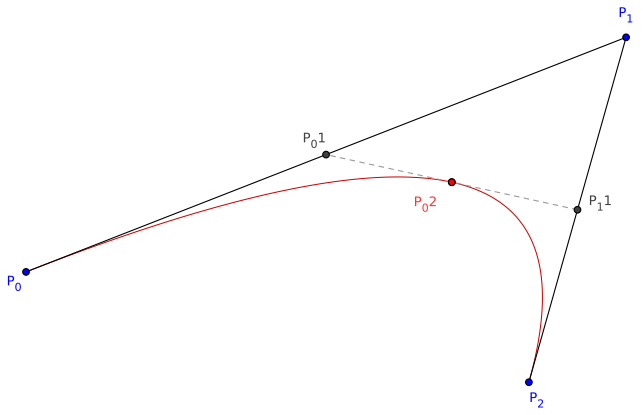
7.1. Bézier görbe szerkesztése "divide et impera" algoritmussal

Tekintsük a P_0, P_1, P_2 pontokat. Egy P_0 és P_2 pontok közé húzódó görbét szeretnénk szerkeszteni, ugyanakkor a görbe alakját a P_1 ponttal fogjuk befolyásolni.

A P_i pontokat kontroll pontoknak, a $\overline{P_0P_1P_2}$ kontroll poligonnak nevezzük.

Az algoritmus abban áll hogy a görbe pontjait sorozatos felezéssel hozzuk létre.

Legyen $P_0^{(1)}$ a $\overline{P_0P_1}$ szakasz, $P_1^{(1)}$ pedig a $\overline{P_1P_2}$ szakasz felezőpontja.



A $\overline{P_0^{(1)}P_1^{(1)}}$ szakaszt felezzük a $P_0^{(2)}$ ponttal.

Az így megszerkesztett $P_0^{(2)}$ pont, rajta lesz a görbén.

Hogy a görbe többi pontjait megkapjuk a kapott pontokat: átnevezzük majd folytatjuk az osztási algoritmust. Jelöljük a P_0, $P_0^{(1)}$, $P_0^{(2)}$ pontokat P_0, P_1, P_2-vel majd alkalmazzuk az előbb ismertetett eljárást.

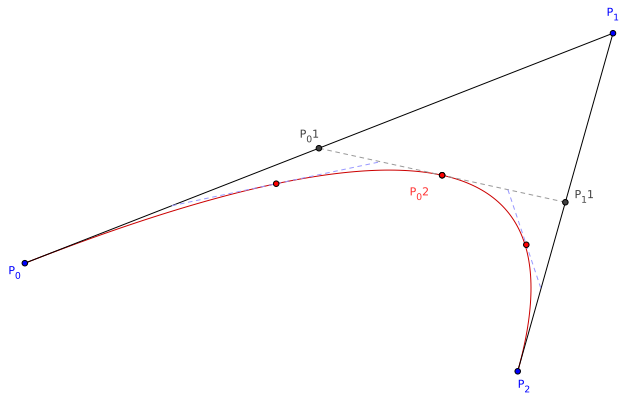
Az újonnan kapott $P_0^{(2)}$ pont úgyszintén rajta lesz a görbén.

Hasonlóan járunk el az első felezésből származó $P_0^{(2)}, P_1^{(1)}, P_2$ pontokkal.

A három $P_0^{(2)}$ pont illetve az algoritmus további alkalmazása nyomán létrejövő $P_0^{(2)}$ pontok alkotják a négyzetes Bézier görbét.

7.2. Négyzetes Bézier görbék

Az előbb említett algoritmusnál egy sajátos esetet használtunk nevezetesen a felezést; ezt most általánosítani fogjuk olyan értelemben hogy bevezetünk egy $t \in [0, 1]$ paramétert ami a szakaszok felosztásának arányait adja meg. Minden egyes t értékre kapunk egy pontot a görbén. Vegyük például a $t = 3/4$ értéket. Minden egyes lépésben

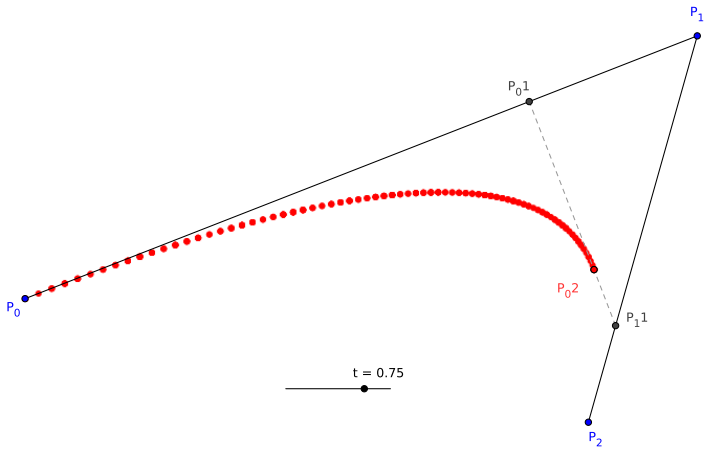


7.1.1. ábra. Négyzetes Bézier görbe

a további pontokat a meglévő pontok konvex kombinációjaként szerkesztjük meg:

$$P_0^{(1)} = (1-t)\,P_0 + tP_1 \tag{7.2.1}$$

$$P_1^{(1)} = (1-t)\,P_1 + tP_2. \tag{7.2.2}$$



A görbén elhelyezkedő $P_0^{(2)}$pontot a már meglévő $P_0^{(1)}, P_1^{(1)}$ pontok konvex kombinációjaként számítjuk ki:

$$P_0^{(2)} = (1-t)\, P_0^{(1)} + tP_1^{(1)}. \tag{7.2.3}$$

Módosítva t paraméter értékeit 0-tól 1-ig az előbbi konvex kombinációk megkapjuk a görbe összes pontját.

Mivel a pontok helyzete függ a t paramétertől, a további jelöléseknél ezt figyelembe vesszük. Azt hogy a $P_0^{(2)}(t)$ pontok írják le a görbét a következőképpen jelöljük:

$$P_0^{(2)}(t) = P(t), \ t \in [0,1].$$

Felhasználva a (7.2.3),(7.2.2) képleteket azt kapjuk, hogy:

$$\begin{aligned} P(t) = P_0^{(2)}(t) = (1-t)\, P_0^{(1)}(t) + tP_1^{(1)}(t) \\ = (1-t)\left[(1-t)\, P_0 + tP_1\right] + t\left[(1-t)\, P_1 + tP_2\right] \\ = (1-t)^2 P_0 + 2t(1-t)\, P_1 + t^2 P_2 \end{aligned}$$

tehát a görbe egyenlete:

$$P(t) = (1-t)^2 P_0 + 2t(1-t) P_1 + t^2 P_2,\ t \in [0,1]. \tag{7.2.4}$$

A görbe egyenletéből kiolvasható, hogy ez egy másodfokú görbe.

A görbe legfontosabb tulajdonságai a következők:

- P_0, P_2 a görbe két végpontja:
$$P(0) = P_0, P(1) = P_2;$$
- A polinom alakjából következik a görbe folytonossága;
- A görbe tangense a P_0, P_2 végpontokban megegyezik a $\overline{P_0P_1}, \overline{P_1P_2}$ szakaszokkal:
$$\frac{d}{dt}P(t) = -2(1-t) P_0 + 2(1-2t) P_1 + 2tP_2$$
ahonnan:
$$\frac{d}{dt}P(0) = 2(P_1 - P_0),\ \frac{d}{dt}P(1) = 2(P_2 - P_1);$$
- A kontroll pontok által alkotott konvex burkoló (a mi esetünkben a $P_0P_1P_2$ háromszög) tartalmazza a görbét.

Az (7.2.4) egyenletben az együtthatók összege:

$$(1-t)^2 + 2t(1-t) + t^2 = 1$$

vagyis a görbe konvex kombinációja a kontroll pontoknak.

Összeillesztési szempontokat tartva szem előtt hatékonyabb a harmadfokú Bézier görbék használata.

7.3. Harmadfokú Bézier görbék

A (7.2.4) képletben szereplő együtthatókat:

$$(1-t)^2 = b_0^2(t),\ 2t(1-t) = b_1^2(t),\ t^2 = b_2^2(t)$$

Bernstein-féle másodfokú alap-polinomoknak nevezzük. A fenti jelöléseket felhasználva a másodfokú Bézier görbe egyenlete felírható:

$$P(t) = (1-t)^2 P_0 + 2t(1-t) P_1 + t^2 P_2 = \sum_{i=0}^{2} b_i^2(t) P_i,\ t \in [0,1].$$

Hasonlóan értelmezhetők a harmadfokú Bernstein-féle alap polinomok:
(7.3.1)

$$b_0^3(t) = (1-t)^3,\ b_1^3 = 3t(1-t)^2,\ b_2^3 = 3t^2(1-t),\ b_3^3(t) = t^3,$$

és általánosan az n-ed fokú Bernstein-féle alap-polinomok:

$$b_i^n(t) = C_n^i t^i (1-t)^{n-i},\ i = 0, ..., n, \tag{7.3.2}$$

ahol $t \in [0,1]$. Ezeknek az alap-polinomoknak a főtulajdonsága, hogy az egységnek egy partícióját alkotják, vagyis:

$$b_i^n(t) \geq 0,\ \sum_{i=0}^{n} b_i^n(t) = 1.$$

A harmadfokú Bézier -féle görbék esetében ugyanúgy járhatunk el mint a másodfokúak esetében, nevezetesen értelmezhetjük őket rekurrens képlettel vagy analitikusan.

Legyen P_0, P_1, P_2, P_3 kontroll pontok.

135. Definíció. Mértanilag a harmadfokú Bézier görbe a $P(t) = P_3^{(3)}(t)$ képlettel szerkeszthető meg ahol:

$$P_i^{(k)}(t) = \begin{cases} (1-t)\,P_i^{(k-1)}(t) + tP_{i+1}^{(k-1)}(t),\ \text{ha } k = 1,2,3,\ i = 0,..,3-k \\ P_i \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad ,\ \text{ha } k = 0 \end{cases}$$

és $t \in [0,1]$.

136. Definíció. Analitikusan a harmadfokú $P(t)$ Bézier görbe a következőképpen értelmezhető:

$$P(t) = \sum_{i=0}^{3} b_i^3(t) P_i = b_0^3(t) P_0 + b_1^3 P_1 + b_2^3 P_2 + b_3^3(t) P_3 \tag{7.3.3}$$

$$= (1-t)^3 P_0 + 3t(1-t)^2 P_1 + 3t^2(1-t) P_2 + t^3 P_3 \tag{7.3.4}$$

ahol $t \in [0,1]$.

A másodfokú görbék esetében említett tulajdonságok itt is érvényesek.

A Bézier görbék előnye hogy a *görbe alakja követi a kontroll poligon alakját*, ebből kifolyólag egy görbe szerkesztéséhez elégséges a kontroll pontok módosítása.

A gyakorlatban, a görbén lévő - t paraméternek megfelelő- pontnak a megszerkesztéséhez, a következő táblázatot használjuk (de Casteljau algoritmus):

$$\begin{array}{llll} P_0 & & & \\ P_1 & P_0^{(1)}(t) & & \\ P_2 & P_1^{(1)}(t) & P_0^{(2)}(t) & \\ P_3 & P_2^{(1)}(t) & P_1^{(2)}(t) & P_0^{(3)}(t) \end{array}$$

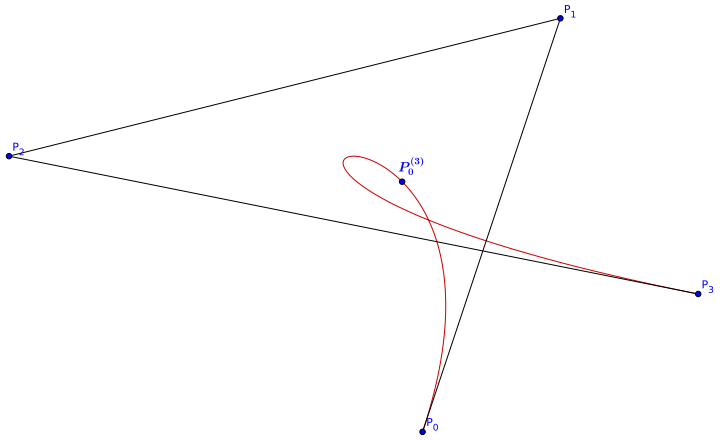
A táblázat jobb-alsó sarkában lévő érték adja meg a pont helyzetét a görbén t paraméternek megfelelően.

137. Példa. Szerkesszük meg az alábbi pontok esetében $t = 1/3$ -nek megfelelő pontot a harmadfokú Bézier görbén: $P_0(1,-1), P_1(2,2), P_2(-2,1), P_3(3,0)$.

Oszlopként írva a koordinátákat a de Casteljau algoritmus a következőképpen alakul:

$$\begin{array}{llll} P_0\begin{pmatrix}1\\-1\end{pmatrix} & & & \\ P_1\begin{pmatrix}2\\2\end{pmatrix} & P_0^{(1)} = \frac{2}{3}P_0 + \frac{1}{3}P_1 = \begin{pmatrix}\frac{4}{3}\\0\end{pmatrix} & & \\ P_2\begin{pmatrix}-2\\1\end{pmatrix} & P_1^{(1)} = \frac{2}{3}P_1 + \frac{1}{3}P_2 = \begin{pmatrix}\frac{2}{3}\\\frac{5}{3}\end{pmatrix} & P_0^{(2)} = \begin{pmatrix}\frac{10}{9}\\\frac{5}{9}\end{pmatrix} & \\ P_3\begin{pmatrix}3\\0\end{pmatrix} & P_2^{(1)} = \frac{2}{3}P_2 + \frac{1}{3}P_3 = \begin{pmatrix}\frac{-1}{3}\\\frac{2}{3}\end{pmatrix} & P_1^{(2)} = \begin{pmatrix}\frac{3}{9}\\\frac{12}{9}\end{pmatrix} & P_0^{(3)} = \begin{pmatrix}\frac{23}{27}\\\frac{22}{27}\end{pmatrix}. \end{array}$$

Az alábbi ábrán a harmadfokú Bézier görbe látható, illetve a $t = \frac{1}{3}$-nak megfelelő $P_0^{(3)}(t)$ pont.



Komplexebb görbék esetében több pontra van szükségünk. Ebben az esetében magasabb fokú Bézier görbét szerkeszthetünk vagy- és ez a gyakoribb- összetett görbéket használunk. Ez abban áll hogy a pontokat csoportosítjuk, például négyesével ha köbös görbét szeretnénk szerkeszteni, majd minden csoportra megszerkesztünk egy görbét (természetesen ebben az esetben a $[0,1]$ intervallum átalakítható tetszőleges $[a,b]$ intervallumra). Ezeket összetett Bézier görbéknek nevezzük. Ezen görbék esetében figyelembe kell venni a P_{3n} összeilleszkedési pontokban a simasági fokot.

Ha a P_2, P_3, P_4 pontok kollineárisak G^1 mértani (geometriai) folytonosságról beszélünk. Ahhoz hogy a görbe C^1 osztályú legyen még szükséges hogy a $\overline{P_2P_3}$, illetve $\overline{P_3P_4}$ szakaszok hossza arányos legyen a két görbe értelmezési tartományával.

7.4. Bézier felületek

A Bézier felületek szerkesztésére -úgy mint a görbék esetében- két mód van, egy mértani szerkesztés, illetve analitikus képleten alapuló. A felület generálását a *tenzoriális szorzat* esetben fogjuk vizsgálni ami azt jelenti hogy az elemzés lebontható a két (fő)irányra. Ennek érdekében

a felületeket a következő formális definícióval értelmezzük: felületnek nevezzük a *térben mozgó és alakját változtató görbe pontjait.*

A mozgó görbét konstans, m-ed fokú Bézier görbének fogjuk tekinteni:

$$P^m(u) = \sum_{i=0}^{m} P_i b_i^m(u), \ u \in [0,1]. \tag{7.4.1}$$

Ezt a görbét minden pozícióban (minden u-ra) a kontroll pontok határozzák meg. Feltételezzük, hogy minden eredeti P_i kontroll pont úgyszintén egy Bézier görbén mozog, ezúttal egy n-ed fokún:

$$P_i = P_i(v) = \sum_{j=0}^{n} P_{ij} b_j^n(v), \ v \in [0,1]. \tag{7.4.2}$$

Összekombinálva a két képletet megkapjuk a $P^{m,n}$ felület képletét az (u,v) paraméternek megfelelően:

$$P^{m,n}(u,v) = \sum_{i=0}^{m}\sum_{j=0}^{n} P_{ij} b_j^n(v)\, b_i^m(u), \ (u,v) \in [0,1]\times[0,1]. \tag{7.4.3}$$

A (7.4.3) képlet mátrix alakja a következő:

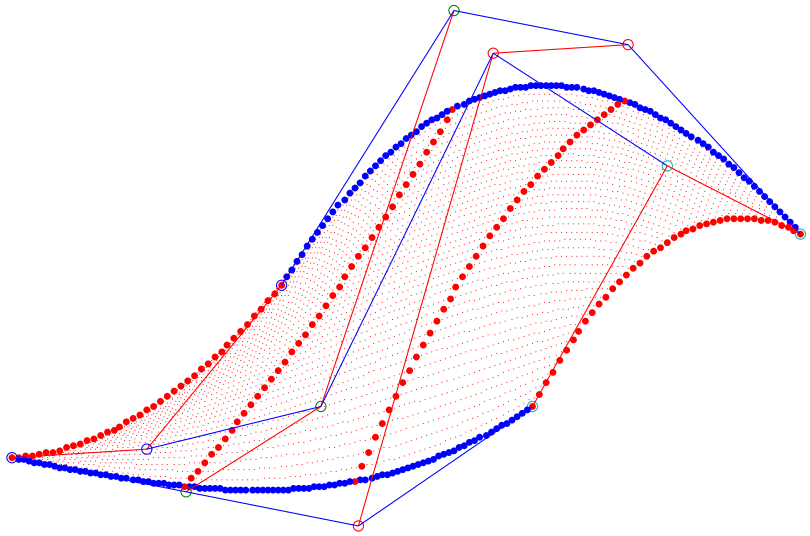
(7.4.4)

$$P^{m,n}(u,v) = \begin{pmatrix} b_0^m(u) & \dots & b_m^m(u) \end{pmatrix} \begin{pmatrix} P_{00} & \dots & P_{0n} \\ \vdots & & \vdots \\ P_{m0} & \dots & P_{mn} \end{pmatrix} \begin{pmatrix} b_0^n(v) \\ \vdots \\ b_n^n(v) \end{pmatrix}.$$

Konkrétan, ha egy $(2,3)$ fokú (másodfokú az egyik irányban, illetve harmadfokú a másik irányban) Bézier felületet szeretnénk szerkeszteni akkor a $P_{00}, P_{01}, P_{02}, P_{03},\ P_{10}, P_{11}, P_{12}, P_{13},\ P_{20}, P_{21}, P_{22}, P_{23}$ kontroll pontok segítségével meghatározzuk a felület analitikus képletét

$$P^{2,3}(u,v) = \sum_{i=0}^{2}\sum_{j=0}^{3} P_{ij} b_i^2(v)\, b_j^3(u), \ (u,v) \in [0,1]\times[0,1].$$

Az alábbi ábrán a $(2,3)$ fokú felület látható:



Rögzített v-re a felület u-tól függő, m-ed fokú Bézier görbe lesz. Ezt nevezik v(=konstans)-nek megfelelő izoparametrikus görbének.

Amellett hogy aránylag egyszerű képleteket eredményez, a tenzoriális szorzat előnye hogy a felület vizsgálata akármelyik iránnyal kezdhető.

138. Példa. Adottak az alábbi pontok: $P_{00}\begin{pmatrix} 0 & 0 & 0 \end{pmatrix}, P_{01}\begin{pmatrix} 2 & 0 & 0 \end{pmatrix}, P_{02}\begin{pmatrix} 4 & 0 & 0 \end{pmatrix}$,
$P_{10}\begin{pmatrix} 0 & 2 & 0 \end{pmatrix}, P_{11}\begin{pmatrix} 2 & 2 & 0 \end{pmatrix}, P_{12}\begin{pmatrix} 4 & 2 & 2 \end{pmatrix}$,
$P_{20}\begin{pmatrix} 0 & 4 & 0 \end{pmatrix}, P_{21}\begin{pmatrix} 2 & 4 & 4 \end{pmatrix}, P_{22}\begin{pmatrix} 4 & 4 & 4 \end{pmatrix}$. Számítsuk ki az $(u, v) =$

$\left(\frac{1}{2},\frac{1}{2}\right)$ paraméternek megfelelő pontot a $(2,2)$ bikvadratikus Bézier felületen. A (7.4.4) képletet felhasználva

$$P^{m,n}\left(\frac{1}{2},\frac{1}{2}\right) = \begin{pmatrix} (1-u)^2 & 2u(1-u) & u^2 \end{pmatrix} \begin{pmatrix} P_{00} & P_{01} & P_{02} \\ P_{10} & P_{11} & P_{12} \\ P_{20} & P_{21} & P_{22} \end{pmatrix} \begin{pmatrix} (1-v)^2 \\ 2v(1-v) \\ v^2 \end{pmatrix}$$

$$= \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}$$

ahol $(u,v) = \left(\frac{1}{2},\frac{1}{2}\right)$.

8. FEJEZET

Numerikus deriválás, numerikus integrálás

8.1. Numerikus deriválás

Tekintsük az $f : I \to \mathbb{R}$ folytonos és deriválható függvényt. Legyen $x_0 \in \mathring{I}$ egy pont az I belsejében. Értelmezés szerint a derivált x_0-ban

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

vagy $h = x - x_0$ jelöléssel:

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

Numerikusan a derivált értékét a jobboldalon szereplő törttel közelítjük meg:

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h}, \tag{8.1.1}$$

ahol h egy nullához közeli érték: $h = 10^{-3}, 10^{-4}$ (a továbbiakban $h > 0$); kerülni kell a túlzottan kicsi értéket mivel a nullával való osztás hibához vezethet.

139. Tétel. *Ha $f \in C^2(I)$ akkor a (8.1.1) képletben a következő hibabecslést adhatjuk:*

$$\left| f'(x_0) - \frac{f(x_0 + h) - f(x_0)}{h} \right| \leq h \cdot \frac{M_2}{2}, \tag{8.1.2}$$

ahol $M_2 = \sup_{x \in I} |f''(x)|$.

Bizonyítás. A Taylor képletből:

$$f(x_0 + h) = f(x_0) + \frac{f'(x_0)}{1!} h + \frac{f''(\xi)}{2!} h^2, \ \xi \in (x_0, x_0 + h).$$

Átrendezve azt kapjuk hogy:

$$\frac{f(x_0+h)-f(x_0)}{h}-f'(x_0)=\frac{f''(\xi)}{2}h,$$

vagyis:

$$\left|f'(x_0)-\frac{f(x_0+h)-f(x_0)}{h}\right|=h\left|\frac{f''(\xi)}{2}\right|\leq h\frac{M_2}{2}.$$

□

Ha a derivált kiszámításához adva van egy bizonyos $\epsilon>0$ pontosság, a h értékét a következő kikötésből számítjuk ki:

$$h\frac{M_2}{2}<\epsilon \implies h<\frac{2\epsilon}{M_2}.$$

Mivel az M_2 értéknek a kiszámításához nehézségekbe ütközhetünk, ezért a h érték numerikus meghatározása a következő módon történhet: a derivált kezdő értékét egy aránylag kis $h=10^{-3}$ értékkel számítjuk majd csökkentjük h-t (felezés, ...) ameddig két egymásutáni derivált (abszolút vagy relatív) különbsége kisebb mint ϵ. Ugyanakkor meg kell jegyezni hogy az említett kilépési kritérium nem garantálja azt hogy az elméleti derivált értéket ϵ-nyira közelítettük meg.

A (8.1.2)-es képletben szereplő hiba rendjét a következőképpen jelöljük:

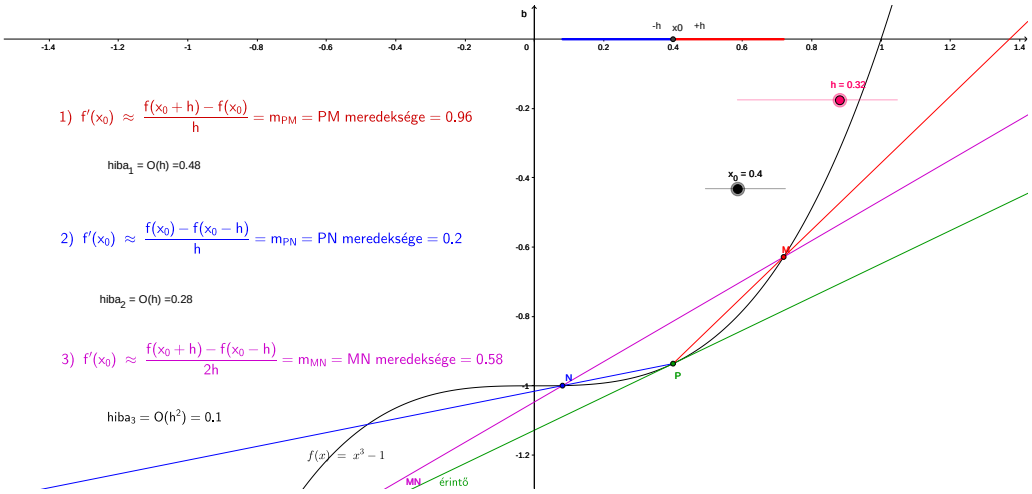
$$\left|f'(x_0)-\frac{f(x_0+h)-f(x_0)}{h}\right|=O(h), \tag{8.1.3}$$

vagyis az abszolút hiba lineárisan arányos h-val.

Mértanilag a $P(x_0,f(x_0))$ pontban húzott e érintő iránytényezője (lásd alábbi ábra), megközelíthető a PM, PN, MN húrok iránytényezőivel. A (8.1.1) képlet jobboldalán a PM húr iránytényezője szerepel. Minél kisebb a h lépés annál közelebb kerül az M pont P-hez és annál jobban közelíti az MN húr iránytényezője az e tangens iránytényezőjét.

Hasonló eset áll fenn a PN, illetve MN húrok esetében és a következő képleteket kapjuk:

$$f'(x_0)=\frac{f(x_0)-f(x_0-h)}{h}, \tag{8.1.4}$$



8.1.1. ábra. Derivált közelítése

vagy:

$$f'(x_0) = \frac{f(x_0+h) - f(x_0-h)}{2h}. \tag{8.1.5}$$

A (8.1.4) képletben a közelítés hibarendje lineáris:

$$\left| f'(x_0) - \frac{f(x_0) - f(x_0-h)}{h} \right| = O(h), \tag{8.1.6}$$

míg a (8.1.5) képletben négyzetes:

$$\left| f'(x_0) - \frac{f(x_0+h) - f(x_0-h)}{2h} \right| = O\left(h^2\right). \tag{8.1.7}$$

Ez utóbbi igazolható a Taylor képlet kétszeri alkalmazásával:

$$\begin{aligned} f(x_0+h) &= f(x_0) + \frac{f'(x_0)}{1!}h + \frac{f''(x_0)}{2!}h^2 + \frac{f'''(\xi_1)}{3!}h^3, \ \xi_1 \in (x_0, x_0+h) \\ f(x_0-h) &= f(x_0) - \frac{f'(x_0)}{1!}h + \frac{f''(x_0)}{2!}h^2 - \frac{f'''(\xi_2)}{3!}h^3, \ \xi_2 \in (x_0-h, x_0) \end{aligned}$$

majd az egyenletek különbségéből:

$$f(x_0+h) - f(x_0-h) = 2f'(x_0)h + \left(f'''(\xi_1) + f'''(\xi_2)\right)h^3/6,$$

következik, hogy:

$$\left| f'(x_0) - \frac{f(x_0+h) - f(x_0-h)}{2h} \right| = \frac{\left|f'''(\xi_1) + f'''(\xi_2)\right|}{12}h^2 = O\left(h^2\right).$$

Természetesen minél több tagot veszünk figyelembe a Taylor képletben, annál pontosabb lesz a derivált értéke.

A másodrendű derivált kiszámításához kétszer alkalmazzuk az elsőrendű derivált képletét, például a (8.1.1) kétszeri alkalmazása a következő képletet eredményezi:

$$\begin{aligned} f''(x_0) &= \frac{f'(x_0+h)-f'(x_0)}{h} = \frac{\frac{f(x_0+2h)-f(x_0+h)}{h} - \frac{f(x_0+h)-f(x_0)}{h}}{h} \\ f''(x_0) &= \frac{f(x_0+2h)-2f(x_0+h)+f(x_0)}{h^2}, \end{aligned}$$

ha pedig kombináltan a (8.1.1) és a (8.1.5) képletet, akkor:

$$f''(x_0) = \frac{f(x_0+h)-2f(x_0)+f(x_0-h)}{h^2}, \tag{8.1.8}$$

aminek a hibarendje négyzetes:

$$\left| f''(x_0) - \frac{f(x_0+h)-2f(x_0)+f(x_0-h)}{h^2} \right| = O\left(h^2\right).$$

140. Példa. Ha $f(x) = e^x$ számítsuk ki $f'(1)$, $f''(1)$ értékeket különböző h lépéseket használva.

Legyen $h = 10^{-2}$, a (8.1.1) képletből

$$f'(1) = \frac{f(1+h)-f(1)}{h} = \frac{e^{1.01}-e}{0.01} = 2.7319,$$

ebben az esetben a hiba

$$\left| \frac{f(1+h)-f(1)}{h} - e \right| = 2.7319 - 2.7183 = 0.0136 = 1.36 \cdot h.$$

A (8.1.5) képletet használva

$$f'(1) = \frac{f(1+h)-f(1-h)}{2h} = \frac{e^{1.01}-e^{0,99}}{0.02} = 2.718327$$

a hiba pedig

$$\left| \frac{f(1+h)-f(1-h)}{2h} - e \right| = 2.718327 - 2.718\,281 = 4.6\times10^{-5} = 0.46\times10^{-4}.$$

A másodfokú derivált értéke:

$$f''(1) = \frac{f(1+h)-2f(1)+f(1-h)}{h^2} = \frac{e^{1.01}-e+e^{0,99}}{10^{-4}} = 2.718304$$

a hiba

$$\left|\frac{f(1+h)-2f(1)+f(1-h)}{h^2}-e\right|=2.26\times10^{-5}=0.226\times10^{-4}.$$

Több ismeretlenes függvények esetében $f:\mathbb{R}^m\to\mathbb{R}$ a parciális deriváltak kiszámítása hasonlóan történik a már ismertetett módszerekkel azzal a megjegyzéssel hogy minden változót konstansnak tekintünk kivéve azt a változót ami szerint történik a deriválás. A (8.1.1),(8.1.5),(8.1.8) képleteknek megfelelően a parciális deriváltak a következőképpen alakulnak:

(8.1.9)
$$\frac{\partial f}{\partial x_i}(x_1,...,x_m)=\frac{f(x_1,...,x_i+h,...,x_m)-f(x_1,...,x_i,...,x_m)}{h}$$

(8.1.10)
$$\frac{\partial f}{\partial x_i}(x_1,...,x_m)=\frac{f(x_1,...,x_i+h,...,x_m)-f(x_1,...,x_i-h,...,x_m)}{2h}$$

$$\frac{\partial^2 f}{\partial x_i^2}(x_1,...,x_m)=\frac{1}{h^2}(f(x_1,...,x_i+h,...,x_m)-2f(x_1,...,x_i,...,x_m)+$$

(8.1.11)
$$+f(x_1,...,x_i-h,...,x_m)).$$

$$\frac{\partial^2 f}{\partial x_i\partial x_j}(x_1,...,x_m)=\frac{1}{4hk}(f(x_1,...,x_i+h,...,x_j+k,...,x_m)-f(x_1,...,x_i+h,...,x_j-k,...,x_m)$$

(8.1.12)
$$-f(x_1,...,x_i-h,...,x_j+k,...,x_m)+f(x_1,...,x_i-h,...,x_j-k,...,x_m)).$$

ahol h illetve k az x_i illetve x_j irányban megtett lépés.

141. Példa. Számítsuk ki a $f(x,y)=e^{xy}$ függvény parciális deriváltjait $(1,2)$ pontban.

Legyen $h=k=10^{-2}$. A (8.1.10) képletnek megfelelően:

$$\frac{\partial f}{\partial x}(1,2)=\frac{f(1+h,2)-f(1-h,2)}{2h}=14.7791$$

és a hiba

$$\left|\frac{\partial f}{\partial x}(1,2)-2e^2\right|=\left|\frac{\partial f}{\partial x}(1,2)-14.7781\right|=9.85\times10^{-4};$$

A függvény y szerinti parciális deriált:

$$\frac{\partial f}{\partial y}(1,2) = \frac{f(1,2+k) - f(1,2-k)}{2k} = 7.3892$$

és a hiba

$$\left|\frac{\partial f}{\partial y}(1,2) - e^2\right| = \left|\frac{\partial f}{\partial y}(1,2) - 7.3891\right| = 1.23 \times 10^{-4}.$$

8.2. Numerikus integrálás (kvadratúra képletek)

Általános értelemben kvadratúra határozott integrál kiszámítását jelenti.

Ha ismert az $f : [a,b] \to \mathbb{R}$ folytonos függvény és az $[a,b]$ intervallum egy felosztása:

$$a = x_0 < x_1 < ... < x_n = b,$$

akkor a

$$I = \int_a^b f(x)\,\mathrm{d}x,$$

határozott integrál kiszámítható a Riemann-féle összeggel:

$$I = \lim_n \sigma_n,$$

ahol

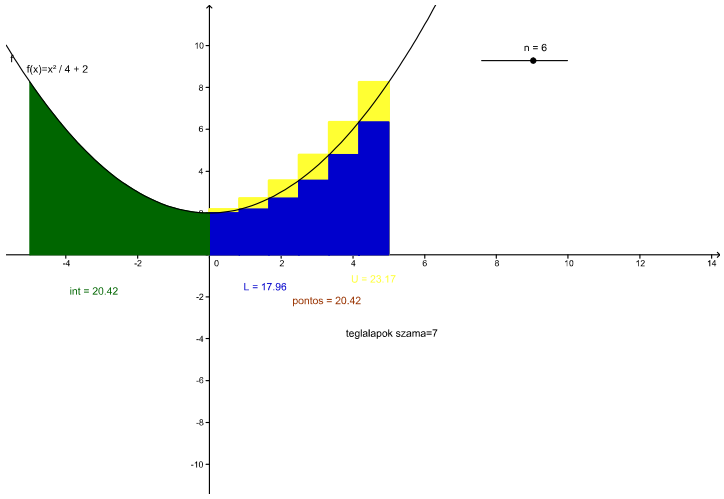
$$\sigma_n = \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(\xi_i),\ \xi_i \in [x, x_{i+1}].$$

142. Definíció. Numerikus kvadratúrán a

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^{n-1} A_i \cdot f(x_i) + R,$$

képletet értjük, ahol az A_i együtthatók függetlenek $f(x)$ függvénytől, x_i a csomópontok, R pedig az elkövetett hiba.

8.2.0.1. Téglalap módszer Tételezzük fel hogy az x_i csomópontok ekvidisztánsak: $x_i = x_0 + ih,\ h = \frac{b-a}{n}$. Akkor az $\int_a^b f(x)\,\mathrm{d}x$ határozott integrál-ami nem más mint az f függvény és az Ox tengely által bezárt terület a-tól b-ig megközelíthető a téglalapok összterületével.



8.2.1. ábra. Numerikus integrálás: téglalap módszer

$$\int_a^b f(x)\,\mathrm{d}x = h\left(f(x_0)+\ldots+f(x_{n-1})\right), \tag{8.2.1}$$

$$\int_a^b f(x)\,\mathrm{d}x = h\left(f(x_1)+\ldots+f(x_n)\right), \tag{8.2.2}$$

vagy ha ismertek a $f\left(\frac{x_i+x_{i+1}}{2}\right)$ értékek akkor:

$$\int_a^b f(x)\,\mathrm{d}x = h\left(f\left(\frac{x_0+x_1}{2}\right)+\ldots+f\left(\frac{x_{n-1}+x_n}{2}\right)\right). \tag{8.2.3}$$

143. TÉTEL. *Az (8.2.1), illetve (8.2.2) képlet esetében az abszolút hiba:*

$$\left|\int_a^b f(x)\,dx - h\left(f(x_0)+\ldots+f(x_{n-1})\right)\right| \le hM_1(b-a),$$

ahol $M_1 = \sup_{x\in[a,b]}|f'(x)|$.

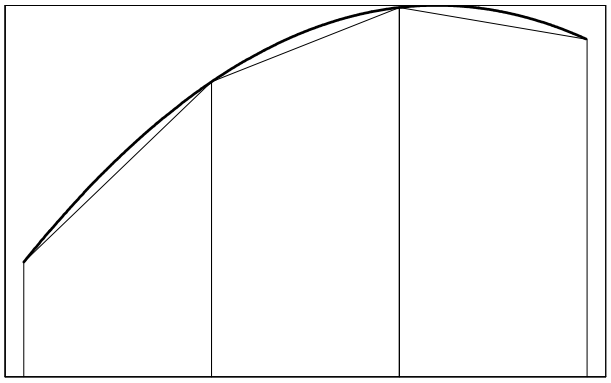
Az (8.2.1), illetve (8.2.2) képlet esetében a hibarend lineáris h-ra nézve:

$$\left|\int_a^b f(x)\,\mathrm{d}x - h\left(f(x_0)+\ldots+f(x_{n-1})\right)\right| = O(h),$$

vagy figyelembe véve hogy $h = \frac{b-a}{n}$ felírhatjuk hogy a hibarend $= O\left(n^{-1}\right)$ ahol n a téglalapok száma. A (8.2.3) képlet esetében a hiba négyzetes:

$$\left| \int_a^b f(x)\,\mathrm{d}x - h\left(f\left(\frac{x_0 + x_1}{2} \right) + ... + f\left(\frac{x_{n-1} + x_n}{2} \right) \right) \right| = O\left(h^2\right) = O\left(n^{-2}\right).$$

8.2.0.2. A trapéz módszer Egy aránylag egyszerű ugyanakkor a téglalap módszernél pontosabb módszer. Az x_i csomópontok úgyszintén ekvidisztánsak az $(x_i, f(x_i))_{i=0}^n$ pontokat pedig lineárisan kötjük össze:



8.2.2. ábra. Numerikus integrálás: trapéz módszer

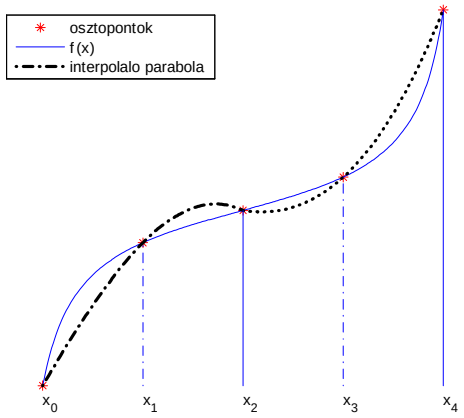
Az így keletkezet n trapéz összterülete megközelíti az integrál értékét:

$$\int_a^b f(x)\,\mathrm{d}x = \frac{h}{2}\left(f(x_0) + 2f(x_1) + ... + 2f(x_{n-1}) + f(x_n) \right), \tag{8.2.4}$$

a hiba pedig négyzetes:

$$\left| \int_a^b f(x)\,\mathrm{d}x - \frac{h}{2}\left(f(x_0) + 2f(x_1) + ... + 2f(x_{n-1}) + f(x_n) \right) \right| = O\left(n^{-2}\right).$$

8.2.0.3. A Simpson módszer Ebben az esetben az $(x_i, f(x_i))_{i=0}^{2n}$ pontokat parabolákkal (hármasával) interpoláljuk, ezért a csomópontok száma páratlan kell legyen, a lépés pedig $h = \frac{b-a}{2n}$.



8.2.3. ábra. Numerikus integrálás: Simpson módszer

Az integrál értéke:

(8.2.5)
$$\int_a^b f(x)\,\mathrm{d}x = \frac{h}{3}\left(f(x_0) + 4\sum_{i=1}^{n} f(x_{2i-1}) + 2\sum_{i=1}^{n-1} f(x_{2i}) + f(x_{2n})\right)$$

a hiba pedig 4-ed rendű:

$$\left|\int_a^b f(x)\,\mathrm{d}x - \frac{h}{3}\left(f(x_0) + 4\sum_{i=1}^{n} f(x_{2i-1}) + 2\sum_{i=1}^{n-1} f(x_{2i}) + f(x_{2n})\right)\right| = O\left(n^{-4}\right).$$

Bizonyítás. Az $[x_0, x_2]$ intervallumban az $(x_0, f(x_0)), (x_1, f(x_1)), (x_2, f(x_2))$, pontokat interpoláló parabola kifejezhető a Lagrange interpoláló függvénnyel:

$$P_1(x) = f(x_0)\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + f(x_1)\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + f(x_2)\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}.$$

Az alatta elterülő terület:

$$\begin{aligned}\int_{x_0}^{x_2} P_1(x)\,\mathrm{d}x &= \frac{f(x_0)}{(x_0-x_1)(x_0-x_2)}\int_{x_0}^{x_2}(x-x_1)(x-x_2)\,\mathrm{d}x + \\ &+\frac{f(x_1)}{(x_1-x_0)(x_1-x_2)}\int_{x_0}^{x_2}(x-x_0)(x-x_2)\,\mathrm{d}x \\ &+\frac{f(x_2)}{(x_2-x_0)(x_2-x_1)}\int_{x_0}^{x_2}(x-x_0)(x-x_1)\,\mathrm{d}x = \\ &= \frac{f(x_0)}{(x_0-x_1)(x_0-x_2)}2\frac{h^3}{3} - \frac{f(x_1)}{(x_1-x_0)(x_1-x_2)}4\frac{h^3}{3} + \frac{f(x_2)}{(x_2-x_0)(x_2-x_1)}2\frac{h^3}{3} = \\ &= \frac{h}{3}\left[f(x_0)+4f(x_1)+f(x_2)\right].\end{aligned}$$

Hasonlóan kifejezhetők a többi parabola alatti területek, majd ezeknek az összege

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \int_{x_0}^{x_2} P_1(x)\,\mathrm{d}x + \int_{x_2}^{x_4} P_2(x)\,\mathrm{d}x + \ldots + \int_{x_{2n-2}}^{x_{2n}} P_n(x)\,\mathrm{d}x = \\ &= \frac{h}{3}\left[f(x_0)+4f(x_1)+f(x_2)\right] + \frac{h}{3}\left[f(x_2)+4f(x_3)+f(x_4)\right] + \ldots \\ &= \frac{h}{3}\left[f(x_0)+4f(x_1)+2f(x_2)+4f(x_3)+2f(x_4)+\ldots+f(x_{2n})\right].\end{aligned}$$

□

Habár a Simpson algoritmus komplexitása csak egy fokkal nő, a trapéz módszerhez képest a hibarend négyzetes rendről 4-ed rendűre ugrik. Emiatt ez az eljárás nagyon gyakran használt.

144. Példa. Számítsuk ki $\int_1^3 x^2\mathrm{d}x$ értéket a trapéz illetve Simpson képlet segítségével ($n=4$).

$f(x)=x^2$. A megadott n-ből kiszámítjuk a trapéz módszernek megfelelő lépést:$h=\frac{3-1}{4}=0.5$ majd a (8.2.4) képletből:

$$\begin{aligned}\int_1^3 x^2\mathrm{d}x &= \frac{h}{2}\left(f(1)+2f(1+0.5)+2f(1.5+0.5)+2f(2+0.5)+f(3)\right) = \\ &= 0.25\left(1+4.5+8+12.5+9\right) = 8.75.\end{aligned}$$

A hiba $\left|8.75 - \frac{x^3}{3}\,|_1^3\right| = |8.75 - 8.66| = 0.09\,(O\,(n^{-2}) = 1/16 = 0.0625)$.
Hasonlóan a Simpson módszerből $h = \frac{3-1}{2\cdot 4} = 0.25$ és a (8.2.5) képletből:

$$\int_1^3 x^2 \mathrm{d}x =$$

$$\frac{h}{3}\,(f\,(1) + 4(f\,(1.25) + f\,(1.75) + f\,(2.25) + f\,(2.75)) + 2\,(f\,(1.5) + f\,(2) + f\,(2.5)) + f\,(3))$$

$$= \frac{0.25}{3}\,(1 + 4\,(1.5625 + 3.0625 + 5.0625 + 7.5625) + 2\,(2.25 + 4 + 6.25) + 9) = 8.6666.$$

Mivel a függvény másodfokú polinom az integrál nulla hibával állítható elő a Simpson képlettel.

Az említett képletek interpolációs sémákból erednek és Newton-Cotes néven ismertek. Az alábbi táblázat összehasonlítja a interpoláló polinom fokát illetve a hibarendet:

Módszer	Interp. pol. fokszáma	Hibarend
Téglalap	0	$O\,(h)$
Trapéz	1	$O\,(h^2)$
Simpson	2	$O\,(h^4)$

.

9. FEJEZET

Differenciálegyenletek numerikus megoldása

9.1. Elsőrendű differenciálegyenletek

A gyakorlatban gyakran szükséges olyan függvények meghatározása (jel. y) amelyeknek csak a változását ismerjük. Mivel a változás általában időben történik, független változóként t-t használunk.

145. Példa. Tanulmányozzuk egy csésze kávé lehűlését egy negyed órán keresztül $[t_0, t_f] = [0, 15]$, ha kezdetben a kávé hőmérséklete $y_0 = 100°C$ fok volt. Ha ismert a csésze hővezetési tényezője $k = \frac{1}{20} = 0.05$, illetve a terem hőmérséklete $y_{környezet} = 20°C$ fok, határozzuk meg hány fokos lesz a kávé a folyamat végén?

Megoldás. Newton-féle lehűlési törvény szerint a lehűlés sebessége ($\frac{dy}{dt} = \dot{y}$) arányos a hővezetési tényezővel, illetve a környezet és a kávé hőmérsékletének különbségével:

$$\dot{y}(t) = k \cdot (y_{környezet} - y(t)).$$

Ugyanakkor, ismert a kezdeti feltétel:

$$y(t_0) = y_0.$$

A feladat analitikus megoldása

$$y(t) = y_{környezet} + (y_0 - y_{környezet})e^{-kt},\ t \in [t_0, t_f],$$

tehát a folyamat végén a kávé hőmérséklete $y(t_f) \approx 57.789$. Az analitikus megoldástól eltérően, a numerikus megoldás csak diszkrét közelítést ad.

Elsőrendű differenciálegyenletnek nevezzük a

$$\frac{dy}{dt}(t) = f(t, y(t)),$$

vagy egyszerűsítve

$$\dot{y} = f(t, y).$$

típusú egyenletet, ahol t a független változó, f pedig egy előre megadott függvény.

A differenciálegyenlet

$$y = y(t)$$

megoldását integrál görbének nevezzük. Ha f sajátos alakú akkor elemi úton is megoldható, de gyakran ez annyira bonyolult, hogy sokkal kényelmesebb közelítő módszerekkel egy partikuláris megoldás előállítása. Ugyanakkor bizonyos egyenleteket csak numerikusan lehet megoldani, például

$$\dot{y}(t) = e^{t^2}.$$

146. Példa. Az

$$\dot{y}(t) = 2 \cdot t \cdot y = f(t, y),\ t \in [t_0, t_f],$$

differenciálegyenlet megoldása

$$y(t) = ce^{t^2}\ (c = \text{konstans}).$$

Az elsőrendű közönséges differenciálegyenleteknek végtelen sok megoldása van, amelyek egy konstansban különböznek egymástól. Ha egy plusz feltételt rendelünk a feladathoz, akkor a megoldás egyértelművé válik. Ezt nevezzük Cauchy-féle feladatnak:

$$\begin{cases} \dot{y}(t) = f(t, y), & t \in [t_0, t_f] \\ y(t_0) = y_0 \end{cases}. \tag{9.1.1}$$

Mértanilag ez azt jelenti, hogy a görbeseregből azt a görbét választjuk amelyik keresztül halad a $P_0(t_0, y_0)$ ponton.

147. Példa. Ha az előbbi feladathoz hozzárendeljük a $y(0) = 1$ kezdeti feltételt, akkor a c konstans egyenlő 1-gyel, vagyis a

$$\begin{cases} \dot{y}(t) = 2ty \\ y(0) = 1 \end{cases}$$

Cauchy feladatnak

$$y(t) = e^{t^2}$$

a megoldása.

A differenciálegyenletek numerikus megoldásának alapelve, hogy az integrál görbét pontszerűen közelítjük meg

$$y(t_i) = y_i,$$

ahol t_i a $[t_0, t_f] \ni t$ tanulmányozott intervallum egy felosztása.

A módszereket feloszthatjuk egylépéses, illetve többlépéses módszerekre, attól függően hogy az y_{i+1} ordináta kiszámításához egy vagy több előzetes ordinátát használunk. Az egy lépéses módszerek közé tartozik az Euler, Runge-Kutta, Taylor, fokozatos közelítések módszere, míg a prediktor-korrektor többlépéses módszer.

9.1.1. Az Euler-féle (törtvonal) módszer Tekintsük a Cauchy-féle feladatot:

$$\begin{cases} \dot{y}(t) = f(t, y) \quad , \ t \in [t_0, t_f] \\ y(t_0) = y_0 \end{cases} , \tag{9.1.2}$$

illetve a $[t_0, t_f]$ intervallum egy ekvidisztáns felosztását:

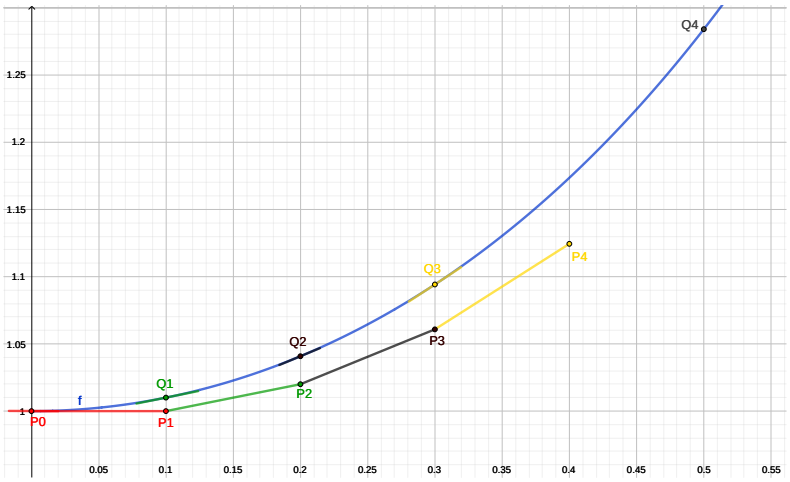
$$t_0 < t_1 < ... < t_n = t_f, \ t_{i+1} = t_i + h \tag{9.1.3}$$

ahol

$$h = \frac{t_f - t_0}{n} \tag{9.1.4}$$

a lépést jelöli.

Az Euler módszer lényege, hogy az elméleti görbét -pontról pontra haladva- lineáris szakaszokkal közelítjük meg és eredményül egy $P_0P_1...P_n$ törtvonalat kapunk; innen származik a módszer elnevezése.



9.1.1. ábra. Törtvonal módszer

Kiinduló pontként a $P_0(t_0, y_0)$ használjuk, majd rendre megszerkesztjük a $P_i(t_i, y_i)$, $i = \overline{1,n}$ pontokat.

Az y_1 ordináta kiszámításához figyelembe vesszük a t_0 időpontban ismert y_0 függvény értékét és feltételezzük hogy a $[t_0, t_1]$ intervallumban a függvény növekedése (meredeksége) konstans, vagyis $\dot{y}(t_0) = f(t_0, y_0)$:

$$m_{P_0P_1} = \frac{y_1 - y_0}{t_1 - t_0} = f(t_0, y_0) \Rightarrow y_1 = y_0 + h \cdot f(t_0, y_0). \tag{9.1.5}$$

Hasonlóan szerkesztjük meg az összes többi y_i ordinátát:

$$y_{i+1} = y_i + h \cdot f(t_i, y_i) \quad i = 1, ..., n-1. \tag{9.1.6}$$

Habár $h \to 0$ -ra a módszer konvergens, a konvergenciarend lineáris.

A (9.1.6) relációt a Taylor képletből kapjuk első két tagjára korlátozva:

$$y(t+h) = y(t) + \frac{1}{1!}\dot{y}(t) \cdot h + O\left(h^2\right), \tag{9.1.7}$$

majd $t = t_i$-re

$$\begin{aligned} y(t_{i+1}) &= y(t_i + h) = y(t_i) + \dot{y}(t_i) \cdot h + O\left(h^2\right) \Leftrightarrow \\ y_{i+1} &= y_i + h \cdot f(t_i, y_i) + O\left(h^2\right). \end{aligned}$$

Tehát minden iterációban az elkövetett hiba négyzetes $O\left(h^2\right)$, de n lépés után (ennyi lépés szükséges P_n kiszámításához) a hibarend lineárissá válik:

$$n \times O\left(h^2\right) = O\left(h^{-1}\right) \times O\left(h^2\right) = O\left(h\right).$$

148. Példa. Oldjuk meg a következő Cauchy-féle feladatot az Euler módszerrel ($h = 0.1$)

$$\begin{cases} \dot{y} = 2ty, \quad t \in [0,1] \\ y\left(0\right) = 1 \end{cases}.$$

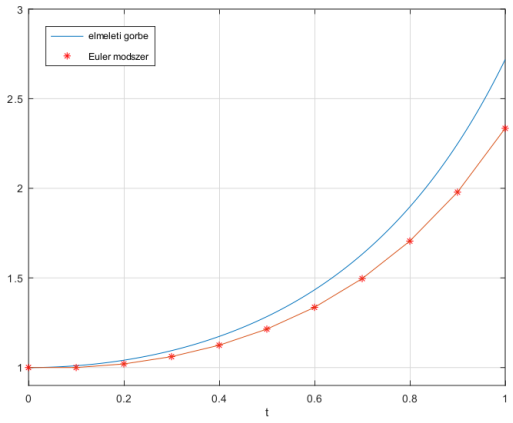
Bizonyítás. $f\left(t,y\right) = 2ty$, $t_0 = 0$, $y_0 = 1 \Rightarrow \ P_0\left(0,1\right)$.
$t_i = i * 0.1$,

$$y_1 = y_0 + hf\left(t_0, y_0\right) = y_0 + h\left(2t_0 \cdot y_0\right) = 1 + 0.1 \cdot 2 \cdot 0 \cdot 1 = 1$$

$\Longrightarrow P_1\left(0.1, 1\right)$

$$y_2 = y_1 + hf\left(t_1, y_1\right) = 1 + 0.1 \cdot 2 \cdot 0.1 \cdot 1 = 1.02$$

$\Longrightarrow P_2\left(0.2, 1.02\right), ...$



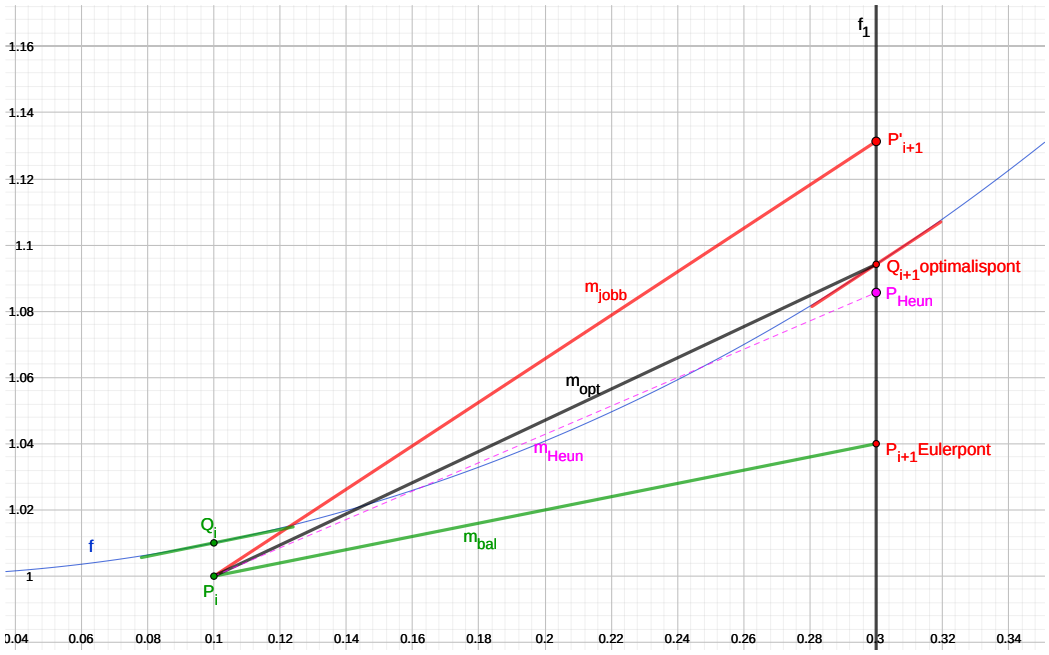
9.1.2. ábra. A pontos megoldás és az Euler közelítés

□

9.1.2. Heun módszer Az Euler módszerben a (t_i, y_i) pontból kiindulva a következő $(t_{i+1}y_{i+1})$ közelítő pontot állítottuk elő. Ezt figyelembe véve az Euler módszert prediktor (előre jelző) módszernek

nevezzük. A pont előállításához az (t_i, y_i) pontban-az y-hoz- húzott érintő meredekségét használja. A Heun módszer egy korrekciót végez az Euler eljáráson, ezáltal egy javított módszert kapunk; ezt prediktor-korrektor módszernek nevezzük.

Egy konvex függvény esetén (lásd alábbi ábrát) a baloldali (t_i, y_i) pontban húzott érintő meredeksége alul értékelt. Ugyanakkor – a (t_i, y_i) pontból kiindulva – de a (t_{i+1}, y_{i+1}) pontban (jobboldali pont) húzott meredekséggel haladva egy felül értékelt ordinátát kapunk (konkáv függvény esetén fordítva történik).



9.1.3. ábra. Heun korrekció

Az optimális meredekség a két érték között van, ezért a Heun korrekció abban áll, hogy az ajánlott meredekséget a baloldali és jobboldali meredekség átlagaként határozza meg:

$$m_{Heun} = \frac{1}{2}\left(m_{bal} + m_{jobb}\right), \tag{9.1.8}$$

ahol

$$m_{bal} = \dot{y}(t_i) = f(t_i, y_i), \tag{9.1.9}$$

$$m_{jobb} = \dot{y}(t_{i+1}) = f(t_{i+1}, y_{i+1}) \tag{9.1.10}$$

A (9.1.10) képletben az y_{i+1} érték nem áll rendelkezésünkre, ezért ezt a (9.1.6) Euler predikcióból számítjuk ki:

$$y_{i+1} = y_i + h \cdot f(t_i, y_i),$$

tehát a Heun módszer

(9.1.11)

$$y_{i+1} = y_i + h \cdot m_{Heun} = y_i + h \cdot \frac{1}{2}\left(f(t_i, y_i) + f(t_{i+1}, y_{i+1})\right) =$$
$$= y_i + h \cdot \frac{1}{2}\left(f(t_i, y_i) + f(t_{i+1}, y_i + h \cdot f(t_i, y_i))\right).$$

Bevezetve a

$$k_1 = h \cdot f(t_i, y_i),$$
$$k_2 = h \cdot f(t_i + h, y_i + k_1),$$

jelöléseket az Euler, illetve Heun módszerben az ordinátákat a következő képletekkel számítható ki:

$$y_{i+1} = y_i + k_1, \text{ Euler}$$
$$y_{i+1} = y_i + \frac{1}{2}(k_1 + k_2), \text{ Heun}$$

A Heun módszer hibarendje négyzetes $O(h^2)$.

9.1.3. Taylor módszer Az ismertettet Euler és Heun módszer levezethető a Taylor sorfejtésből (lásd (9.1.7) képletet), sőt több tag figyelembe vételével pontosabb képletek is levezethetők. Az Euler módszert - mint lineáris tagú Taylor sorfejtés - pontosíthatjuk további tagok figyelembevételével, például négyzetes tagig:

$$y(t+h) = y(t) + \frac{1}{1!}h \cdot \dot{y}(t) + +\frac{1}{2!}h^2 \cdot \ddot{y}(t) + O\left(h^3\right). \tag{9.1.12}$$

Az $\ddot{y}(t)$ kiszámításához használjuk a (9.1.2) képletet:

(9.1.13)
$$\ddot{y}(t) = \frac{\partial}{\partial t}\dot{y}(t) = \frac{\partial f}{\partial t}(t,y) + \frac{\partial f}{\partial y}(t,y)\cdot\dot{y}(t) = \frac{\partial f}{\partial t}(t,y) + f(t,y)\frac{\partial f}{\partial y}(t,y).$$

Tehát

(9.1.14)
$$y_{i+1} = y_i + h\cdot f(t_i,y_i) + \frac{1}{2}h^2\left[\frac{\partial f}{\partial t}(t_i,y_i) + f(t_i,y_i)\frac{\partial f}{\partial y}(t_i,y_i)\right], i = 0,\ldots,n.$$

és a módszer hibarendje $n \times O(h^3) = O(h^2)$.

Az eljárás általánosítható viszont egyre komplikáltabb képleteket kapunk az y deriváltakra, például a harmadrendű derivált:

$$\dddot{y}(t) = \frac{\partial^2 f}{\partial t^2}(t,y) + 2\frac{\partial^2 f}{\partial t\partial y}(t,y)\cdot\dot{y}(t) + \frac{\partial^2 f}{\partial y^2}(t,y)\cdot(\dot{y}(t))^2 + \frac{\partial f}{\partial y}(t,y)\cdot\ddot{y}(t).$$

9.1.3.1. Téglalap módszer Az Euler módszer pontosságának javítására használjuk újból a Taylor sorfejtést integrál maradékkal:

$$y(t+h) = y(t) + \int_0^h \dot{y}(t+s)\,ds, \tag{9.1.15}$$

majd az integrálra különböző numerikus közelítést használunk.

Az integrál közelítésére használjuk a (8.2.3) (középponti) téglalap módszert:

$$\int_0^h \dot{y}(t+s)\,ds = h\dot{y}\left(t+\frac{h}{2}\right).$$

A (9.1.2)-ből

$$\dot{y}\left(t+\frac{h}{2}\right) = f\left(t+\frac{h}{2}, y\left(t+\frac{h}{2}\right)\right),$$

és figyelembe véve a Taylor képletet:

$$y\left(t+\frac{h}{2}\right) = y(t) + \frac{h}{2}\dot{y}(t) = y(t) + \frac{h}{2}f(t,y(t)),$$

következik, hogy

$$\dot{y}\left(t+\frac{h}{2}\right) = f\left(t+\frac{h}{2}, y(t) + \frac{h}{2}f(t,y(t))\right).$$

Visszahelyettesítve a (9.1.15) képletbe a kapott integrált azt kapjuk, hogy:

$$y(t+h) = y(t) + hf\left(t + \frac{h}{2}, y(t) + \frac{h}{2} f(t, y(t))\right), \tag{9.1.16}$$

vagyis $t = t_i$ -re

$$y(t_{i+1}) = y(t_i) + hf\left(t_i + \frac{h}{2}, y(t_i) + \frac{h}{2} f(t_i, y(t_i))\right), \tag{9.1.17}$$

ami $y(t_i) = y_i$ közelítést használva a következő képlethez vezet:

$$y_{i+1} = y_i + hf\left(t_i + \frac{h}{2}, y_i + \frac{h}{2} f(t_i, y_i)\right),\ i = \overline{0, n-1}. \tag{9.1.18}$$

9.1.3.2. Trapéz módszer (Heun módszer) A trapéz módszert használva a (9.1.15) képletben szereplő integrál kiszámításához :

$$\int_0^h \dot{y}(t+s)\, ds = \frac{h}{2}\left(\dot{y}(t) + \dot{y}(t+h)\right),$$

azt kapjuk, hogy:

$$\begin{aligned} y(t+h) &= y(t) + \frac{h}{2}\left[\dot{y}(t) + \dot{y}(t+h)\right] = y(t) + \frac{h}{2}\left[f(t, y(t)) + f(t+h, y(t+h))\right] = \\ &= y(t) + \frac{h}{2}\left[f(t, y(t)) + f(t+h, y(t) + hf(t, y(t)))\right], \end{aligned}$$

és $t = t_i$-re

$$y(t_{i+1}) = y(t_i) + \frac{h}{2}\left[f(t_i, y(t_i)) + f(t_i + h, y(t_i) + hf(t_i, y(t_i)))\right], \tag{9.1.19}$$

vagyis

$$y_{i+1} = y_i + \frac{h}{2}\left[f(t_i, y_i) + f(t_i + h, y_i + hf(t_i, y_i))\right],\ i = \overline{0, n-1}. \tag{9.1.20}$$

A (9.1.20) képlet átírható

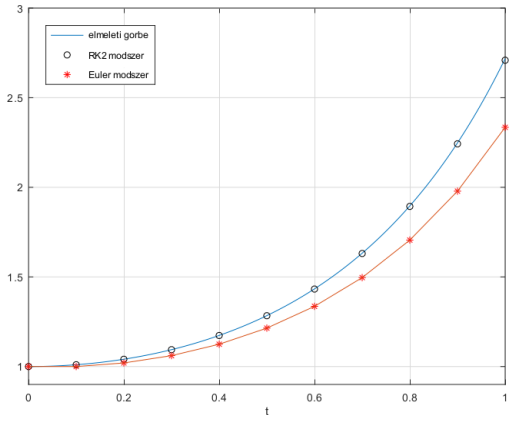
$$y_{i+1} = y_i + \frac{1}{2}\left[k_1 + k_2\right],\ i = \overline{0, n-1} \tag{9.1.21}$$

alakra, ahol:

$$\begin{aligned} k_1 &= h \cdot f(t_i, y_i), \\ k_2 &= h \cdot f(t_i + h, y_i + k_1), \end{aligned}$$

és Heun vagy Runge-Kutta2 néven ismert.

A két javított módszer: (9.1.18) és (9.1.20) hibarendje $O\left(h^2\right)$.



9.1.4. ábra. Euler vs. RK2 módszer

9.1.3.3. Runge-Kutta4 módszer Az eddigi ismertetett módszereknél igényesebb, ám pontosabb az úgy nevezett Runge-Kutta4 módszer:

$$y_{i+1} = y_i + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right), \tag{9.1.22}$$

ahol

$$\begin{aligned} k_1 &= h \cdot f\left(t_i, y_i\right), \; k_2 = h \cdot f\left(t_i + \frac{h}{2}, y_i + \frac{k_1}{2}\right), \\ k_3 &= h \cdot f\left(t_i + \frac{h}{2}, y_i + \frac{k_2}{2}\right), \; k_4 = h \cdot f\left(t_i + h, y_i + k_3\right). \end{aligned}$$

Minden részintervallumon a hiba $O\left(h^5\right)$, tehát n lépés után a hibarend:

$$n \times O\left(h^5\right) = O\left(h^{-1}\right) \times O\left(h^5\right) = O\left(h^4\right).$$

149. Példa. Az alábbi táblázatban a Példa 146 különböző módszereknek megfelelő összehasonlító adatai szerepelnek (négy tizedes pontossággal):

$t_i \diagdown y_i$	$(Euler)$	$(RK2)$	$(RK4)$	$(pontos:\ e^{t^2})$
0	1.0000	1.0000	1.0000	1.0000
0.1	1.0000	1.0100	1.0101	1.0101
0.2	1.0200	1.0407	1.0408	1.0408
0.3	1.0608	1.0940	1.0942	1.0942
0.4	1.1244	1.1732	1.1735	1.1735
0.5	1.2144	1.2835	1.2840	1.2840
0.6	1.3358	1.4324	1.4333	1.4333
0.7	1.4961	1.6306	1.6323	1.6323
0.8	1.7056	1.8934	1.8965	1.8965
0.9	1.9785	2.2426	2.2479	2.2479
1	2.3346	2.7091	2.7183	2.7183

1. táblázat. Módszerek közötti eltérés

9.1.4. Fokozatos közelítések módszere

$$y_{i+1}(t) = Fy_i(t),$$

ahol

$$Fy_i(t) = \int f(t, y(s))\, ds...$$

9.2. Elsőrendű, differenciál egyenletrendszerek

A differenciál egyenletrendszerek tanulmányozása több szempontból is fontos. Egyfelől bizonyos gyakorlati folyamatok leírásánál jutunk ilyen matematikai modellhez, másfelől az elsőrendűnél magasabb differenciálegyenletek visszavezethetők differenciál egyenletrendszerekre.

Tekintsük az alábbi, $x = x(t)$ és $y = y(t)$ függvényeket tartalmazó differenciál egyenletrendszert:

$$\begin{cases} \dot{x}(t) = f(t, x, y) \\ \dot{y}(t) = g(t, x, y) \end{cases}, \ t \in [t_0, t_f]. \tag{9.2.1}$$

Az adott egyenleteken kívül a függvények eleget tesznek az alábbi kezdeti feltételeknek:

$$\begin{cases} x(t_0) = x_0 \\ y(t_0) = y_0 \end{cases}. \tag{9.2.2}$$

Az x, y függvények megszerkesztéséhez bármelyik ismert módszerhez folyamodhatunk: Euler, Runge-Kutta, Taylor, stb.

Az alábbiakban az Euler módszert mutatjuk be.

A tanulmányozott intervallumot felosztjuk egyenközű osztópontokra: $t_0 < t_1 < ... < t_n = t_f$, $t_{i+1} - t_i = h$.

A (9.2.1) és (9.2.2) feltételekből következik, hogy:

$$\begin{aligned} \dot{x}(t_0) &= f(t_0, x(t_0), y(t_0)) = f(t_0, x_0, y_0) \\ \dot{y}(t_0) &= g(t_0, x(t_0), y(t_0)) = g(t_0, x_0, y_0). \end{aligned}$$

Ezekben az osztópontokban megszerkesztjük az x illetve y függvény közelítő értékét $x_i = x(t_i)$, $y_i = y(t_i)$:

$$\begin{aligned} \frac{x_1 - x_0}{t_1 - t_0} &= \dot{x}(t_0) = f(t_0, x_0, y_0), \\ \frac{y_1 - y_0}{t_1 - t_0} &= \dot{y}(t_0) = g(t_0, x_0, y_0), \end{aligned}$$

ahonnan:

$$\begin{aligned} x_1 &= x_0 + h \cdot f(t_0, x_0, y_0), \\ y_1 &= y_0 + h \cdot g(t_0, x_0, y_0). \end{aligned}$$

Hasonlóan kapjuk általánosan:

$$x_{i+1} = x_i + h \cdot f(t_i, x_i, y_i), \tag{9.2.3}$$

$$y_{i+1} = y_i + h \cdot g(t_i, x_i, y_i), \ i = \overline{1, n-1}. \tag{9.2.4}$$

A Heun módszer esetén az iteráció a következő:

$$x_{i+1} = x_i + \tfrac{1}{2}[k_1 + k_2], \ i = \overline{0, n-1} \tag{9.2.5}$$

$$y_{i+1} = y_i + \tfrac{1}{2}[l_1 + l_2], \ i = \overline{0, n-1} \tag{9.2.6}$$

ahol

$$\begin{aligned} k_1 &= h \cdot f(t_i, x_i, y_i), & l_1 &= h \cdot g(t_i, x_i, y_i), \\ k_2 &= h \cdot f(t_i + h, x_i + k_1, y_i + l_1), & l_2 &= h \cdot g(t_i + h, x_i + k_1, y_i + l_1). \end{aligned}$$

Többfüggvényes egyenletrendszerek esetében hasonlóan járunk el.

150. PÉLDA. $\begin{cases} \dot{x}(t) = -y(t) \,(= f(t,x,y)) \\ \dot{y}(t) = x(t) (= g(t,x,y)) \\ x(0) = 1 \\ y(0) = 0 \end{cases}$, $t \in [0, 2\pi]$ differenciál egyenletrendszer feladat pontos megoldása a trigonometrikus kör: $\begin{cases} x(t) = \cos(t) \\ y(t) = \sin(t) \end{cases}$, $t \in [0, 2\pi]$. A (9.2.3) képleteket használva a következő (diszkrét) közelítést kapjuk az x illetve y függvényekre ($h = 0.1$):

t_i	x_i	y_i
0	1	0
0.1	1	0.1
0.2	0.99	0.2
0.3	0.97	0.299
0.4	0.9401	0.396
$\vdots$	$\vdots$	$\vdots$

A differenciál egyenletrendszerek egyik legegyszerűbb alakja a lineáris alak:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots \\ a_{21} & a_{22} & \\ \vdots & & \ddots \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix} + \begin{pmatrix} b_1(t) \\ b_2(t) \\ \vdots \end{pmatrix},$$

vagy mátrix alakban

$$\dot{X} = AX + B, \tag{9.2.7}$$

ahol A konstansokat tartalmaz, a B oszlopvektor pedig t függvénye.

151. PÉLDA. Az előbbi példa felírható az alábbi alakba ($x_1 := x$, $x_2 := y$):

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

vagyis $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Ha az A mátrix sajátvektorai lineárisan függetlenek, akkor A diagonalizálható és ebben az esetben a differenciál egyenletrendszer által

összekapcsolt függvények leválaszthatóak egymástól egyszerű differenciálegyenleteket eredményezve.

Legyen $\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots \\ 0 & \lambda_2 & \\ \vdots & & \ddots \end{pmatrix}$ az A mátrix átlós alakja, vagyis $\Lambda = S^{-1}AS$ ahol $S = [S_1\ S_2\ \ldots]$ az A mátrix S_i (lineárisan független) sajátvektoraiból alkotott mátrix. Ha $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix}$ úgy, hogy

$$X = SY \tag{9.2.8}$$

akkor a (9.2.7) differenciál egyenletrendszer átírható:

$$S\dot{Y} = ASY + B,$$

vagyis

$$\dot{Y} = S^{-1}ASY + S^{-1}B = \Lambda Y + \overline{B}, \tag{9.2.9}$$

ahol

$$\overline{B} = S^{-1}B = \begin{pmatrix} \overline{b}_1(t) \\ \overline{b}_2(t) \\ \vdots \end{pmatrix}.$$

A (9.2.9) relációt komponensekre lebontva a következő differenciálegyenletekhez jutunk:

$$\dot{y}_i(t) = \lambda_i y_i(t) + \overline{b_i}(t), \ i = 1, 2, \ldots$$

aminek általános megoldása:

$$y_i(t) = e^{\lambda_i t}\left(c_i + \int e^{-\lambda_i t}\overline{b_i}(t)\ dt\right).$$

Visszahelyettesítve a (9.2.8) képletbe megkapjuk az X megoldást.

152. Példa. Az előbbi példát folytatva $\Lambda = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$, $S = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix}$ és (9.2.9)-nek megfelelően

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \Leftrightarrow \begin{cases} \dot{y}_1 = iy_1 \\ \dot{y}_2 = -iy_2 \end{cases}$$

aminek megoldása

$$\begin{cases} y_1(t) = c_1 e^{it} \\ y_2(t) = c_2 e^{-it} \end{cases}.$$

Visszahelyettesítve a (9.2.8) képletbe és felhasználva a kezdeti feltételeket következik, hogy $c_1 = c_2 = \frac{\sqrt{2}}{2}$ és

$$\begin{cases} x_1 = \frac{\left(e^{it}+e^{-it}\right)}{2} = \cos(t) \\ x_2 = \frac{\left(e^{it}-e^{-it}\right)}{2} = \sin(t) \end{cases}.$$

9.3. Magasabb-rendű differenciálegyenletek

Tekintsük a következő harmadrendű differenciálegyenletet:

$$\dddot{y} = f(t, y, \dot{y}, \ddot{y}), \tag{9.3.1}$$

amihez az alábbi kezdeti feltételeket társítjuk:

$$\begin{cases} y(t_0) = y_0 \\ \dot{y}(t_0) = \dot{y}_0 \\ \ddot{y}(t_0) = \ddot{y}_0 \end{cases}.$$

9.3.1. Visszavezetés differenciál egyenletrendszerekre Magasabb-rendű differenciálegyenletek megoldása visszavezethető elsőrendű differenciál egyenletrendszerek megoldására. Ennek érdekében a differenciálegyenlet rendjének megfelelő számú függvényt vezetünk be:

$$y_1 := y,\ y_2 := \dot{y}_1,\ y_3 := \dot{y}_2.$$

Akkor a (9.3.1) differenciálegyenlet a következő differenciál egyenletrendszerrel ekvivalens:

$$\begin{cases} \dot{y}_1 = y_2 \\ \dot{y}_2 = y_3 \\ \dot{y}_3 = f(t, y_1, y_2, y_3) \end{cases}$$

és a kezdeti feltételek:

$$\begin{aligned} y_1(t_0) &= y(t_0) = y_0 := y_1^0 \\ y_2(t_0) &= \dot{y}_1(t_0) = \dot{y}(t_0) = \dot{y}_0 := y_2^0 \\ y_3(t_0) &= \dot{y}_2(t_0) = \ddot{y}_1(t_0) = \ddot{y}_0 := y_3^0. \end{aligned}$$

153. Példa. Vizsgáljuk az alábbi tompítatlan, linearizált, harmonikus mozgást:

$$\begin{cases} \ddot{y} + \frac{1}{10} y = 0 \\ y(0) = 1 \\ \dot{y}(0) = 2 \end{cases}, \ t \in [0, 10].$$

Bevezetve az $y_1 := y$, $y_2 := \dot{y}_1$ függvényeket, az adott differenciálegyenlet átalakítható az alábbi (lineáris) differenciál egyenletrendszerré:

$$\begin{cases} \dot{y}_1 = y_2 \\ \dot{y}_2 = -\frac{1}{10} y_1 \\ y_1(0) = 1 \\ y_2(0) = 2 \end{cases}$$

amit a már ismertetett módszerrel oldunk meg.

9.3.2. Másodrendű perem differenciálegyenlet megoldása véges differenciákkal Tekintsük az általános másodrendű differenciálegyenletet:

$$\ddot{y}(t) + p(t)\dot{y}(t) + q(t)y(t) = r(t), \ t \in [a, b] \tag{9.3.2}$$

ahol $p, q, r \in C[a, b]$. A lényeg újból megközelíteni a tényleges $y(t)$ értéket bizonyos $(t_i)_{i=0}^n$ ekvidisztáns csomópontokban: $y(t_i) = y_i$.

Ennek érdekében felhasználjuk a numerikus deriválásból ismert képleteket, például:

$$\begin{aligned}\dot{y}(t_i) &= \frac{y(t_{i+1})-y(t_{i-1})}{2h}=\frac{y_{i+1}-y_{i-1}}{2h}\\ \ddot{y}(t_i) &= \frac{y(t_{i+1})-2y(t_i)+y(t_{i-1})}{h^2}=\frac{y_{i+1}-2y_i+y_{i-1}}{h^2}\end{aligned}$$

ahol $y_i := y(t_i)$ jelölést használtuk. Behelyettesítve a (9.3.2) képletbe és a $p_i = p(t_i), q_i = q(t_i), r_i = r(t_i)$ jelöléseket használva a következő lineáris egyenletrendszert kapjuk:

$$\frac{y_{i+1}-2y_i+y_{i-1}}{h^2}+p_i\frac{y_{i+1}-y_{i-1}}{2h}+q_iy_i=r_i,\ \ i=1,...,n-1$$

ami átrendezve:

(9.3.3)

$$\left(1-\frac{p_ih}{2}\right)y_{i-1}-\left(2-q_ih^2\right)y_i+\left(1+\frac{p_ih}{2}\right)y_{i+1}=r_ih^2,\ \ i=1,...,n-1.$$

Figyelembe véve, hogy y_0, y_n adott, a (9.3.3) lineáris egyenletrendszer mátrix alakja:

$$\begin{pmatrix} -(2-q_1h^2) & \left(1+\frac{p_1h}{2}\right) & 0 & 0 & 0 & 0\\ \left(1-\frac{p_2h}{2}\right) & -(2-q_2h^2) & \left(1+\frac{p_2h}{2}\right) & 0 & 0 & 0\\ 0 & \left(1-\frac{p_3h}{2}\right) & -(2-q_3h^2) & \left(1+\frac{p_3h}{2}\right) & 0 & 0\\ & & & \ddots & & \\ 0 & 0 & 0 & 0 & \left(1-\frac{p_{n-2}h}{2}\right) & -(2-q_{n-2}h^2) \quad \left(1\right.\\ 0 & 0 & 0 & 0 & 0 & \left(1-\frac{p_ih}{2}\right) \quad -(2 \end{pmatrix}$$

$$\cdot\begin{pmatrix} y_1\\ y_2\\ \\ \vdots\\ \\ y_{n-2}\\ y_{n-1}\end{pmatrix}=\begin{pmatrix} r_0h^2-\left(1-\frac{p_1h}{2}\right)y_0\\ r_1h^2\\ \\ \vdots\\ \\ r_{n-2}h^2\\ r_{n-1}h^2-\left(1-\frac{p_{n-1}h}{2}\right)y_n\end{pmatrix}.$$

Irodalomjegyzék

[1] Bahvalov, N.Sz., *A gépi matematika numerikus módszerei*, Műszaki Könyvkiadó, Budapest, 1977.

[2] Bajcsay, P., *Numerikus analízis*, Tankönyvkiadó, Budapest, 1978.

[3] Bjezikovics, Ja.Sz., *Közelítő számítások*, Tankönyvkiadó, Budapest, 1952.

[4] Bjoerck A., Dahlquist G., *Numerical mathematics and scientific computation*, vol.1, SIAM, 2008.

[5] de Boor, C.: *A practical guide to splines*, Springer-Verlag New York Heidelberg Berlin, 1978.

[6] Conte S.D., de Boor C., *Elementary numerical analysis. An algorithmic approach,* 3ed., McGraw-Hill, 1980.

[7] Coman, Gh.: *Analiză numerică*, Editura Libris, Cluj-Napoca, 1995.

[8] Duff, I.S., Erisman, A.M., Reid, J.K.: *Direct methods for sparse matrices,* Oxford Science Publications, 1989.

[9] Farin, G., Hansford, D.: *The essentials of CAGD*, A.K. Peters, 2000.

[10] Forsythe G.E., Malcolm M.A., Moler C.B., *Computer methods for mathematical computations,* Prentice-Hall, 1977.

[11] Golub, G.H., van Loan, C.F.: *Matrix computations*, The Johns Hopkins University Press, (3ed.) 1996.

[12] Hoffman J.D.: *Numerical methods for engineers and scientists,* (2ed.), M.Dekker, 2001.

[13] Kahaner, D., Moler, C., Nash, S.: *Numerical methods and software*, Prentice Hall, 1988.

[14] Lazăr, I.: *Metode numerice cu funcţii în C++*, Presa universitară clujeană, Cluj-Napoca, 2001.

[15] Mathews, J.H, Fink, K.D., *Numerical methods using Matlab, (3,ed.)* Prentice-Hall, 1999.

[16] Moler, C.: *Numerical Computing with Matlab*, SIAM, 2008.

[17] Obádovics, J. Gy.: *Gyakorlati számítási eljárások*, Gondolat Kiadó, 1972.

[18] Popper, Gy., Csizmás, F.: *Numerikus módszerek mérnököknek*, Akadémiai Kiadó, Typotex, Budapest, 1993.

[19] Press W., Teukolsky S., Vetterling W., Flannery B., *Numerical recipes in C*, Cambridge Univ. Press, 1992.

[20] Singiresu, R.: *Applied numerical methods for engineers and scientists*, New Jersey Prentice-Hall, 2002.

[21] Stoer J., Bulirsch, R., *Introduction to Numerical Analysis*, Springer New York, 2002.

[22] Stoyan, G., Takó, G.: *Numerikus módszerek*, Typotex, 2005.

[23] Strang G.: *Introduction to applied mathematics,* Wellesley-Cambridge, 1986.

[24] Trefethen, L.N., Bau, D.: Numerical linear algebra, SIAM, 1997.

[25] Van Loan, Ch.F., *Introduction to Scientific Computing*, Prentice-Hall, Upper Sad- dle River, NJ, 2000.

[26] Virágh, J.: *Numerikus matematika*, Szeged JATE Press, 2003.

[27] Vladislav, T., Raşa, I.: *Analiză numerică*, Editura Tehnică, Bucureşti, 1997.

[28] ***, *Analiză numerică-Lucrări de laborator*, Univ. Babeş-Bolyai, Fac. de Matematică şi Informatică, Cluj Napoca, 1994.