Numerikus matematika

Baran Ágnes

Előadás
Lebegőpontos számok

Lebegőpontos számok

Lebegőpontos számok

Példa.

$a = 10$

$$0.3721 = \frac{3}{10} + \frac{7}{10^2} + \frac{2}{10^3} + \frac{1}{10^4}$$

$$21.65 = 0.2165 \cdot 10^2 = \left(\frac{2}{10} + \frac{1}{10^2} + \frac{6}{10^3} + \frac{5}{10^4}\right) \cdot 10^2$$

$a = 2$

$$0.1101 = \frac{1}{2} + \frac{1}{2^2} + \frac{0}{2^3} + \frac{1}{2^4}$$

$$0.001011 = 0.1011 \cdot 2^{-2} = \left(\frac{1}{2} + \frac{0}{2^2} + \frac{1}{2^3} + \frac{1}{2^4}\right) \cdot 2^{-2}$$

Lebegőpontos számok

A nemnulla lebegőpontos számok alakja:

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

ahol

$a > 1$ egész, a számábrázolás alapja

$t > 1$, egész, a mantissza hossza

$k_- \leq k \leq k_+$ egész, a karakterisztika, ahol $k_- < 0$ és $k_+ > 0$ adott

$1 \leq m_1 \leq a - 1$, egész (a szám normalizált)

$0 \leq m_i \leq a - 1$, egész, ha $i = 2, \ldots, t$

röviden: $[\pm|k|m_1,\dots,m_t]$
ahol $(m_1,\dots,m_t)$ a mantissza.

Az a, t, k_-, k_+ értékek egyértelműen leírják az ábrázolható számok halmazát.

Példa.

Legyen $a = 2,\ t = 4,\ k_- = -3,\ k_+ = 2.$

(a) Írjuk fel a következő számok lebegőpontos alakját:

$$0.6875, \quad 0.8125, \quad 3.25$$

(b) Hány pozitív normalizált lebegőpontos szám ábrázolható ilyen jellemzők mellett?

A legnagyobb ábrázolható szám:

$$M_\infty = a^{k_+}\left(\frac{a-1}{a}+\frac{a-1}{a^2}+\cdots+\frac{a-1}{a^t}\right)$$

$$= a^{k_+}\left(1-\frac{1}{a}+\frac{1}{a}-\frac{1}{a^2}+\cdots+\frac{1}{a^{t-1}}-\frac{1}{a^t}\right)$$
$$= a^{k_+}\left(1-a^{-t}\right)$$

A legkisebb pozitív normalizált ábrázolható szám:

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

Szubnormális számok: ha $k = k_-$, akkor $m_1 = 0$ is lehet.

Az 1 mindig lebegőpontos szám:

$$1 = a^1 \cdot \frac{1}{a}, \qquad \text{vagy röviden: } 1 = [+|1|1, 0, \dots, 0]$$

Az 1 jobboldali szomszédja:

$$1 + \varepsilon_1 = [+|1|1, 0, \dots, 0, 1]$$

másképp:

$$1 + \varepsilon_1 = a\left(\frac{1}{a} + 0 + \cdots + 0 + \frac{1}{a^t}\right) = 1 + a^{1-t}$$

azaz $\varepsilon_1 = a^{1-t}$ (**gépi epszilon**)

Az IEEE lebegőpontos aritmetikai szabvány:

	egyszeres pontosság	dupla pontosság
méret	32 bit	64 bit
mantissza	23+1 bit	52+1 bit
karakterisztika	8 bit	11 bit
ε_1	$\approx 1.19 \cdot 10^{-7}$	$\approx 2.22 \cdot 10^{-16}$
M_∞	$\approx 10^{38}$	$\approx 10^{308}$

mivel m_1 mindig 1, ezért nem ábrázoljuk
az előjel ábrázolására 1 bit

Adott a, t, k_+, k_- mellett az ábrázolható lebegőpontos számok a $[-M_\infty, M_\infty]$ intervallum egy megszámlálható részhalmazát alkotják.

Példa

(a) Ábrázoljuk számegyenesen az $a = 2$, $t = 4$, $k_- = -3$, $k_+ = 2$ jellemzők mellett felírható összes pozitív normalizált lebegőpontos számot.

(b) A fenti számábrázolási jellemzők mellett mennyi lesz M_∞, ε_0 és ε_1 értéke?

(c) Mit mondhatunk két szomszédos szám távolságáról?

(d) Mit mondhatunk a szomszédos számok távolságáról, ha k_+ értékét 4-re módosítjuk?

(e) Mi lenne, ha $k_+ > 4$ teljesülne?

Példa.

A pozitív normalizált lebegőpontos számok $a = 2$, $t = 4$, $k_- = -3$, $k_+ = 2$ esetén.

	$k=0$	$k=1$	$k=2$	$k=-1$	$k=-2$	$k=-3$
0.1000	$\frac{8}{16}$	$\frac{8}{8}$	$\frac{8}{4}$	$\frac{8}{32}$	$\frac{8}{64}$	$\frac{8}{128}$
0.1001	$\frac{9}{16}$	$\frac{9}{8}$	$\frac{9}{4}$	$\frac{9}{32}$	$\frac{9}{64}$	$\frac{9}{128}$
0.1010	$\frac{10}{16}$	$\frac{10}{8}$	$\frac{10}{4}$	$\frac{10}{32}$	$\frac{10}{64}$	$\frac{10}{128}$
0.1011	$\frac{11}{16}$	$\frac{11}{8}$	$\frac{11}{4}$	$\frac{11}{32}$	$\frac{11}{64}$	$\frac{11}{128}$
0.1100	$\frac{12}{16}$	$\frac{12}{8}$	$\frac{12}{4}$	$\frac{12}{32}$	$\frac{12}{64}$	$\frac{12}{128}$
0.1101	$\frac{13}{16}$	$\frac{13}{8}$	$\frac{13}{4}$	$\frac{13}{32}$	$\frac{13}{64}$	$\frac{13}{128}$
0.1110	$\frac{14}{16}$	$\frac{14}{8}$	$\frac{14}{4}$	$\frac{14}{32}$	$\frac{14}{64}$	$\frac{14}{128}$
0.1111	$\frac{15}{16}$	$\frac{15}{8}$	$\frac{15}{4}$	$\frac{15}{32}$	$\frac{15}{64}$	$\frac{15}{128}$

$M_\infty = 2^2(1 - 2^{-4}) = \frac{15}{4}$ és $\varepsilon_0 = 2^{-3-1} = \frac{1}{16} \left(= \frac{8}{128}\right)$

Legyen $y = a^k \cdot 0.m_1 m_2 ... m_t$.

A legközelebbi nála nagyobb szám

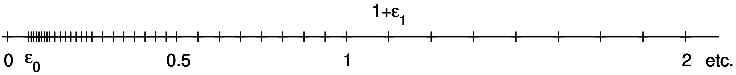
$$a^k \cdot \frac{1}{a^t} = a^{k-t}$$

távolságra van tőle.

Nagyobb karakterisztika $\rightarrow$ nagyobb lépésköz.

Ha $k > t$, akkor a lépésköz nagyobb mint 1.

$a = 2,\ t = 4,\ k_- = -3$ esetén



$\varepsilon_0 = a^{k_- - 1} = 2^{-4} = \frac{1}{16}$,

$\varepsilon_1 = a^{1-t} = 2^{-3} = \frac{1}{8}$

Példa

Vizsgáljuk meg számítógépünkön a $2^{66} + 1 == 2^{66}$, $2^{66} + 10 == 2^{66}$, $2^{66} + 100 == 2^{66}$, $2^{66} + 1000 == 2^{66}$ és $2^{66} + 10000 == 2^{66}$ logikai kifejezések értékét!

Dupla pontosság esetén ($t = 53$):

y	a jobboldali szomszéd távolsága
1	$\approx 2.22 \cdot 10^{-16}$
16	$\approx 3.5527 \cdot 10^{-15}$
1024	$\approx 2.27 \cdot 10^{-13}$
$2^{20} \approx 10^6$	$\approx 2.33 \cdot 10^{-10}$
$2^{52} \approx 4.5 \cdot 10^{15}$	1
$2^{60} \approx 1.15 \cdot 10^{18}$	256
$2^{66} \approx 7.38 \cdot 10^{19}$	16384

Kerekítés

A $[-M_\infty, M_\infty]$ intervallumból nem minden szám írható fel lebegőpontos alakban.

Példa

A 0.1 kettes számrendszerbeli alakja:

$$0.00011001100110011 00....$$

Az $\frac{1}{3}$ kettes számrendszerbeli alakja:

$$0.0101010101010....$$

Kerekítés

Legyen $x \in [-M_\infty, M_\infty]$ egy valós szám, $fl(x)$ pedig a hozzárendelt lebegőpontos szám.

Szabályos kerekítés esetén:

$$fl(x) = \begin{cases} 0, \quad \text{ha } |x| < \varepsilon_0 \\ \text{az } x\text{-hez legközelebbi lebegőpontos számok} \\ \text{közül a nagyobb abszolút értékű,} & \text{ha } |x| \geq \varepsilon_0 \end{cases}$$

Levágás esetén:

$$fl(x) = \begin{cases} 0, \quad \text{ha } |x| < \varepsilon_0 \\ \text{az } x\text{-hez legközelebbi lebegőpontos szám a 0 felé, ha } |x| \geq \varepsilon_0 \end{cases}$$

Megjegyzés

Ha az ábrázolni kívánt szám két szomszédos lebegőpontos szám között félúton helyezkedik el, akkor a valóságban az előzőnél bonyolultabb kerekítési szabály alapján történik a kerekítés.

Példa

Legyen $a = 2$, $t = 4$, $k_- = -3$, $k_+ = 2$. Mi lesz a 0.1-hez rendelt lebegőpontos szám szabályos kerekítés, illetve levágás esetén?

A 0.1 kettes számrendszerben, normalizálva:

$$2^{-3} \cdot 0.1100110011001100\ldots.$$

Szabályos kerekítés:

$$fl(0.1) = 2^{-3} \cdot 0.1101$$

Levágás:

$$fl(0.1) = 2^{-3} \cdot 0.1100$$

Kerekítés

Az **abszolút hiba** becslése

szabályos kerekítésnél:

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

levágásnál:

$$|fl(x) - x| \leq \begin{cases} \varepsilon_0, & \text{ha } |x| < \varepsilon_0 \\ \varepsilon_1 |x|, & \text{ha } |x| \geq \varepsilon_0 \end{cases}$$

Kerekítés

A **relatív hiba** becslése, ha $|x| \geq \varepsilon_0$

szabályos kerekítésnél:

$$\frac{|fl(x) - x|}{|x|} \leq \frac{1}{2}\varepsilon_1$$

levágásnál:

$$\frac{|fl(x) - x|}{|x|} \leq \varepsilon_1$$

Gépi epszilon (ε_1)

Adott számábrázolási jellemzők mellett az 1 és a jobboldali lebegőpontos szomszédjának a távolsága.

Alapműveleteknél:

1. példa:

$a = 10,\ t = 3$
$x = 0.425 \cdot 10^{-1},\ y = 0.677 \cdot 10^{-2}$

$fl(x + y) =?$

$y \to y = 0.0677 \cdot 10^{-1}$ (**tartalék számjegyek**)

$x + y = 0.425 \cdot 10^{-1} + 0.0677 \cdot 10^{-1} = 0.4927 \cdot 10^{-1}$

$$fl(x+y) = \begin{cases} 0.492 \cdot 10^{-1}, & \text{levágás} \\ 0.493 \cdot 10^{-1}, & \text{szabályos kerekítés} \end{cases}$$

2. példa:

$a = 10,\ t = 3$
$x = 0.367 \cdot 10^{-2},\ y = 0.682 \cdot 10^{-2}$

$fl(x + y) = ?$

$x + y = 0.367 \cdot 10^{-2} + 0.682 \cdot 10^{-2} = 1.049 \cdot 10^{-2} = 0.1049 \cdot 10^{-1}$

$$fl(x+y) = \begin{cases} 0.104 \cdot 10^{-1}, & \text{levágás} \\ 0.105 \cdot 10^{-1}, & \text{szabályos kerekítés} \end{cases}$$

Alapműveleteknél:

Jelölje $\triangle$ a négy alapművelet valamelyikét, legyen x és y lebegőpontos szám. Tfh a gép a műveletet pontosan végrehajtja és az eredményhez hozzárendel egy lebegőpontos számot. Ekkor

szabályos kerekítés esetén:

$$|fl(x \triangle y) - x \triangle y| \leq \begin{cases} \varepsilon_0, & \text{ha } |x \triangle y| < \varepsilon_0 \\ \frac{1}{2}\varepsilon_1 |x \triangle y|, & \text{ha } |x \triangle y| \geq \varepsilon_0 \end{cases}$$

levágás esetén:

$$|fl(x \triangle y) - x \triangle y| \leq \begin{cases} \varepsilon_0, & \text{ha } |x \triangle y| < \varepsilon_0 \\ \varepsilon_1 |x \triangle y|, & \text{ha } |x \triangle y| \geq \varepsilon_0 \end{cases}$$

Összefoglalva:

ha $|x \triangle y| > M_\infty$, akkor **túlcsordulás**,

ha $|x \triangle y| < \varepsilon_0$, akkor **alulcsordulás** ($fl(x \triangle y) = 0$)

ha $\varepsilon_0 \le |x \triangle y| \le M_\infty$, akkor az előző reláció átírható:

$$fl(x \triangle y) = (x \triangle y) \cdot (1 + \varepsilon_\triangle), \quad \text{ahol } |\varepsilon_\triangle| \le \varepsilon_1 \begin{cases} 1, & \text{levágás} \\ \frac{1}{2}, & \text{szabályos kerekítés} \end{cases}$$

A hibák terjedése

Legyenek $x_0, x_1, \ldots, x_n$ lebegőpontos számok.

$S_n = \sum_{i=0}^{n} x_i =?$, ha az összeadás algoritmusa:

$$S_0 = x_0, \qquad S_k = S_{k-1} + x_k, \quad k = 1, \ldots, n.$$

A hiba becslése:

$$|fl(S_n) - S_n| \leq n\varepsilon_1 |x_0| + n\varepsilon_1 |x_1| + (n-1)\varepsilon_1 |x_2| + \cdots + \varepsilon_1 |x_n|$$

Egy durvább becslés:

$$|fl(S_n) - S_n| \leq n\varepsilon_1 \sum_{k=0}^{n} |x_k|$$

Ha minden x_k pozitív, akkor

$$\left| \frac{fl(S_n) - S_n}{S_n} \right| \leq n\varepsilon_1$$

Megjegyzések

- A lebegőpontos összeadás nem asszociatív

Példa

Vizsgálja meg számítógépén a $0.4 - 0.5 + 0.1 == 0$ és a $0.1 - 0.5 + 0.4 == 0$ logikai kifejezések értékét.

- Az elvégzett műveletek számának növekedésével a kerekítési hiba tipikusan nő. Matematikailag ekvivalens kifejezések értékére lényegesen különböző értékeket kaphatunk a gépi számítás során.

Példa

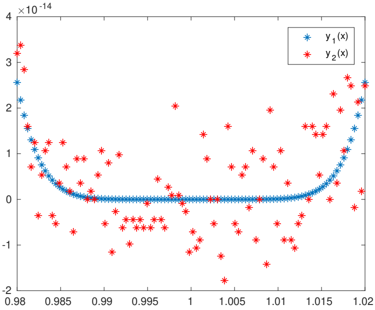
Az alábbi algoritmus végrehajtása után mennyi az x elméleti, illetve a gépi számítás után adódó értéke?

```
x=1/3;
for i=1:40
   x=4*x-1;
end
```

Példa

Számítógépén határozza meg és ábrázolja az 1 egy kis környezetében az $(x-1)^8$ kifejezés értéket az alábbi két (matematikailag ekvivalens) módon:

$$y_1(x) = (x-1)^8,$$
$$y_2(x) = x^8 - 8x^7 + 28x^6 - 56x^5 + 70x^4 - 56x^3 + 28x^2 - 8x + 1$$



Megjegyzések

- A kifejezések alkalmas átalakításával elkerülhető, hogy a köztes eredmények (és így a végeredmény is) túlcsorduljanak.

Példa

Legyen $x = (10^{200}, 1)$. Számítsa ki gépén az x normáját az alábbi két módon.

(a)

$$\|x\| = \sqrt{x_1^2 + x_2^2}$$

(b)

$$c = \max\{|x_1|, |x_2|\}, \quad \|x\| = c \cdot \sqrt{\left(\frac{x_1}{c}\right)^2 + \left(\frac{x_2}{c}\right)^2}$$

Lineáris egyenletrendszerek

Példa (Delta fedezet)

Egy jövőbeli kötelezettségünk, piaci folyamatoktól függően, kétféleképpen realizálódhat: vagy 1000\$-t kell fizetnünk, vagy 0\$-t. Erre felkészülve, a kockázatokat előre kezelve, be akarunk fektetni valamennyi pénzt. Két befektetési lehetőségünk van: a pénz egy részét leköthetjük a bankszámlánkon 2% kamatozással, másik részéből 100\$ darabáron részvényeket vásárolhatunk. A részvénynek két lehetséges hozama van: $+6\%$, vagy -6%, a kötelezettségeink: ha a részvény hozama pozitív, akkor 1000\$-t kell fizetnünk, ha negatív, akkor 0\$-t. Megoldható-e a feladat, ha igen, akkor mekkora összeget kell befektetnünk?

Példa

Egy kisvállalkozás háromféle terméket gyárt (I., II., és III.), mindháromhoz szükség van az N_1, N_2 és N_3 nyersanyagokra. Az egyes termékekből 1 csomag legyártásához a táblázatban adott egységek szükségesek az adott nyersanyagokból.

	I.	II.	III.
N_1	2	1	2
N_2	4	4	5
N_3	2	5	5

Ha tudjuk, hogy egy adott napon az egyes nyersanyagokból rendre 171, 431 és 376 egység fogyott, akkor melyik termékből hány csomagot gyártottak?

Megoldás:

x_1, x_2, x_3: az I., II., III. termékből legyárott csomagok száma

$$\begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix} x_1 + \begin{bmatrix} 1 \\ 4 \\ 5 \end{bmatrix} x_1 + \begin{bmatrix} 2 \\ 5 \\ 5 \end{bmatrix} x_3 = \begin{bmatrix} 171 \\ 431 \\ 376 \end{bmatrix}$$

Hogy lehet kikombinálni az I., II., III. termékek egy csomagjához szükséges nyersanyagok vektorából az összes nyersanyag vektorát?

Mátrix-vektor alakban:

$$\begin{bmatrix} 2 & 1 & 2 \\ 4 & 4 & 5 \\ 2 & 5 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 171 \\ 431 \\ 376 \end{bmatrix},$$

azaz $Ax = b$.

Gauss-elimináció

$$\left[\begin{array}{ccc|c} 2 & 1 & 2 & 171 \\ 4 & 4 & 5 & 431 \\ 2 & 5 & 5 & 376 \end{array}\right] \to \left[\begin{array}{ccc|c} 2 & 1 & 2 & 171 \\ 0 & 2 & 1 & 89 \\ 0 & 4 & 3 & 205 \end{array}\right] \to \left[\begin{array}{ccc|c} 2 & 1 & 2 & 171 \\ 0 & 2 & 1 & 89 \\ 0 & 0 & 1 & 27 \end{array}\right]$$

Visszahelyettesítéssel: (alulról felfelé)

$$x_3 = 27$$
$$2x_2 + x_3 = 89 \to x_2 = 31$$
$$2x_1 + x_2 + 2x_3 = 171 \to x_1 = 43$$

A megoldás: 43 csomag I. termék, 31 csomag II. termék, 27 csomag III. termék.

A visszahelyettesítés helyett folytathattuk volna az eliminációt (Gauss-Jordan elimináció):

$$\left[\begin{array}{ccc|c} 2 & 1 & 2 & 171 \\ 0 & 2 & 1 & 89 \\ 0 & 0 & 1 & 27 \end{array}\right] \to \left[\begin{array}{ccc|c} 2 & 1 & 0 & 117 \\ 0 & 2 & 0 & 62 \\ 0 & 0 & 1 & 27 \end{array}\right] \to \left[\begin{array}{ccc|c} 2 & 0 & 0 & 86 \\ 0 & 1 & 0 & 31 \\ 0 & 0 & 1 & 27 \end{array}\right] \to$$

$$\to \left[\begin{array}{ccc|c} 1 & 0 & 0 & 34 \\ 0 & 1 & 0 & 31 \\ 0 & 0 & 1 & 27 \end{array}\right]$$

Ekkor a jobb oldalon a megoldásvektort kapjuk.

Matlab-ban

- A backslash operátorral: Az A mátrix és b vektor megadása után

```
>> x=A\b
x =
   43
   31
   27
```

- Az rref függvénnyel:

```
  >> rref([A b])
ans =
     1     0     0    43
     0     1     0    31
     0     0     1    27
```

Ez a Gauss-Jordan elimináció végén kapott mátrixot adja vissza.

Példa

Egy kisvállalkozás háromféle terméket gyárt (I., II., és III.), mindháromhoz szükség van az N_1, N_2 és N_3 nyersanyagokra. Az egyes termékekből 1 csomag legyártásához a táblázatban adott egységek szükségesek az adott nyersanyagokból.

	I.	II.	III.
N_1	2	1	5
N_2	4	4	8
N_3	2	5	1

Miután a nap végén a raktáros jelenti, hogy aznap az egyes nyersanyagokból rendre 252, 512 és 266 egység fogyott, a gyártásvezető elrendelt egy ellenőrzést. Miért?

Most a megfelelő egyenletrendszer kibővített mátrixával elvégezve a Gauss-eliminációt:

$$\left[\begin{array}{ccc|c} 2 & 1 & 5 & 252 \\ 4 & 4 & 8 & 512 \\ 2 & 5 & 1 & 266 \end{array}\right] \to \left[\begin{array}{ccr|r} 2 & 1 & 5 & 252 \\ 0 & 2 & -2 & 8 \\ 0 & 4 & -4 & 14 \end{array}\right] \to \left[\begin{array}{ccr|r} 2 & 1 & 5 & 252 \\ 0 & 2 & -2 & 8 \\ 0 & 0 & 0 & -2 \end{array}\right]$$

A 3. egyenlet jelentése: $0x_1 + 0x_2 + 0x_3 = -2$, ami ellentmondás.

Ha folytatnánk az eliminációt:

$$\left[\begin{array}{ccr|c} 2 & 1 & 5 & 0 \\ 0 & 2 & -2 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccr|c} 2 & 0 & 6 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccr|c} 1 & 0 & 3 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

Matlab-ban

- A backslash operátorral: Az A mátrix és b vektor megadása után

```
>> x=A\b
```

Warning: Matrix is close to singular or badly scaled. Results may be inaccurate. RCOND = 7.930164e-19.

```
x =
   1.0e+16 *
    3.6029
   -1.2010
   -1.2010
```

Figyelmeztet, hogy az eredmény pontatlan lehet. Valóban, ha ellenőrzésképpen kiszámítjuk Ax értékét, akkor

```
>> A*x
ans =
   256
   528
   274
```

ami nem egyenlő b-vel.

Matlab-ban

- Az rref függvénnyel:

```
  >> rref([A b])
ans =
     1     0     3     0
     0     1    -1     0
     0     0     0     1
```

Ez a Gauss-Jordan elimináció végén kapott mátrixot adja vissza.

Innen azonnal látjuk, hogy a rendszer ellentmondásos, nincs olyan x_1, x_2, x_3, mely kielégíti a megadott egyenleteket.

Példa

Egy kisvállalkozás háromféle terméket gyárt (I., II., és III.), mindháromhoz szükség van az N_1, N_2 és N_3 nyersanyagokra. Az egyes termékekből 1 csomag legyártásához a táblázatban adott egységek szükségesek az adott nyersanyagokból.

	I.	II.	III.
N_1	2	1	5
N_2	4	4	8
N_3	2	5	1

Ha tudjuk, hogy egy adott napon az egyes nyersanyagokból rendre 109, 308 és 289 egység fogyott, akkor melyik termékből hány csomagot gyártottak?

Gauss-eliminációval:

$$\left[\begin{array}{ccc|c} 2 & 1 & 5 & 109 \\ 4 & 4 & 8 & 308 \\ 2 & 5 & 1 & 289 \end{array}\right] \to \left[\begin{array}{ccr|r} 2 & 1 & 5 & 109 \\ 0 & 2 & -2 & 90 \\ 0 & 4 & -4 & 180 \end{array}\right] \to \left[\begin{array}{ccr|r} 2 & 1 & 5 & 109 \\ 0 & 2 & -2 & 90 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

A 3. egyenlet jelentése: $0x_1 + 0x_2 + 0x_3 = 0$, ami semmilyen korlátozást nem jelent az ismeretlenekre.
Folytatva az eliminációt:

$$\left[\begin{array}{ccr|r} 2 & 1 & 5 & 109 \\ 0 & 1 & -1 & 45 \\ 0 & 0 & 0 & 0 \end{array}\right] \to \left[\begin{array}{ccr|r} 2 & 0 & 6 & 64 \\ 0 & 1 & -1 & 45 \\ 0 & 0 & 0 & 0 \end{array}\right] \to \left[\begin{array}{ccr|r} 1 & 0 & 3 & 32 \\ 0 & 1 & -1 & 45 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Ha eltekintünk attól a feltételtől, hogy nemnegatív egész megoldásokat keresünk, akkor a rendszernek végtelen sok megoldása van. Pl. egy megoldás: $x_1 = 32$, $x_2 = 45$, $x_3 = 0$

Matlab-ban

- A backslash operátorral: Az A mátrix és b vektor megadása után

```
>> x=A\b
```

Warning: Matrix is close to singular or badly scaled. Results may be inaccurate. RCOND = 7.930164e-19.

```
x =
   167
     0
   -45
```

Újra figyelmeztetést kaptunk, de Ax most egyenlő b-vel:

```
>> A*x
ans =
   109
   308
   289
```

Matlab-ban

- Az rref függvénnyel:

```
  >> rref([A b])
ans =
     1     0     3    32
     0     1    -1    45
     0     0     0     0
```

Ez a Gauss-Jordan elimináció végén kapott mátrixot adja vissza, innen egy megoldást azonnal leolvashatunk.

Gyengén meghatározott lineáris egyenletrendszerek

Példa. Az

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

egy.rendszer megoldása:

$$x = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

Az

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2.0001 \end{bmatrix}$$

egy.rendszer megoldása:

$$y = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Gyengén meghatározott lineáris egyenletrendszerek

Példa. Tekintsük a következő 100×100-as lineáris egyenletrendszert:

$$\begin{bmatrix} 1 & -1 & -1 & -1 & \cdots & -1 & -1 \\ 0 & 1 & -1 & -1 & \cdots & -1 & -1 \\ 0 & 0 & 1 & -1 & \cdots & -1 & -1 \\ \vdots & & & & & & \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{99} \\ x_{100} \end{bmatrix} = \begin{bmatrix} -98 \\ -97 \\ -96 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

Ez egyértelműen megoldható:

$$x_1 = x_2 = x_3 = \cdots = x_{99} = x_{100} = 1.$$

Perturbáljuk egy kicsit a rendszert!

Gyengén meghatározott lineáris egyenletrendszerek

$$\begin{bmatrix} 1 & -1 & -1 & -1 & \cdots & -1 & -1 \\ 0 & 1 & -1 & -1 & \cdots & -1 & -1 \\ 0 & 0 & 1 & -1 & \cdots & -1 & -1 \\ \vdots & & & & & & \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{99} \\ x_{100} \end{bmatrix} = \begin{bmatrix} -98 \\ -97 \\ -96 \\ \vdots \\ 0 \\ \color{red}{1.00001} \end{bmatrix}$$

Ez is egyértelműen megoldható, de

$$x_1 \approx 3.1691 \cdot 10^{24}.$$

Egy pici perturbáció az adatokban $\rightarrow$ hatalmas különbség a megoldásban.

Normák, kondíciószámok

$A \in \mathbb{R}^{n\times n}$ invertálható, $b \in \mathbb{R}^n$, $b \neq 0$
az $Ax = b$ lin. egyenletrendszer megoldását keressük.

Tfh b hibával terhelten ismert: b helyett $b + \delta b$ adott. Ekkor a lineáris egyenletrendszer:

$$Ay = b + \delta b$$

vagy

$$A(x + \delta x) = b + \delta b$$

A kérdés: mekkora lehet a megoldás hibája?

A jobb oldali vektor változása mekkora hatással van a megoldás változására?

vektorokat kell mérnünk $\to$ normák

Norma

Legyen X egy lineáris tér $\mathbb{R}$ felett. Az $d : X \to \mathbb{R}$ leképezés **norma**, ha

1. $d(x) \geq 0$ minden $x \in X$ esetén
2. $d(x) = 0 \iff x = 0$
3. $d(\lambda x) = |\lambda| d(x)$, minden $\lambda \in \mathbb{R}$ és $x \in X$ esetén
4. $d(x + y) \leq d(x) + d(y)$ minden $x, y \in X$ esetén (háromszög-egyenlőtlenség)

A továbbiakban $d(x)$ helyett $\|x\|$

Példák:

Legyen $X = \mathbb{R}^n$

1. Az 1-norma, vagy oktaéder norma:

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

2. A 2-norma, vagy euklideszi norma:

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

3. A ∞-norma, vagy maximum norma:

$$\|x\|_\infty = \max_{1 \le i \le n} |x_i|$$

Példa.

Ha

$$x = \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix}$$

akkor

$$\|x\|_1 = |-3| + |0| + |1| = 4$$
$$\|x\|_2 = \left(|-3|^2 + |0|^2 + |1|^2\right)^{1/2} = \sqrt{10}$$
$$\|x\|_\infty = \max\{|-3|, |0|, |1|\} = 3$$

Abszolút hiba, relatív hiba

Az $A(x + \delta x) = b + \delta b$ rendszerben:

- $\|\delta b\|$: a jobb oldal abszolút hibája
- $\|\delta x\|$: a megoldás abszolút hibája

Ezek önmagukban nem elég informatívak.

Sokkal érdekesebb számunkra:

- $\frac{\|\delta b\|}{\|b\|}$: a jobb oldal relatív hibája
- $\frac{\|\delta x\|}{\|x\|}$: a megoldás relatív hibája

Meg lehet mutatni, hogy a megoldás relatív hibája a jobb oldali vektor relatív hibájától függ, és egy olyan mennyiségtől, ami csak az A mátrixtól függ. Ez utóbbi a mátrix kondíciószáma, ami egy 1-nél nem kisebb valós szám.

A kondíciószám azt mutatja meg, hogy adott mátrix esetén hányszor nagyobb lehet a megoldás relatív hibája a jobboldali vektor relatív hibájánál.

$$\boxed{\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\delta b\|}{\|b\|}}$$

$\mathrm{cond}(A) \geq 1$ minden A invertálható mátrixra.

A kondíciószám meghatározására a Matlabot használhatjuk.

A kondíciószám értéke függ attól, hogy milyen vektornormát használunk.

Legyen b relatív hibája $\frac{\|\delta b\|}{\|b\|} \approx \varepsilon_1$ (inputhiba nagyságrendű).
Ekkor ha

$$cond(A) \geq \frac{1}{\varepsilon_1}$$

akkor

$$cond(A)\frac{\|\delta b\|}{\|b\|} \geq 1$$

azaz a megoldásra rakódó hiba ugyanakkora lehet, mint maga a megoldás.
Az egyenletrendszer rosszul kondícionált.

Ahhoz, hogy a megoldásnak legalább 1 helyes számjegye legyen

$$cond(A) \leq \frac{1}{a\varepsilon_1}$$

kell, mert ekkor

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{1}{a}$$

Numerikus matematika

Baran Ágnes

Előadás
Legkisebb négyzetek módszere

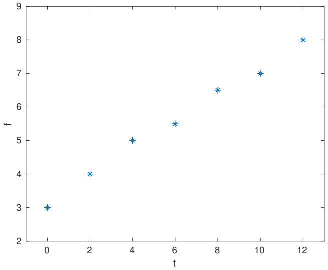
Legkisebb négyzetek módszere

Példa

Egy fél méter magas, téglatest alakú víztartályt egyenletes sebességgel töltenek fel vízzel. Amikor a tartályban 3 cm magasan áll a víz Péter elhatározza, hogy megméri a vízszint változását az idő függvényében. A következő méréseket végezte:

t_i (min)	0	2	4	6	8	10	12
f_i (cm)	3	4	5	5.5	6.5	7	8

Becsülje meg milyen magasan lesz a víz 20 perccel azután, hogy Péter elindította a mérést! Mikor indították el a tartály feltöltését? Kb mikor lesz tele a tartály?

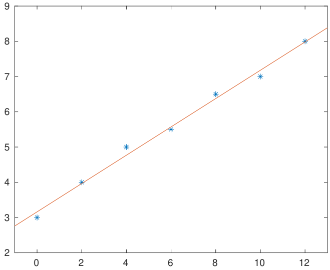


A feltöltés sebessége egyenletes $\Longrightarrow$ a víz magassága az idő lineáris függvénye:

$$F(t) = x_1 + x_2 t,$$

ahol x_1 és x_2 értékét a mérések alapján határozzuk meg.

A méréseink esetlegesen hibával terheltek, így nem biztos, hogy a pontok egy egyenesre illeszkednek.

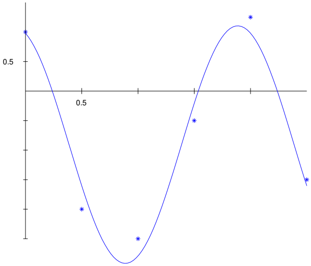


Példa

Megfigyelünk egy periodikus folyamatot, a méréseinkre egy

$$F(t) = x_1 + x_2 \cos \pi t + x_3 \sin \pi t$$

alakú modellt szeretnénk illeszteni, ahol x_1, x_2, x_3 értékét a mérések alapján határozzuk meg.



Mérési hibák miatt a modell nem biztos, hogy pontosan illeszkedik az adatokra.

Hogyan válasszuk meg a modell paramétereit, ha az adataink esetlegesen hibával terheltek?

t_i:	az i-edik megfigyelési hely
f_i:	az i-edik helyen megfigyelt érték
$F(t_i)$:	a modellünk értéke az i-edik helyen

Az i-edik helyen a modellünk értékének és a megfigyelt értéknek a négyzetes eltérése:

$$(F(t_i) - f_i)^2$$

Olyan paramétereket fogunk választani, melyre ezen négyzetes eltérések összege minimális:

$$\sum_i (F(t_i) - f_i)^2,$$

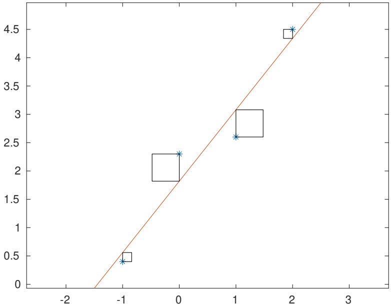
ahol az összegzést az összes megfigyelésre végezzük.

Olyan paramétereket fogunk választani, melyre ezen négyzetes eltérések összege minimális:

$$\sum_i (F(t_i) - f_i)^2,$$

ahol az összegzést az összes megfigyelésre végezzük.

Szemléletesen: négyzetek területösszegét minimalizáljuk



A modell

Olyan modellekkel foglalkozunk, melyek valamilyen adott $\varphi_1(t), \ldots, \varphi_n(t)$ függvények lineáris kombinációi:

$$F(t) = x_1\varphi_1(t) + \cdots + x_n\varphi_n(t) = \sum_{j=1}^{n} x_j\varphi_j(t)$$

Példa

$$F(t) = x_1 \cdot \underbrace{1}_{\varphi_1(t)} + x_2 \underbrace{t}_{\varphi_2(t)}$$

$$F(t) = x_1 \cdot \underbrace{1}_{\varphi_1(t)} + x_2 \underbrace{\cos \pi t}_{\varphi_2(t)} + x_3 \underbrace{\sin \pi t}_{\varphi_3(t)}$$

$$F(t) = x_1 \underbrace{\sin t}_{\varphi_1(t)} + x_2 \underbrace{\sin 2t}_{\varphi_2(t)} + x_3 \underbrace{\sin 3t}_{\varphi_3(t)}$$

Legkisebb négyzetes közelítések

Adott m mérés:
a $t_1, t_2, \ldots, t_m$ helyeken az
a $f_1, f_2, \ldots, f_m$ megfigyelések.

A folyamatot leíró

$$F(t) = \sum_{j=1}^{n} x_j \varphi_j(t)$$

modell n darab ismeretlen paraméterét keressük úgy, hogy

$$J(x) = \sum_{i=1}^{m} (F(t_i) - f_i)^2$$

minimális legyen. Tipikusan $m \gg n$.

x_j: ismeretlen paraméterek $(j = 1, \ldots, n)$
$\varphi_j(t)$: adott függvények $(j = 1, \ldots, n)$

Legyen

$$A = \begin{bmatrix} \varphi_1(t_1) & \varphi_2(t_1) & \dots & \varphi_n(t_1) \\ \varphi_1(t_2) & \varphi_2(t_2) & \dots & \varphi_n(t_2) \\ \vdots & & & \\ \varphi_1(t_m) & \varphi_2(t_m) & \dots & \varphi_n(t_m) \end{bmatrix} \in \mathbb{R}^{m \times n},$$

$$f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} \in \mathbb{R}^m, \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n,$$

Ekkor

$$Ax = \begin{bmatrix} F(t_1) \\ F(t_2) \\ \vdots \\ F(t_m) \end{bmatrix}$$

A minimalizálandó függvény:

$$J(x) = \sum_{i=1}^{m} (F(t_i) - f_i)^2 = \|Ax - f\|_2^2$$

Minimum csak ott lehet, ahol

$$\frac{\partial J(x)}{\partial x_k} = 0, \quad k = 1, \ldots, n.$$

Ez az alábbi lineáris egyenletrendszerre vezet:

$$A^T A x = A^T f$$

(Gauss-féle normálegyenlet)

Gauss-féle normálegyenlet

$$A^T A x = A^T f$$

- a Gauss-féle normálegyenlet mindig megoldható
- A megoldás a legkisebb négyzetes értelemben legjobban közelítő modell paramétereit adja.
- Ha az A mátrix oszlopvektorai lineárisan függetlenek, akkor a Gauss-féle normálegyenletnek egyetlen megoldása van.

Ha az A oszlopvektorai függőek (az $A^T A$ mátrix szinguláris), akkor végtelen sok megoldás van.

Szingularitás esetén javasolható:

- több adat felvétele
- a modell egyszerűsítése

Példa

Ha az illesztett függvény egy egyenes: $F(t) = x_1 + x_2 t$, akkor $\varphi_1(t) \equiv 1$ és $\varphi_2(t) = t$

$$A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \\ 1 & t_m \end{bmatrix}$$

$$A^T A = \begin{bmatrix} m & \sum_{i=1}^m t_i \\ \sum_{i=1}^m t_i & \sum_{i=1}^m t_i^2 \end{bmatrix}, \qquad A^T f = \begin{bmatrix} \sum_{i=1}^m f_i \\ \sum_{i=1}^m t_i f_i \end{bmatrix}$$

szingularitás: az A oszlopvektorai lineárisan függőek, azaz

$$t_1 = t_2 = \cdots = t_m$$

1. Ha van legalább két különböző t_i érték, akkor a rendszer egyértelműen megoldható. A megoldás a J minimumhelye lesz.

2. Ha $t_1 = t_2 = \cdots = t_m =: t_0$, akkor

$$\begin{bmatrix} m & mt_0 \\ mt_0 & mt_0^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum\limits_{i=1}^{m} f_i \\ t_0 \sum\limits_{i=1}^{m} f_i \end{bmatrix}$$

a 2. egyenlet az első t_0-szorosa $\quad \rightarrow \quad$ végtelen sok megoldás.

$$b = s \in \mathbb{R}, \quad a = \frac{1}{m} \sum_{i=1}^{m} f_i - st_0$$

Ha $b = 0$

$$a = \frac{1}{m} \sum_{i=1}^{m} f_i$$

Példa

Határozzuk meg az alábbi adatokat négyzetesen legjobban közelítő egyenes egyenletét!

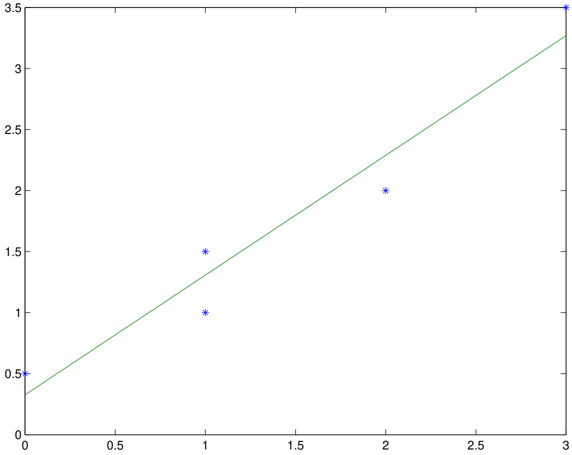
t_i	0	1	1	2	3
f_i	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{7}{2}$

A modell: $F(t) = a + b \cdot t$

$$\begin{bmatrix} 5 & 7 \\ 7 & 15 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \frac{17}{2} \\ 17 \end{bmatrix}$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{26} \begin{bmatrix} 15 & -7 \\ -7 & 5 \end{bmatrix} \begin{bmatrix} \frac{17}{2} \\ 17 \end{bmatrix} = \begin{bmatrix} \frac{17}{52} \\ \frac{51}{52} \end{bmatrix}$$

Az illesztett modell: $F(t) = \frac{17}{52} + \frac{51}{52} t$


3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3

Példa

Határozzuk meg az alábbi adatokat négyzetesen legjobban közelítő egyenes egyenletét!

t_i	2	2	2	2	2
f_i	1	1	2	2	2

A modell: $F(t) = a + b \cdot t$

$$\begin{bmatrix} 5 & 10 \\ 10 & 20 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 8 \\ 16 \end{bmatrix}$$

$$5a + 10b = 8$$

$$b = s \in \mathbb{R}, \qquad a = \frac{8}{5} - 2s$$

Ha $s = 0$, akkor $F(t) \equiv \frac{8}{5}$

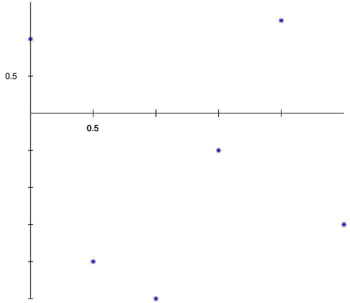
Példa

Határozzuk meg az alábbi adatokat négyzetesen legjobban közelítő

$$F(t) = x_1 + x_2 \cos(\pi t) + x_3 \sin(\pi t)$$

alakú modellt!

t_i	0	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$
f_i	1	-2	$-\frac{5}{2}$	$-\frac{1}{2}$	$\frac{5}{4}$	$-\frac{3}{2}$



$\varphi_1(t) \equiv 1$, $\varphi_2(t) = \cos(\pi t)$, $\varphi_3(t) = \sin(\pi t)$

$$A = \begin{bmatrix} 1 & \cos(\pi t_1) & \sin(\pi t_1) \\ 1 & \cos(\pi t_2) & \sin(\pi t_2) \\ \vdots & & \\ 1 & \cos(\pi t_6) & \sin(\pi t_6) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$A^T A = \begin{bmatrix} 6 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & 3 \end{bmatrix}, \qquad A^T f = \begin{bmatrix} -\frac{17}{4} \\ \frac{19}{4} \\ -3 \end{bmatrix}$$

Az $A^T A x = A^T f$ Gauss-féle normálegyenlet megoldása:

$$x = \begin{bmatrix} -\frac{29}{32} \\ \frac{181}{96} \\ -\frac{67}{96} \end{bmatrix}$$

Példa

Határozzuk meg az alábbi adatokat négyzetesen legjobban közelítő

$$F(t) = x_1 + x_2 \cos(\pi t) + x_3 \sin(\pi t)$$

alakú modellt!

t_i	0	$\frac{1}{2}$	2	$\frac{5}{2}$
f_i	1	-2	$\frac{5}{4}$	$-\frac{3}{2}$

Az előző példából:

t_i	0	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$
f_i	1	-2	$-\frac{5}{2}$	$-\frac{1}{2}$	$\frac{5}{4}$	$-\frac{3}{2}$

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

$$\begin{bmatrix} 1 \\ -2 \\ -\frac{5}{2} \\ -\frac{1}{2} \\ \frac{5}{4} \\ -\frac{3}{2} \end{bmatrix} \to f = \begin{bmatrix} 1 \\ -2 \\ \frac{5}{4} \\ -\frac{3}{2} \end{bmatrix}$$

Az A oszlopai lineárisan függőek $\to A^T A$ szinguláris

A szingularitás kezelése:

1. több adat felvétele (ld. előző példa)
2. a modell egyszerűsítése:

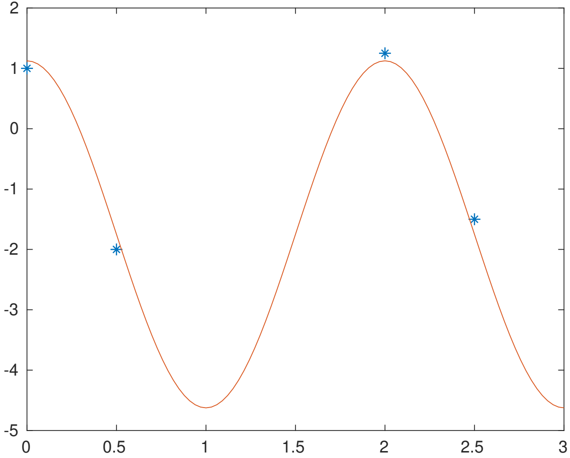
$$F(t) = x_1 + x_2 \cos(\pi t)$$

Ekkor

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}, \qquad f = \begin{bmatrix} 1 \\ -2 \\ \frac{5}{4} \\ -\frac{3}{2} \end{bmatrix}$$

Az $A^T A x = A^T f$ Gauss-féle normálegyenlet megoldása:

$$x = \begin{bmatrix} -1.7500 \\ 2.8750 \end{bmatrix}$$



2
1
0
-1
-2
-3
-4
-5
0
0.5
1
1.5
2
2.5
3

Példa

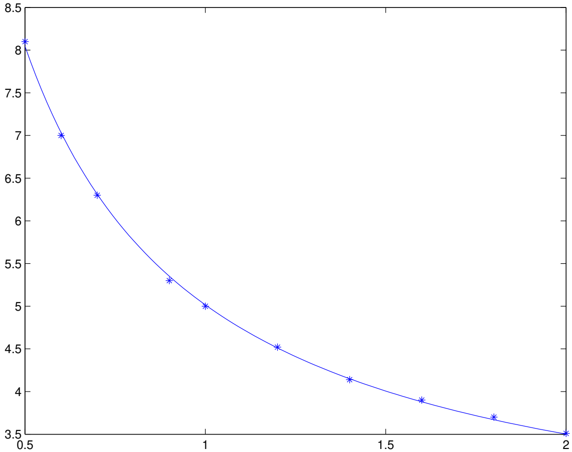
Határozzuk meg az alábbi adatokat négyzetesen legjobban közelítő $F(t) = a + \frac{b}{t}$ alakú modell paramétereit!

t_i	0.5	0.6	0.7	0.9	1	1.2	1.4	1.6	1.8	2
f_i	8.1	7	6.3	5.3	5	4.52	4.14	3.9	3.7	3.51

$$m = 10, \qquad A = \begin{bmatrix} 1 & \frac{1}{t_1} \\ 1 & \frac{1}{t_2} \\ \vdots & \\ 1 & \frac{1}{t_{10}} \end{bmatrix}$$

$$A^T A = \begin{bmatrix} 10 & \sum_{i=1}^{10} \frac{1}{t_i} \\ \sum_{i=1}^{10} \frac{1}{t_i} & \sum_{i=1}^{10} \frac{1}{t_i^2} \end{bmatrix}, \qquad A^T f = \begin{bmatrix} \sum_{i=1}^{10} f_i \\ \sum_{i=1}^{10} \frac{f_i}{t_i} \end{bmatrix}$$

Megj.: Az $\left(\frac{1}{t_i}, f_i\right)$ adatokra illesztettünk egyenest.

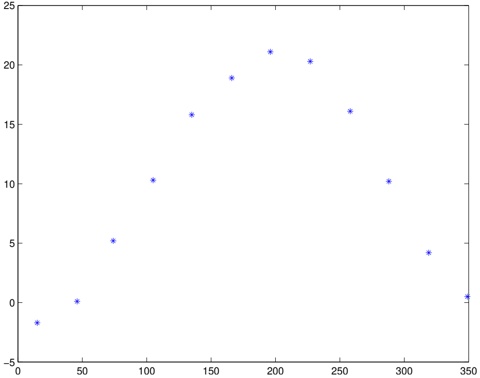


8.5
8
7.5
7
6.5
6
5.5
5
4.5
4
3.5
0.5
1
1.5
2

Példa

Havi középhőmérsékletek átlagai Budapesten (1901-1950)

t_i	15	46	74	105	135	166	196	227	258	288	319	349
f_i	-1.7	0.1	5.2	10.3	15.8	18.9	21.1	20.3	16.1	10.2	4.2	0.5



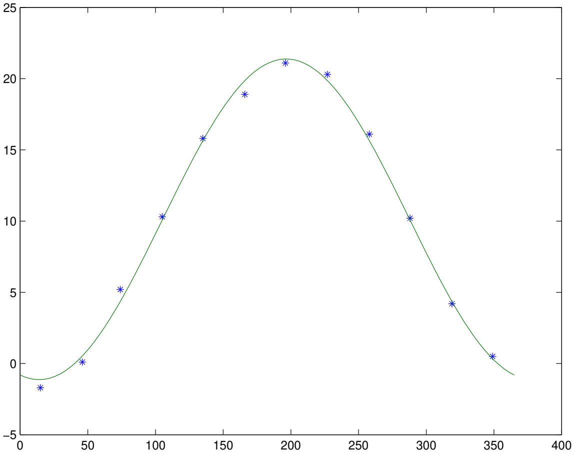
A modell:

$$F(t) = x_1 + x_2 \cos\left(2\pi \frac{t-14}{365}\right)$$

$$A = \begin{bmatrix} 1 & \cos\left(2\pi \frac{t_1-14}{365}\right) \\ 1 & \cos\left(2\pi \frac{t_2-14}{365}\right) \\ \vdots & \\ 1 & \cos\left(2\pi \frac{t_{12}-14}{365}\right) \end{bmatrix}$$

Az $A^T A x = A^T f$ Gauss-féle normálegyenlet megoldása (4 tizedesjegyre kerekítve):

$$x = \begin{bmatrix} 10.1248 \\ -11.2577 \end{bmatrix}$$


25
20
15
10
5
0
−5
0
50
100
150
200
250
300
350
400

Példa

(Matlab, carsmall adathalmaz) 93 autó esetén adott a lóerő, a súly és a gyorsulás értéke. Ezekből az adatokból szeretnénk megbecsülni, hogy az autó 1 gallon üzemanyaggal hány mérföldet tud megtenni (MPG). Feltételezzük, hogy az MPG érték a felsorolt jellemzők lineáris függvénye. Írjuk le ezt a kapcsolatot!

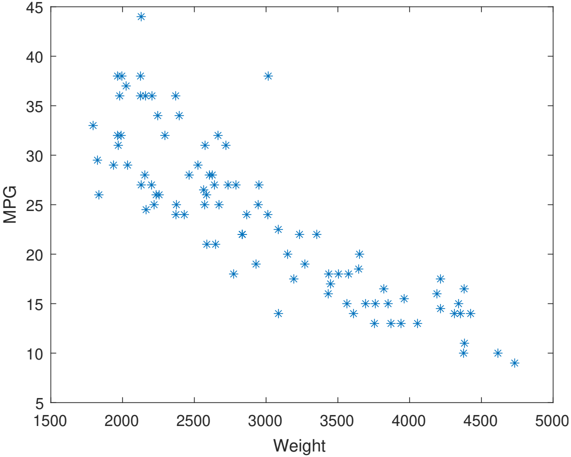
Legyen
$\varphi_1(t)$ a t autó esetén a lóerő
$\varphi_2(t)$ a t autó súlya
$\varphi_3(t)$ a t autó esetén a gyorsulás
$F(t)$ a t autó esetén a MPG érték

Kezdjük egy egyszerű modellel:

$$F(t) \approx x_1 + x_2\varphi_2(t),$$

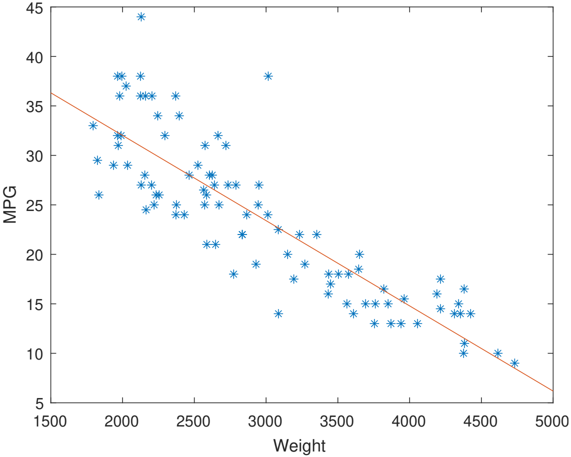
azaz az MPG értéket csak a súly függvényében vizsgáljuk.

Ábrázoljuk a (súly, MPG) párokat!



Látjuk, hogy a két érték között negatív kapcsolat van (minél nagyobb a súly, annál kevesebb mérföldet tud megtenni 1 gallon benzinnel).

Illesszünk egyenest a (súly, MPG) adatokra!



Az illesztett egyenes paraméterei: $x_1 = 49.2383$, $x_2 = -0.0086$.
A négyzetes eltérések összege: 1572.6

Próbálkozzunk egy bonyolultabb modellel:

$$F(t) \approx x_1 + x_2\varphi_1(t) + x_3\varphi_2(t),$$

azaz a lóerő és a súly segítségével becsüljük az MPG értéket.

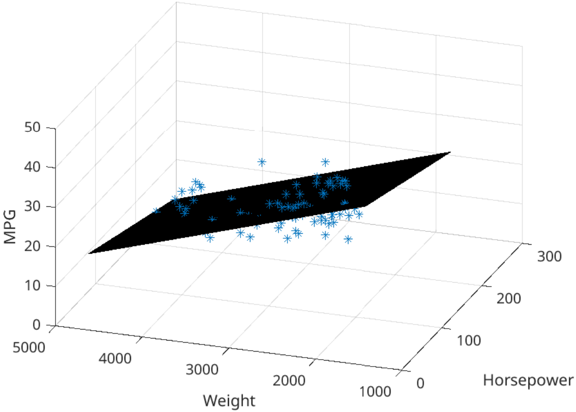
Az $A^T Ax = A^T f$ Gauss-féle normálegyenletet kell megoldanunk, ahol az A mátrixnak most 3 oszlopa van:

1. oszlop: az azonosan 1 vektor
2. oszlop: a lóerő értékek vektora
3. oszlop: a súly értékek vektora

Az f oszlopvektor az MPG értékek vektora

Az egyenlet megoldása után:

$$x_1 = 47.769, \quad x_2 = -0.042018, \quad x_3 = -0.0065651$$



A négyzetes eltérések összege: 1488.9

Ha mindhárom jellemzőt (lóerő, súly, gyorsulás) figyelembe vesszük a becslésnél:

$$F(t) \approx x_1 + x_2\varphi_1(t) + x_3\varphi_2(t) + x_4\varphi_3(t),$$

akkor az A mátrix 4 oszlopból áll:

1. oszlop: az azonosan 1 vektor
2. oszlop: a lóerő értékek vektora
3. oszlop: a súly értékek vektora
4. oszlop: a gyorsulás értékek vektora.

Az egyenlet megoldása után:

$$x_1 = 47.9768, \quad x_2 = -0.0429, \quad x_3 = -0.0065, \quad x_4 = -0.0116,$$

A négyzetes eltérések összege: 1488.8
(A javulás az előző modellhez képest minimális.)

Megjegyzés: a négyzetes hiba helyett gyakran az átlagos négyzetes hibát használjuk (így a hiba nem függ az adatok számától):

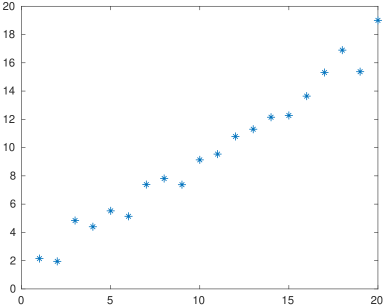
$$MSE = \frac{1}{m}\sum_{i=1}^{m}(F(t_i) - f_i)^2,$$

vagy ennek a négyzetgyökét (így a hibát és a megfigyeléseket ugyanazon a skálán mérjük):

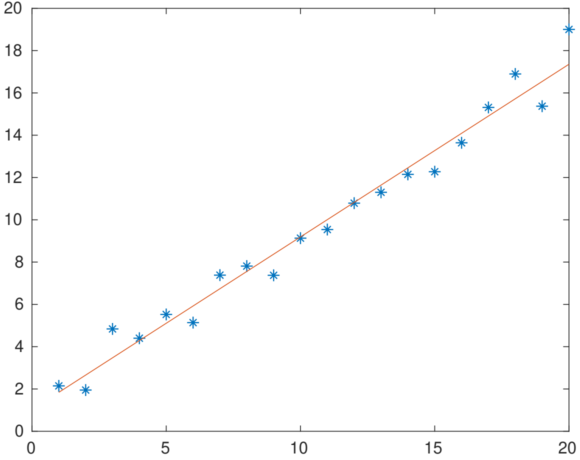
$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(F(t_i) - f_i)^2}.$$

Példa

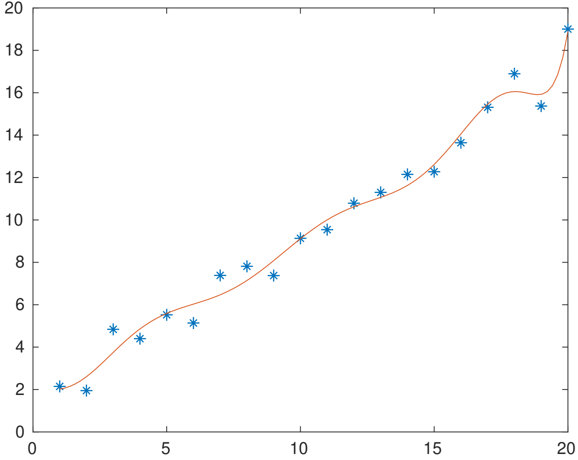
Tegyük fel, hogy megfigyelünk egy folyamatot, amely az $F(t) = 1.1 + 0.8t$ modellel írható le. ,,Felejtsük el" a modellt, és végezzünk méréseket a $t = 1, \dots, 20$ helyeken. A méréseink hibával terheltek, így az ábrán látható megfigyeléseket végeztük. Vizsgáljuk meg mi történik, ha modellt illesztünk a megfigyeléseinkre, de tegyük fel, hogy nincs elképzelésünk az illesztendő modellről, így a négyzetes hiba minimalizálása érdekében különböző fokszámú polinomokkal próbálkozunk. Minden esetben számoljuk ki az átlagos négyzetes hibát.



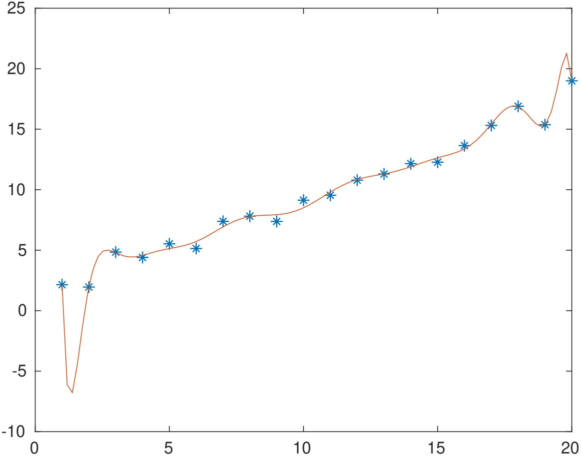
Ha egyenest illesztünk, akkor a megfigyelési helyeken az átlagos négyzetes hiba: 0.5992.



Ha egy kilencedfokú polinomot illesztünk, akkor a megfigyelési helyeken az átlagos négyzetes hiba: 0.3166.



Ha egy tizenötödfokú polinomot illesztünk, akkor a megfigyelési helyeken az átlagos négyzetes hiba: 0.0890.



Melyik a legjobb modell???

A megfigyelt értékekre a harmadik illeszkedik a legjobban, de a megfigyelt folyamatot mégsem jól írja le: rossz az általánosító képessége.

Vizsgáljuk meg az illesztett modellek értékét $t_a = 1.5$-ben és $t_b = 19.2$-ben, és hasonlítsuk össze az „elméleti értékkel" (amiből az adatokat generáltuk).

A polinom fokszáma	MSE (a megfigyelési helyeken)	az abszolút eltérés 1.5-ben	az abszolút eltérés 19.2-ben
1	0.5992	0.0509	0.2431
9	0.3166	0.0860	0.4311
15	0.0890	7.79	0.2469

Ha kellően sok adat áll rendelkezésre és kérdés, hogy milyen modellt válasszunk, akkor érdemes az adatainkat két részre bontani, tanuló- és tesztadatokra. A tanulóadatokra illesztjük a modellt, de az átlagos négyzetes hibát a tesztadatokon is mérjük, ez mutatja a modell általánosító képességét.

Numerikus matematika

Baran Ágnes

Előadás
Interpoláció

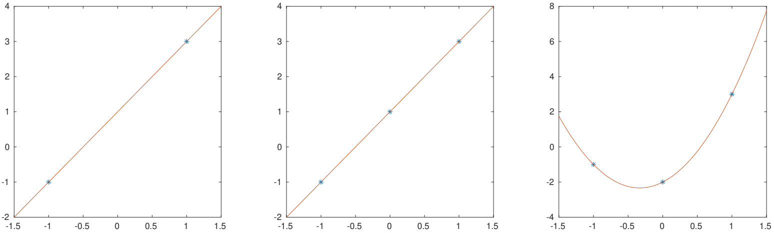
Lagrange-interpoláció

Síkbeli pontokra szeretnénk polinomot illeszteni.

A legkisebb négyzetes közelítéssel ellentétben most

- feltételezzük, hogy az adataink hibamentesek, azaz az illesztett függvénynek pontosan illeszkednie kell a pontokra.
- minden esetben polinomot illesztünk

A lehető legkisebb fokszámú polinomot keressük



Lagrange-interpoláció

A feladat:

Adott $n+1$ pont:
$x_0, x_1, \ldots, x_n$ páronként különböző helyeken az
$f_0, f_1, \ldots, f_n$ értékek.

Olyan minimális fokszámú $\varphi(x)$ polinomot keresünk melyre

$$\varphi(x_i) = f_i, \quad i = 0, \ldots, n$$

Állítás: Egyértelműen létezik olyan legfeljebb n-edfokú polinom, amely teljesíti a

$$\varphi(x_i) = f_i, \quad i = 0, \ldots, n$$

illeszkedési feltételeket.

$n+1$ pont $\implies$ legfeljebb n-edfokú polinom

A Lagrange-polinom rekurzív előállítása (Newton-alak)

Jelölje $L_k(x)$ az (x_0, f_0), (x_1, f_1),..., (x_k, f_k) adatokra illeszkedő Lagrange-polinomot.

- ha csak 1 adat ismert, (x_0, f_0):

  $L_0(x) \equiv f_0$

- ha 2 adat ismert, (x_0, f_0), (x_1, f_1):

  $L_1(x) = L_0(x) + b_1(x - x_0)$

  Ekkor $L_1(x_0) = L_0(x_0) = f_0$. Ezután b_1-et úgy határozzuk meg, hogy $L_1(x_1) = f_1$ teljesüljön:

$$L_1(x_1) = f_0 + b_1(x_1 - x_0) = f_1$$

$$b_1 = \frac{f_1 - f_0}{x_1 - x_0}$$

- ha 3 adat ismert, (x_0, f_0), (x_1, f_1), (x_2, f_2):

  $L_2(x) = L_1(x) + b_2(x - x_0)(x - x_1)$

  Ekkor
  $L_2(x_0) = L_1(x_0) = f_0$ és
  $L_2(x_1) = L_1(x_1) = f_1$.

  b_2-t úgy határozzuk meg, hogy $L_2(x_2) = f_2$ teljesüljön:

$$b_2 = \frac{1}{x_2 - x_0}\left(\frac{f_2 - f_1}{x_2 - x_1} - \frac{f_1 - f_0}{x_1 - x_0}\right)$$

- Ha $k+1$ adat ismert, (x_0, f_0), (x_1, f_1),..., (x_k, f_k):

  $L_k(x) = L_{k-1}(x) + b_k\omega_k(x)$

  ahol $\omega_k(x) = \prod_{i=0}^{k-1}(x - x_i)$.

  Ekkor
  $L_k(x_0) = L_{k-1}(x_0) = f_0,$
  $L_k(x_1) = L_{k-1}(x_1) = f_1,$
  $\vdots$
  $L_k(x_{k-1}) = L_{k-1}(x_{k-1}) = f_{k-1}.$

  b_k-t úgy határozzuk meg, hogy $L_k(x_k) = f_k$ teljesüljön:

  $$b_k = (f_k - L_{k-1}(x_k))/\omega_k(x_k)$$

  Hogyan lehet egyszerűen előállítani a b_k együtthatókat?

Osztott differenciák

Tfh adottak az $x_0, x_1, \dots, x_n$ páronként különböző alappontok és az $f_0, f_1, \dots, f_n$ értékek.
Az x_i, x_{i+1} pontokra támaszkodó elsőrendű osztott differencia:

$$[x_i, x_{i+1}]f := \frac{f_{i+1} - f_i}{x_{i+1} - x_i}$$

Az $x_i, \dots, x_{i+k}$ pontokra támaszkodó k-adrendű osztott differencia:

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

Legyen $[x_i]f = f_i$.

Állítás: A Lagrange-polinom Newton-alakjában

$$b_k = [x_0, \dots, x_k]f$$

Számítási séma

$$
\begin{array}{c|ccccc}
x_0 & [x_0]f \searrow & & & & \\
 & & [x_0,x_1]f \searrow & & & \\
x_1 & [x_1]f \nearrow & & [x_0,x_1,x_2]f \searrow & & \\
 & \searrow & [x_1,x_2]f \nearrow & & \dots & \\
\vdots & & \vdots & & \searrow & \\
\vdots & & \vdots & & & [x_0,\dots,x_n]f \\
\vdots & & \vdots & & \dots \nearrow & \\
x_{n-1} & [x_{n-1}]f \searrow & & & & \\
 & & [x_{n-1},x_n]f \nearrow & & & \\
x_n & [x_n]f \nearrow & & & &
\end{array}
$$

A Lagrange-polinom Newton-alakja

$$
\begin{array}{c|cccccccc}
x_0 & {\color{red}[x_0]f} & \searrow & & & & & & \\
 & & & {\color{red}[x_0,x_1]f} & \searrow & & & & \\
x_1 & [x_1]f & \nearrow & & & {\color{red}[x_0,x_1,x_2]f} & \searrow & & \\
 & & \searrow & [x_1,x_2]f & \nearrow & & & \cdots & \\
\vdots & & \vdots & & & & & & \searrow \\
\vdots & & \vdots & & & & & & & {\color{red}[x_0,\ldots,x_n]f} \\
\vdots & & \vdots & & & & & \cdots & \nearrow \\
x_{n-1} & [x_{n-1}]f & \searrow & & & & & & \\
 & & & [x_{n-1},x_n]f & \nearrow & & & & \\
x_n & [x_n]f & \nearrow & & & & & &
\end{array}
$$

$$
\begin{aligned}
L_n(x) = {\color{red}[x_0]f} + {\color{red}[x_0,x_1]f} \cdot (x-x_0) + {\color{red}[x_0,x_1,x_2]f} \cdot (x-x_0)(x-x_1) \\
+ \cdots + {\color{red}[x_0,\ldots,x_n]f} \cdot (x-x_0)\cdots(x-x_{n-1})
\end{aligned}
$$

Példa

Határozzuk meg a $(-2,-31),(-1,-7),(0,-1),(2,5)$ pontokra illeszkedő minimális fokszámú polinomot!

$$
\begin{array}{r|cccccc}
-2 & -31 & & & & & \\
 & & 24 & & & & \\
-1 & -7 & & -9 & & & \\
 & & 6 & & 2 & & \\
0 & -1 & & -1 & & & \\
 & & 3 & & & & \\
2 & 5 & & & & &
\end{array}
$$

$$L_3(x) = -31 + 24(x+2) - 9(x+2)(x+1) + 2(x+2)(x+1)x$$

Megjegyzés

A Lagrange-polinom nem függ az adatok sorrendjétől, így választhattuk volna a táblázat alsó "élét" is:

$$
\begin{array}{r|rrrr}
-2 & -31 & & & \\
 & & 24 & & \\
-1 & -7 & & -9 & \\
 & & 6 & & 2 \\
0 & -1 & & -1 & \\
 & & 3 & & \\
2 & 5 & & &
\end{array}
$$

$$L_3(x) = 5 + 3(x-2) - 1 \cdot (x-2)x + 2(x-2)x(x+1)$$

Mindkét esetben

$$L_3(x) = 2x^3 - 3x^2 + x - 1$$

Példa

Határozzuk meg a $(-2,-5),(-1,3),(1,-5),(2,-9)$ pontokra illeszkedő minimális fokszámú polinomot!

$$
\begin{array}{r|rrrr}
-2 & -5 & & & \\
 & & 8 & & \\
-1 & 3 & & -4 & \\
 & & -4 & & 1 \\
1 & -5 & & 0 & \\
 & & -4 & & \\
2 & -9 & & &
\end{array}
$$

$$L_3(x) = -5 + 8(x+2) - 4(x+2)(x+1) + (x+2)(x+1)(x-1)$$

Példa

Határozzuk meg azt a minimális fokszámú polinomot, amely az előző adatokon kívül a $(0,9)$ pontra is illeszkedik!

Használjuk fel az előző feladat eredményét!

$$
\begin{array}{r|rrrr}
-2 & -5 & & & \\
 & & 8 & & \\
-1 & 3 & & -4 & \\
 & & -4 & & 1 \\
1 & -5 & & 0 & \\
 & & -4 & & \\
2 & -9 & & & \\
\end{array}
$$

Egészítsük ki a táblázatot az új adattal és számítsuk ki a hiányzó értékeket!

$$
\begin{array}{r|rrrrr}
-2 & -5 & & & & \\
 & & 8 & & & \\
-1 & 3 & & -4 & & \\
 & & -4 & & 1 & \\
1 & -5 & & 0 & & 2 \\
 & & -4 & & 5 & \\
2 & -9 & & 5 & & \\
 & & -9 & & & \\
\color{green}{0} & \color{green}{9} & & & &
\end{array}
$$

$$L_4(x) = L_3(x) + 2(x+2)(x+1)(x-1)(x-2)$$

Feladat

Határozza meg az alábbi adatokra illeszkedő minimális fokszámú polinomot.

(a)

x_i	-2	-1	1	2
f_i	-21	-1	3	23

(b)

x_i	-2	-1	0	2
f_i	-7	1	1	25

(c)

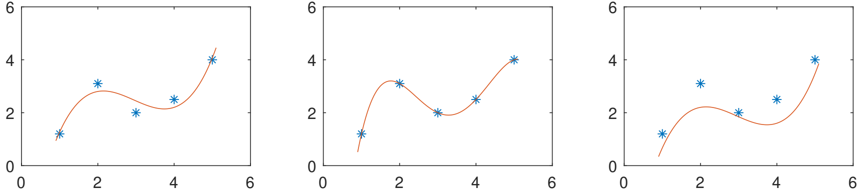
x_i	-2	-1	0	2
f_i	-14	-3	2	-6

Példa

Matlab-ban definiáltuk az x és y változókat, majd lefuttattuk az alábbi két kódot. Melyik kód melyik ábrát állította elő?

```
p=polyfit(x,y,4);
xx=linspace(0.9,5.1);
yy=polyval(p,xx);
figure; plot(x,y,'*',xx,yy)
```

```
p=polyfit(x,y,3);
xx=linspace(0.9,5.1);
yy=polyval(p,xx);
figure; plot(x,y,'*',xx,yy)
```



Horner-algoritmus

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

ahol $a_n \neq 0$. Legyen $x^\star \in \mathbb{R}$ adott, $p(x^\star) = ?$

$$p(x^\star) = (((\cdots(a_n x^\star + a_{n-1})x^\star + \cdots)x^\star + a_2)x^\star + a_1)x^\star + a_0$$

Az algoritmus:

$$\begin{aligned}
c_0 &= a_n \\
c_1 &= c_0 x^\star + a_{n-1} \\
c_2 &= c_1 x^\star + a_{n-2} \\
&\vdots \\
c_n &= c_{n-1} x^\star + a_0 = p(x^\star)
\end{aligned}$$

Táblázatban:

	a_n	a_{n-1}	$\cdots$	a_2	a_1	a_0
$x^\star$	c_0	c_1	$\cdots$	c_{n-2}	c_{n-1}	c_n

$p(x^\star) = c_n$

Példa

$$p(x) = 2x^5 + 3x^4 - 3x^2 + 5x - 1, \qquad p(-2) = ?$$

	2	3	0	-3	5	-1
-2	2	-1	2	-7	19	-39

$p(-2) = -39$

Általánosított Horner-algoritmus

$$\begin{aligned}L_n(x) = &b_0 + b_1 \cdot (x - x_0) + b_2 \cdot (x - x_0)(x - x_1) + \\ &+ \cdots + b_n \cdot (x - x_0)(x - x_1)(x - x_{n-1})\end{aligned}$$

ahol $b_k = [x_0, \ldots, x_k]f$. $L_n(x^\star) = ?$

$$\begin{aligned}c_0 &= b_n \\ c_1 &= c_0(x^\star - x_{n-1}) + b_{n-1} \\ c_2 &= c_1(x^\star - x_{n-2}) + b_{n-2} \\ &\vdots \\ c_n &= c_{n-1}(x^\star - x_0) + b_0 = L_n(x^\star)\end{aligned}$$

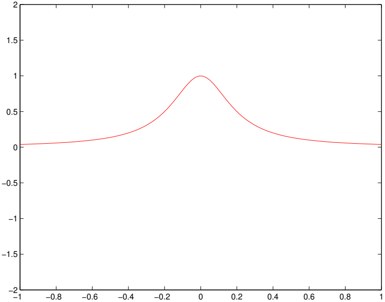
Megjegyzés

Ha nincs szükségünk a Lagrange-polinom együtthatóira, csak bizonyos helyeken a polinom értékeire, akkor nem érdemes a Newton-alakban kibontani a zárójeleket.

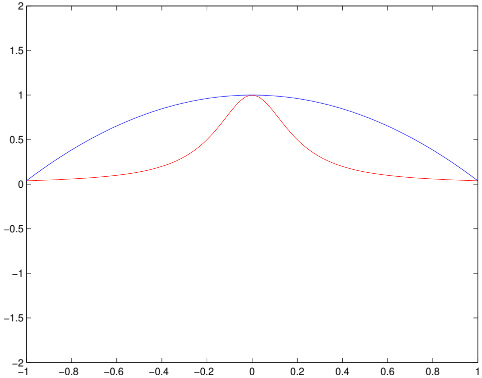
Megjegyzés

Ha egy függvényt szeretnénk közelíteni úgy, hogy elkészítjük adott alappontok esetén az illeszkedő Lagrange-polinomot, akkor az alappontok számának növelésével a hiba nem feltétlenül csökken, sőt akár tetszőlegesen naggyá válhat.

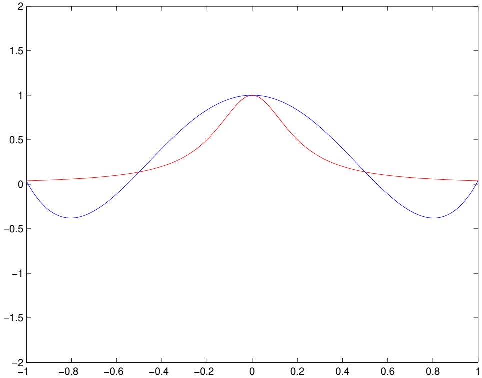
Példa: Az $f(x) = \frac{1}{1+25x^2}$ függvény $[-1, 1]$ fölött



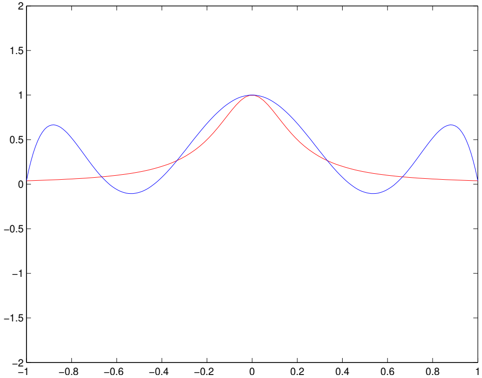
Lagrange-interpoláció, $n = 2$



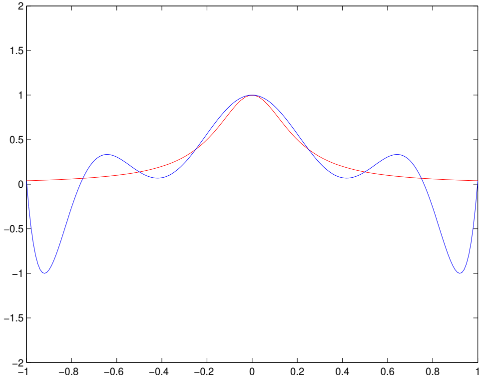
Lagrange-interpoláció, $n = 4$



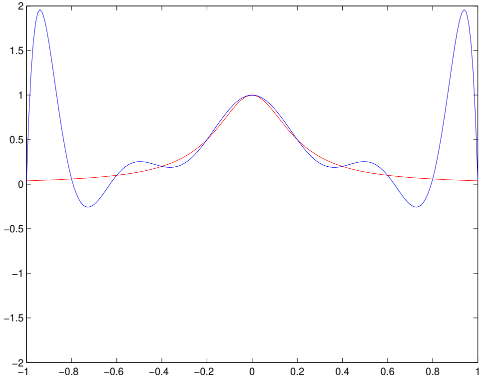
Lagrange-interpoláció, $n = 6$



Lagrange interpoláció, $n = 8$



Lagrange-interpoláció, $n = 10$

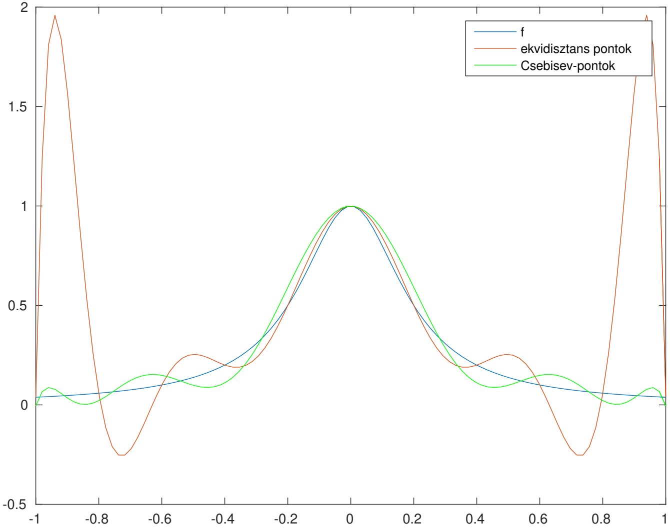


Megjegyzés

Ha az $f : [-1, 1] \to \mathbb{R}$ függvényre n helyen illeszkedő Lagrange-polinomot szeretnénk elkészíteni, akkor az

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

alappontok (Csebisev-pontok) esetén lesz minimális a polinom és a függvény legnagyobb eltérése.
Ha f nem a $[-1, 1]$ intervallumon értelmezett, akkor megfelelő lineáris transzformációval leképezzük a pontokat a megadott intervallumra.


f
ekvidisztans pontok
Csebisev-pontok
2
1.5
1
0.5
0
-0.5
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

Numerikus matematika

Baran Ágnes

Gyakorló feladatok

1. feladat

Adja meg a következő számok kettes számrendszerbeli alakját!

$$143, \quad 85, \quad 1.9375, \quad 0.40625, \quad 7.5625$$

2. feladat

$a = 2$, $t = 4$, $k_- = -6$, $k_+ = 6$ számábrázolási jellemzők mellett mi lesz a 0.15, illetve a 0.55 lebegőpontos alakja szabályos kerekítés, illetve levágás esetén? Mi lesz a 3 jobboldali lebegőpontos szomszédja? Definiálja a gépi epszilont és adja meg az értékét.

3. feladat

Legfeljebb mekkora lehet a megoldás relatív hibája (∞-normában) az $Ax = b$ rendszer megoldásakor a lent adott A és b esetén, ha A-ról tudjuk, hogy pontosan adott, míg a b vektor relatív hibája (∞-normában legfeljebb) $0.5 \cdot 10^{-4}$?

$$A = \begin{bmatrix} 2 & -2 \\ -6 & 4 \end{bmatrix}, \quad b = \begin{bmatrix} -0.982 \\ 1.173 \end{bmatrix}.$$

4. feladat

Adja meg $\|x\|_1$, $\|x\|_2$, $\|x\|_\infty$, $\|A\|_1$, $\|A\|_\infty$ értékét, ha

$$x = \begin{bmatrix} -2 \\ 3 \\ -4 \end{bmatrix}, \quad A = \begin{bmatrix} -4 & 0 & 9 \\ 1 & 3 & 2 \\ -1 & 4 & -5 \end{bmatrix}$$

5. feladat

Oldja meg az $Ax = b$ egyenletrendszereket Matlab-bal, ha

$$A = \begin{bmatrix} 35 & 55 & 10 \\ -14 & -58 & -22 \\ -35 & -75 & -20 \end{bmatrix}, \quad b = \begin{bmatrix} -312 \\ -450 \\ 248 \end{bmatrix}$$

$$A = \begin{bmatrix} 28 & -36 & 8 \\ 14 & -30 & -14 \\ -35 & 41 & -20 \end{bmatrix}, \quad b = \begin{bmatrix} -312 \\ -450 \\ 248 \end{bmatrix}$$

$$A = \begin{bmatrix} 35 & 55 & 10 \\ -14 & -58 & -22 \\ -35 & -75 & -20 \end{bmatrix}, \quad b = \begin{bmatrix} 410 \\ -650 \\ -680 \end{bmatrix}$$

6. feladat

Matlab-ban, az $Ax = b$ egyenletrendszer kibővített mátrixával meghívtuk az `rref` függvényt és a lenti kimenetet kaptuk. Ezek alapján mit mondhatunk az egyenletrendszerről? (Egyenletek száma, ismeretlenek száma, megoldhatóság, megoldások száma, megoldás, az A rangja, a kibővített mátrix rangja?)

(a)

$$\begin{bmatrix} 1 & 0 & -0.1 & 0 \\ 0 & 1 & -0.4 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 1 & 0 & -0.1 & -1.1 \\ 0 & 1 & -0.4 & 1.6 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

7. feladat

Matlab segítségével határozza meg az alábbi adatokra legkisebb négyzetes értelemben legjobban illeszkedő egyenest.

t	1	1	2	3	4
f	2.97	3.20	4.73	6.53	8.10

Ábrázolja az adatokat és az illesztett egyenest.

8. feladat

Matlab segítségével határozza meg az alábbi adatokra legkisebb négyzetes értelemben legjobban illeszkedő másodfokú polinomot.

t	0	1	2	3	4
f	1.50	2.13	1.15	-1.72	-6.07

Ábrázolja az adatokat és az illesztett polinomot.

9. feladat

Milyen értéket vesz fel az alábbi adatokra legkisebb négyzetes értelemben legjobban illeszkedő

$$F(t) = x_1 + x_2\sqrt{1+t^2} + x_3\frac{\sin(\pi t)}{t}$$

alakú modell az 1.6 helyen? Adja meg a modell paramétereit. Válaszait 2 tizedesjegyre kerekítse.

t	1.2	1.3	1.5	1.9	2.0
f	0.81	0.49	0.51	1.86	2.29

Ábrázolja az adatokat és az illesztett függvényt egy közös ábrán!

Numerikus matematika

Baran Ágnes

Nemlineáris egyenletek

Nemlineáris egyenletek

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

Példa:

$$\cos(x) - x = 0$$

vagy

$$x^5 - 3x^4 + x^3 - 5x^2 + 3 = 0$$

vagy

$$e^x - 4x^2 = 0$$

vagy

$$\ln(x) - x + 2 = 0$$

A gyök numerikus közelítése

Az $f(x) = 0$ egyenlet gyökét egy $\{x_k\}$, $k = 0, 1, 2, \ldots$ sorozattal (iteráció) fogjuk közelíteni.

A közelítés adott, ha adott

- az x_0 kiindulópont,
- az algoritmus x_{k+1} meghatározására, ha x_k már ismert,
- a leállási feltétel.

1. Felezési módszer

Tfh $f : [a, b] \to \mathbb{R}$ folytonos, és $f(a) \cdot f(b) < 0$
Ekkor az

$$f(x) = 0$$

egyenletnek van gyöke (a, b)-ben.

Az algoritmus

Adott a maximális iterációszám (maxit) és az ε pontosság.

1. legyen $k = 1$, $x_0 = a$ és $x_1 = b$
2. legyen $x_2 = \frac{x_0 + x_1}{2}$
3. a) ha $f(x_2) = 0$, akkor x_2 gyök $\to$ kilépés (eredmény: x_2)
   b) ha $f(x_0) \cdot f(x_2) < 0$, akkor $x_1 = x_2$
   c) ha $f(x_1) \cdot f(x_2) < 0$, akkor $x_0 = x_2$

   ha $|x_1 - x_0| < \varepsilon \to$ kilépés (eredmény: x_2)

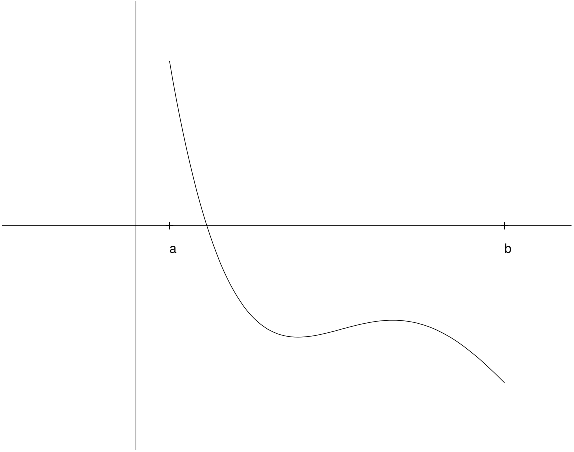
   $k := k + 1$

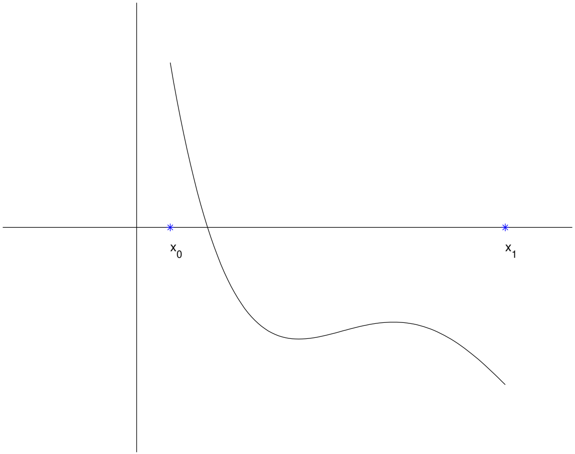
   ha $k = maxit \to$ kilépés (maxit lépésben nem találtunk gyököt)

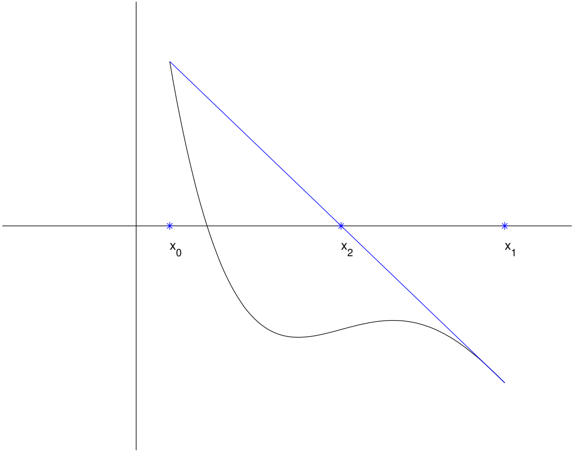
   $\to 2.$

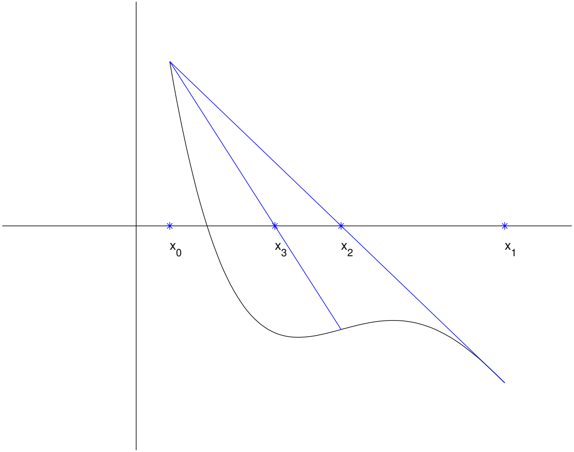
2. Húrmódszer

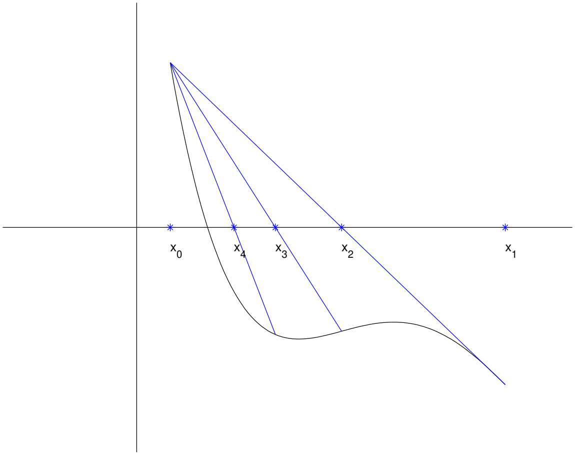
Az $f(x) = 0$ nemlineáris egyenlet gyökét keressük, ahol $f : [a, b] \to \mathbb{R}$, továbbá $f(a) \cdot f(b) < 0$ és f folytonos.

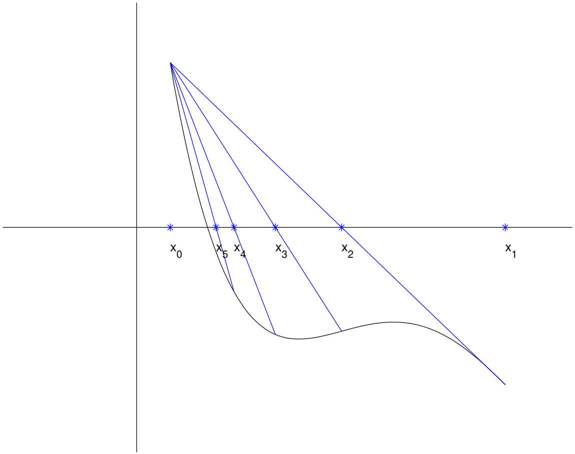



x
0
x
1


x0
x2
x1


x0
x3
x2
x1


x0
x4
x3
x2
x1


x0
x5
x4
x3
x2
x1

Húrmódszer

$x_0 = a$, $x_1 = b$.
Az x_2 pont meghatározása:

Az $(x_0, f(x_0))$ és $(x_1, f(x_1))$ pontokra illeszkedő egyenes egyenlete (Lagrange-interpoláció):

$$\begin{array}{c|cc} x_0 & f(x_0) & \\ & & \frac{f(x_1)-f(x_0)}{x_1-x_0} \\ x_1 & f(x_1) & \end{array}$$

$y(x) = f(x_1) + \frac{f(x_1)-f(x_0)}{x_1-x_0} \cdot (x - x_1)$

Ott metszi az x-tengelyt, ahol $y(x) = 0$:

$$x_2 = x_1 - f(x_1) \cdot \frac{x_1 - x_0}{f(x_1) - f(x_0)}$$

Húrmódszer

x_2 kiszámítása után ismételjük meg az előző lépéseket az $[x_0, x_2]$, illetve $[x_2, x_1]$ intervallumok közül azzal, ahol előjelet vált a függvény.

A húrmódszer esetén

- x_2 kiszámítása jól definiált
- az eljárás minden folytonos f esetén konvergál f egy gyökéhez
- csak páratlan multiplicitású gyök közelítésére
- két pontra támaszkodó iteráció

Az algoritmus:

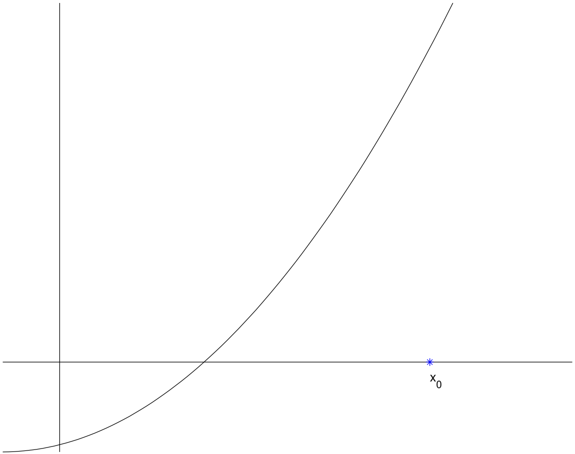
Adott a maximális iterációszám (maxit) és az ε pontosság.

1. $x_0 := a$, $x_1 := b$, $f0 := |f(x_0)|$
2. $$x_2 := x_1 - f(x_1) \cdot \frac{x_1 - x_0}{f(x_1) - f(x_0)}$$
3. a) Ha $f(x_2) = 0$, akkor kilépés (x_2 gyök).
   b) ha $f(x_2) \cdot f(x_1) < 0$, akkor $x_0 = x_2$
   c) ha $f(x_2) \cdot f(x_0) < 0$, akkor $x_1 = x_2$

   ha $|f(x_2)| < \varepsilon * (1 + f0)$, akkor kilépés (eredmény: x_2)
   $k := k + 1$
   ha $k = maxit$, akkor kilépés (maxit lépésben nem találtunk gyököt)
   $\rightarrow 2.$

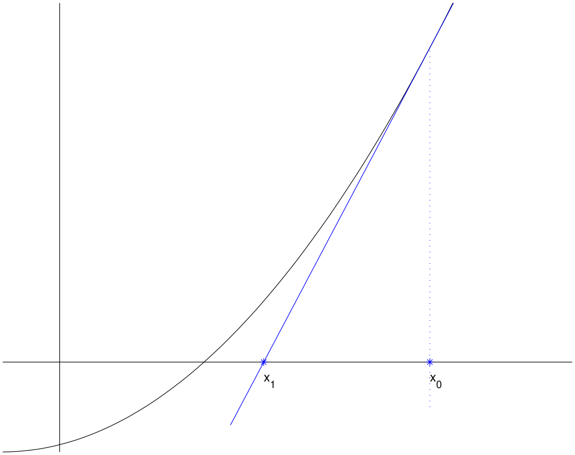
3. Newton-módszer

Az $f(x) = 0$ nemlineáris egyenlet gyökét keressük.



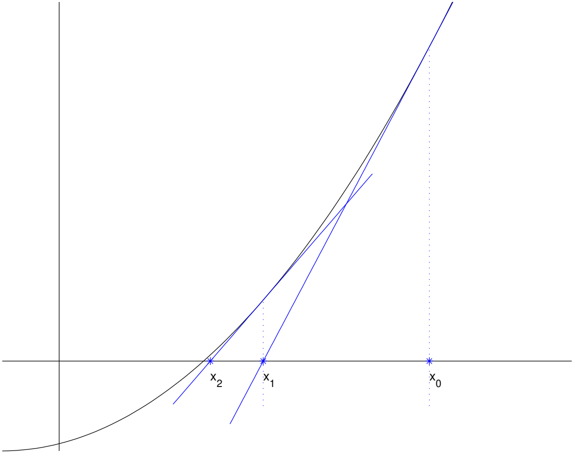
3. Newton-módszer

Az $f(x) = 0$ nemlineáris egyenlet gyökét keressük.



3. Newton-módszer

Az $f(x) = 0$ nemlineáris egyenlet gyökét keressük.



Az algoritmus:

x_0 a gyök egy kezdeti közelítése,
x_{k+1} meghatározása:
Az f függvény x_k-beli érintője (Hermite-interpoláció):

$$\begin{array}{c|cc} x_k & f(x_k) & \\ & & f'(x_k) \\ x_k & f(x_k) & \end{array}$$

$$y(x) = f(x_k) + f'(x_k) \cdot (x - x_k)$$

Ott metszi az x-tengelyt, ahol $y(x) = 0$:

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

A Newton-iteráció:

x_0 kezdőpont,

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \dots$$

- nem feltétlenül definiált
- egy pontra támaszkodó iteráció

Tétel. Legyen x^* az f egy gyöke. Ha

- f kétszer folytonosan diff.ható,
- $|f'(x)| \geq m_1 > 0$,
- $|f''(x)| \leq M_2$,
- $|x_0 - x^*| < \frac{2m_1}{M_2}$,

akkor a Newton-iteráció jól definiált, $x_k \to x^*$, ha $k \to \infty$, továbbá

$$|x_{k+1} - x^*| \leq C|x_k - x^*|^2$$

Mit jelent a gyakorlatban a

$$|x_{k+1} - x^*| \leq C|x_k - x^*|^2$$

becslés?

Ha valamely k-ra $|x_k - x^*| \approx 0.1$, akkor a sorozat következő néhány tagjának a távolsága a gyöktől kb

$$0.01$$
$$0.0001$$
$$0.00000001$$

A Newton-módszer konvergenciája **kvadratikus**, vagy másodrendű.

Példa

Közelítsük az $x^3 - 3x - 2 = 0$ egyenlet gyökét Newton–módszerrel az $x_0 = 1.5$ pontból indulva!

$f(x) = x^3 - 3x - 2$ és $f'(x) = 3x^2 - 3$.

$$x_{k+1} = x_k - \frac{x_k^3 - 3x_k - 2}{3x_k^2 - 3}, \quad k = 0, 1, 2, \ldots$$

$$\begin{aligned}
x_0 &= 1.5\\
x_1 &= 2.33333333333333\\
x_2 &= 2.\underline{0}5555555555556\\
x_3 &= 2.\underline{00}194931773879\\
x_4 &= 2.\underline{00000}252829797\\
x_5 &= 2.\underline{00000000000}426
\end{aligned}$$

2. példa

Közelítsük $\sqrt{a}$, $(a > 0)$ értékét Newton–módszerrel!

$f(x) = x^2 - a$ és $f'(x) = 2x$. Ekkor

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

$a = 5$, $x_0 = 2$ esetén:

$$\begin{aligned}
x_1 &= 2.\underline{2}5 \\
x_2 &= 2.\underline{236}111111111 \\
x_3 &= 2.\underline{23606797}791580 \\
x_4 &= 2.\underline{23606797749979}
\end{aligned}$$

3. példa

Közelítsük az $x^3 - 3x + 2 = 0$ egyenlet gyökét Newton–módszerrel az $x_0 = 1.5$ pontból indulva!

$f(x) = x^3 - 3x + 2$ és $f'(x) = 3x^2 - 3$.

$$x_{k+1} = x_k - \frac{x_k^3 - 3x_k + 2}{3x_k^2 - 3}, \quad k = 0, 1, 2, \ldots$$

$$\begin{aligned}
x_0 &= 1.5 \\
x_1 &= 1.26666666666667 \\
x_2 &= 1.13856209150327 \\
x_3 &= 1.\underline{0}7077733565581 \\
x_4 &= 1.\underline{0}3579185227111 \\
\ldots \\
x_9 &= 1.\underline{00}113136084711
\end{aligned}$$

Hasonlítsuk össze az eredmény az 1. példa eredményével! Bár az egyenlet gyökéhez konvergál a sorozat, de a konvergencia nem kvadratikus. Miért?

A probléma: az 1 kétszeres gyöke f-nek (a konvergenciatétel 2. feltétele nem teljesül).

Ha az

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

iteráció helyett az

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

iterációt alkalmazzuk:

$$\begin{aligned} x_0 &= 1.5 \\ x_1 &= 1.\underline{0}3333333333333 \\ x_2 &= 1.\underline{000}18214936248 \\ x_3 &= 1.\underline{00000000}552926 \end{aligned}$$

A Newton-iteráció nem feltétlenül konvergál, ezért fontos, hogy programozásakor az $\{x_k\}$ sorozatot legfeljebb egy megadott `maxit` iterációszámig határozzuk meg.

4. példa

Vizsgáljuk meg mi történik, ha a Newton-módszert az $f(x) = x^3 - 5x$ függvény gyökének közelítésére alkalmazzuk az $x_0 = 1$ pontból indulva!

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^3 - 5x_k}{3x_k^2 - 5}$$

$$x_0 = 1, \quad x_1 = -1, \quad x_2 = 1, \ldots$$

4. Szelőmódszer

A Newton-iteráció minden lépésében szükséges a derivált adott pontbeli értéke.

Ha a derivált számítása nem lehetséges, vagy túl költséges, akkor az $f'(x_k) \approx [x_{k-1}, x_k]f$ közelítést alkalmazhatjuk.

$$x_{k+1} = x_k - \frac{f(x_k)}{[x_{k-1}, x_k]f} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}$$

Ez a **szelőmódszer**.

Szelőmódszer

x_0, x_1 kezdőpontok,

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}, \quad k = 1, 2, \ldots$$

A képlet hasonló a húrmódszerhez, de itt nem vizsgáljuk az új pontban a függvény előjelét, mindig a 2 utolsó pontból számítjuk a következőt.

- a képlet nem feltétlenül definiált ($f(x_k) = f(x_{k-1})$ lehet)
- 2 pontra támaszkodó

Szelőmódszer

Konvergencia feltételei ugyanazok, mint a Newton-iterációnál, csak még $|x_1 - x^*| < \frac{2m_1}{M_2}$ is kell.

A konvergenciarend alacsonyabb, mint a Newton-iterációnál:

$$|x_{k+1} - x^*| \le C|x_k - x^*|^p,$$

ahol $p = \frac{1+\sqrt{5}}{2} \approx 1.618$.

(Húrmódszernél $p = 1$, Newton-módszernél $p = 2$.)

5. Fixpont-iteráció.

$g(x) = x$ gyökét keressük, ahol $g : [a, b] \to \mathbb{R}$.

Az algoritmus:
x_0 kezdőpont, $x_{k+1} = g(x_k)$, $k = 0, 1, \dots$

Tétel

Ha $g([a, b]) \subseteq [a, b]$, és $\exists \quad 0 \leq \alpha < 1$:

$$|g(x) - g(y)| \leq \alpha \cdot |x - y| \quad \forall x, y \in [a, b], \tag{1}$$

akkor egyértelműen létezik olyan $x^* \in [a, b]$, hogy $g(x^*) = x^*$, továbbá $\forall x_0 \in [a, b]$ esetén az $x_{k+1} = g(x_k)$, $k = 0, 1, \dots$ sorozat tart x^*-hoz.

Megjegyzés: Ha $|g'(x)| \leq \alpha < 1$, akkor (1) teljesül.

Megjegyzés

Ha egy g függvény teljesíti az (1) tulajdonságot, akkor összehúzó leképezésnek (kontrakciónak) nevezzük.

Feladat

Mutassa meg, hogy az $f(x) = e^x - 4x^2$ függvénynek van zérushelye a $[0, 1]$ intervallumban! Igazolja, hogy az

$$x_{k+1} = \frac{1}{2} e^{\frac{x_k}{2}}, \quad k = 0, 1, \ldots$$

iteráció tetszőleges $x_0 \in [0, 1]$ kezdőpont esetén tart ehhez a gyökhöz!

Példa

Az

$$xe^x - 1 = 0, \quad x \in [0.25, 1]$$

egyenlet gyökét szeretnénk közelíteni fixpont-iterációval. Vizsgáljuk meg az

$$x_0 = 0.5, \quad x_{k+1} = g(x_k), \quad k = 0, 1, \ldots$$

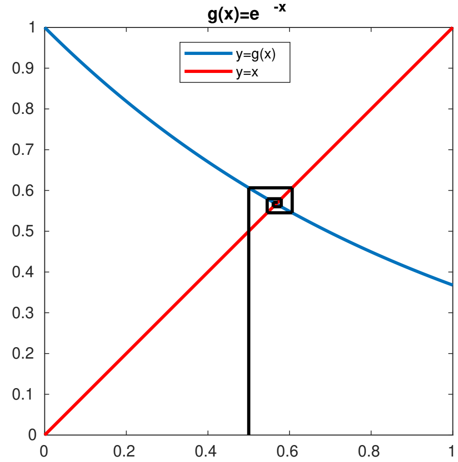
iteráció konvergenciáját, ha

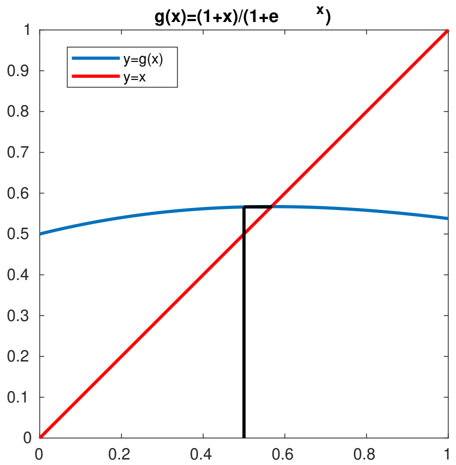
(a) $g(x) = e^{-x}$

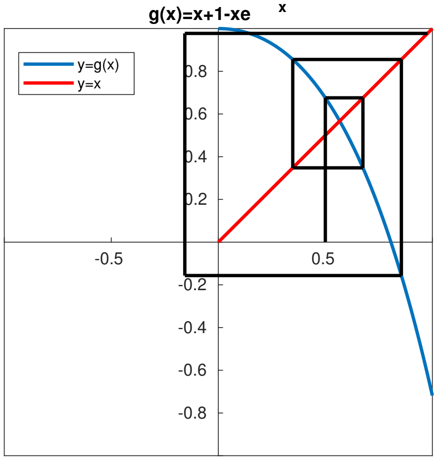
(b) $g(x) = \frac{1+x}{1+e^x}$

(c) $g(x) = x + 1 - xe^x$

	$g(x) = e^{-x}$	$g(x) = \frac{1+x}{1+e^x}$	$g(x) = x + 1 - xe^x$
$x^{(1)}$	0.60653	0.56631	0.67564
$x^{(2)}$	0.54524	0.56714	0.34781
$x^{(3)}$	0.57970	0.56714	0.85532
$x^{(4)}$	0.56006	0.56714	-0.15651
$x^{(5)}$	0.57117	0.56714	0.97733
$x^{(6)}$	0.56486	0.56714	-0.61976
$x^{(7)}$	0.56844	0.56714	0.71371
$x^{(8)}$	0.56641	0.56714	0.25663
$x^{(9)}$	0.56756	0.56714	0.92492
$x^{(10)}$	0.56691	0.56714	-0.40742


g(x)=e -x
y=g(x)
y=x
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1


g(x)=(1+x)/(1+e x)
y=g(x)
y=x
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1



g(x)=x+1-xe^x
y=g(x)
y=x
0.8
0.6
0.4
0.2
-0.5
0.5
-0.2
-0.4
-0.6
-0.8

Nemlineáris egyenletrendszerek.

$f(x) = 0$, ahol $f : \mathbb{R}^n \to \mathbb{R}^n$. Másképpen:

$$\begin{aligned} f_1(x_1, \dots, x_n) &= 0 \\ f_2(x_1, \dots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, \dots, x_n) &= 0 \end{aligned}$$

Példa: $f : \mathbb{R}^2 \to \mathbb{R}^2$,

$$\begin{aligned} x_1^2 + x_1 x_2 - 3x_2 + 3 &= 0 \\ -3x_1^2 + x_2^2 - x_1 &= 0 \end{aligned}$$

Newton-módszer több dimenzióban

$x^{(0)}$ kezdővektor,

$$x^{(k+1)} = x^{(k)} - \left(J(x^{(k)})\right)^{-1} \cdot f(x^{(k)}), \quad k = 0, 1, \dots,$$

ahol J a Jacobi-mátrix:

$$J(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \vdots & & & \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \dots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

A mátrixinvertálás helyett: a

$$J(x^{(k)}) \cdot \underbrace{\left(x^{(k+1)} - x^{(k)}\right)}_{\delta x :=} = -f(x^{(k)})$$

lineáris egyenletrendszert oldjuk meg. Ezután

$$x^{(k+1)} = x^{(k)} + \delta x.$$

Leállási feltétel:

$$\|f(x^{(k+1)})\|_\infty < \varepsilon \cdot (1 + \|f(x^{(0)})\|_\infty)$$

Fixpont-iteráció egyenletrendszerekre

A $g(x) = x$ gyökét keressük, ahol $g : T \to \mathbb{R}^n$, $T \subseteq \mathbb{R}^n$.

Példa:

$$\begin{aligned} \frac{1}{4}\cos(2x_1 - x_2) - \frac{3}{4} &= x_1 \\ \frac{1}{3}\sin(x_1) - \frac{2}{3} &= x_2 \end{aligned}$$

Az algoritmus:
$x^{(0)}$ kezdővektor, $x^{(k+1)} = g(x^{(k)})$, $k = 0, 1, \ldots$.

Fixpont-iteráció egyenletrendszerekre

A $g(x) = x$ gyökét keressük, ahol $g : T \to \mathbb{R}^n$, $T \subseteq \mathbb{R}^n$.

Az algoritmus

$x^{(0)}$ kezdővektor, $x^{(k+1)} = g(x^{(k)})$, $k = 0, 1, \dots$.

Tétel.

Ha T konvex, $g(T) \subseteq T$, és g differenciálható, továbbá $\|J(x)\| \leq \alpha < 1$ minden $x \in T$-re, akkor az egyenletrendszer egyértelműen megoldható, és $\forall x^{(0)} \in T$ esetén az $x^{(k+1)} = g(x^{(k)})$, $k = 0, 1, \dots$ sorozat tart a megoldáshoz.

Feladat

Az

$$\cos(x_1 - x_2) - \sin(x_2) - 4x_1 = 0$$
$$\cos(x_1 + x_2) - \sin(x_1 - x_2) - 5x_2 = 0$$

egyenletrendszer megoldását keressük a $[-1,1]^2$ tartományon. Mit mondhatunk a rendszer megoldhatóságáról és az

$$x_1^{(k+1)} = \frac{1}{4}\cos(x_1^{(k)} - x_2^{(k)}) - \frac{1}{4}\sin(x_2^{(k)}),$$
$$x_2^{(k+1)} = \frac{1}{5}\cos(x_1^{(k)} + x_2^{(k)}) - \frac{1}{5}\sin(x_1^{(k)} - x_2^{(k)})$$

$k = 0, 1, \ldots$ eljárás konvergenciájáról?

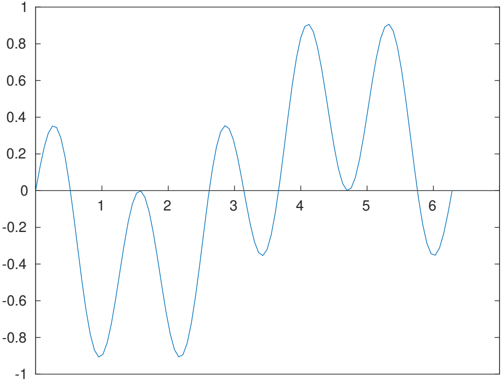
Numerikus matematika

Baran Ágnes

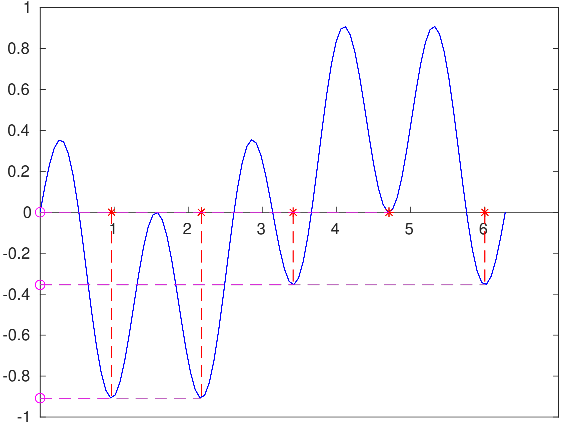
Optimalizálás

Egyváltozós függvény szélsőértéke (emlékeztető)

Az $f : \mathbb{R} \to \mathbb{R}$ függvény szélsőértékhelyeit keressük.



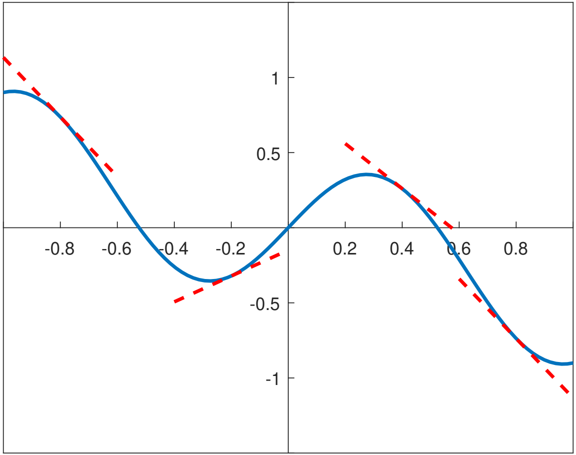
Egy függvénynek több lokális szélsőértéke is lehet.



Az $f(x) = \sin(2x)\cos(3x)$ függvény $[0, 2\pi]$-beli

- lokális minimumhelyei (*) és
- lokális minimumai (o)

Legyen $f : \mathbb{R} \to \mathbb{R}$ egy differenciálható függvény, mit jelent $f'(x_0)$?

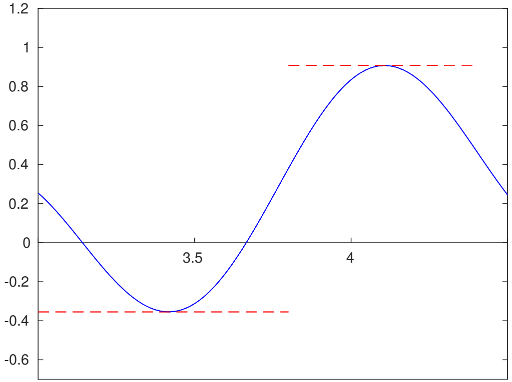


Emlékeztető: az f függvény x_0-beli érintője:

$$y(x) = f(x_0) + f'(x_0)(x - x_0)$$

Szélsőérték, szükséges feltétel

Az $f : \mathbb{R} \to \mathbb{R}$ differenciálható függvénynek csak ott lehet lokális szélsőértéke, ahol $f'(x) = 0$.

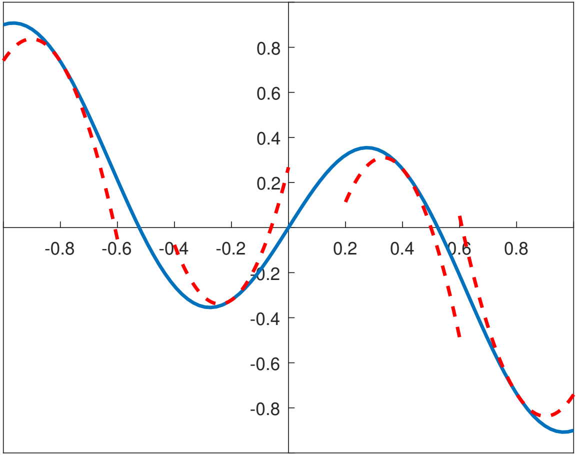


Fordítva nem igaz! Abból, hogy $f'(x) = 0$ NEM következik, hogy a függvénynek ott lokális szélsőértéke van.

Emlékeztető: ha $f : \mathbb{R} \to \mathbb{R}$ egy kétszer differenciálható függvény, akkor x_0 egy kis környezetében közelíthetjük az

$$y(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2$$

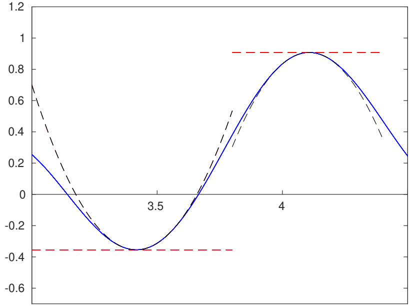
másodfokú polinommal.



Szélsőérték, elégséges feltételek

Ha az $f : \mathbb{R} \to \mathbb{R}$ függvény kétszer differenciálható és

- $f'(x^\star) = 0$, $f''(x^\star) > 0$, akkor f-nek $x^\star$-ban lokális minimuma van.
- $f'(x^\star) = 0$, $f''(x^\star) < 0$, akkor f-nek $x^\star$-ban lokális maximuma van.



piros szaggatott vonal: $f(x_0) + f'(x_0)(x - x_0)$

fekete szaggatott vonal: $f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2$

Egyváltozós függvény szélsőértéke

Az $f : \mathbb{R} \to \mathbb{R}$ függvény szélsőértékhelyeit keressük.

Szükséges feltétel

Az $f : \mathbb{R} \to \mathbb{R}$ differenciálható függvénynek csak ott lehet lokális szélsőértéke, ahol $f'(x) = 0$.

Elégséges feltételek

Ha az $f : \mathbb{R} \to \mathbb{R}$ függvény kétszer differenciálható és

- $f'(x^\star) = 0$, $f''(x^\star) > 0$, akkor f-nek $x^\star$-ban lokális minimuma van.
- $f'(x^\star) = 0$, $f''(x^\star) < 0$, akkor f-nek $x^\star$-ban lokális maximuma van.

Példa

Egy légitársaság A és B város közötti repülőjáratára 500 Euró egy jegy. A két város között egy 300 férőhelyes gép közlekedik, de átlagosan csak 180 utassal. Piackutatások szerint minden 5 Eurós engedmény a jegyárból átlagosan 3 plusz utast jelentene. Milyen jegyár mellett lenne maximális a légitársaság bevétele?

Tegyük fel, hogy a légitársaság $5n$ Eurót enged a jegyárból. Ekkor a várható bevétele:

$$f(n) = (180 + 3n)(500 - 5n) = -15n^2 + 600n + 90000$$

Az f maximumhelyét keressük.

$$f'(n) = -30n + 600$$

$$f'(n) = 0 \iff n = 20$$

Mivel

$$f''(n) = -30,$$

így $f''(20) < 0$, azaz $n = 20$ az f függvény maximumhelye.

1. feladat

Keresse meg az $f(x) = x^3 - 6x^2 + 9x + 15$ függvény lokális szélsőértékhelyeit!

2. feladat

Egy 108 dm^3 térfogatú, négyzet alapú, felül nyitott dobozt akarunk készíteni. Hogyan válasszuk meg a doboz méretét, ha a készítéséhez felhasznált anyag mennyiségét minimalizálni szeretnénk?

3. feladat

Egy folyó melletti telken szeretnénk egy 1800 m^2-es téglalap alakú részt elkeríteni úgy, hogy egyik oldalról a folyó határolja. Milyen méretű részt kerítsünk el, ha a felhasznált kerítés hosszát minimalizálni szeretnénk?

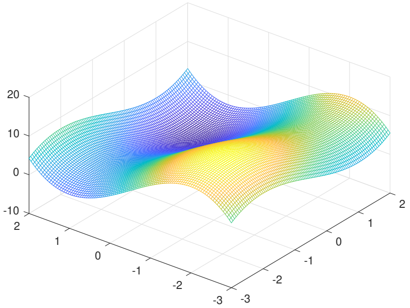
Kétváltozós függvények

Példa

Az

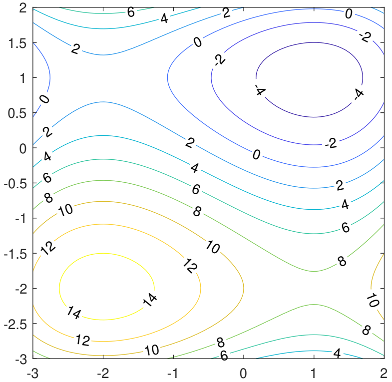
$$f(x_1, x_2) = \frac{1}{4}(2x_1^3 + 3x_1^2 - 12x_1) + \frac{1}{2}(2x_2^3 + 3x_2^2 - 12x_2)$$

függvény a $[-2, 2] \times [-2, 3]$ tartomány felett.



Példa

Rajzoltassuk ki az előző függvény szintvonalait is. (Mikroökonómia: szintvonal = közömbösségi görbe.)



Kétváltozós függvények minimalizálása

Az $f : \mathbb{R}^2 \to \mathbb{R}$ függvény lokális minimumhelyeit keressük.

Gradiens

Az $f : \mathbb{R}^2 \to \mathbb{R}$ függvény x-beli gradiense

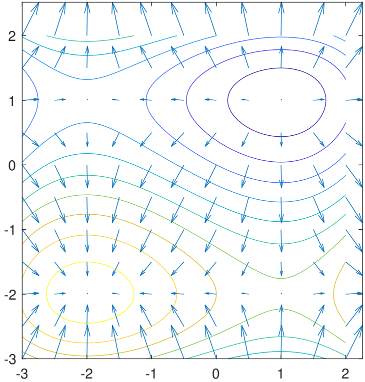
$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \end{bmatrix}$$

Példa

$$f(x_1, x_2) = \frac{1}{4}(2x_1^3 + 3x_1^2 - 12x_1) + \frac{1}{2}(2x_2^3 + 3x_2^2 - 12x_2)$$

$$\nabla f(x) = \begin{bmatrix} \frac{3}{2}x_1^2 + \frac{3}{2}x_1 - 3 \\ 3x_2^2 + 3x_2 - 6 \end{bmatrix}$$

Látjuk, hogy a gradiensvektor értéke pontonként más-más lehet. Rácsozzuk be a $[-3, 2]^2$ tartományt (mindkét tengely mentén 11-11 részre osztva) és számítsuk ki az előző függvény gradiensét ezekben a pontokban, majd rajzoltassuk rá ezeket a vektorokat a szintvonalakra!

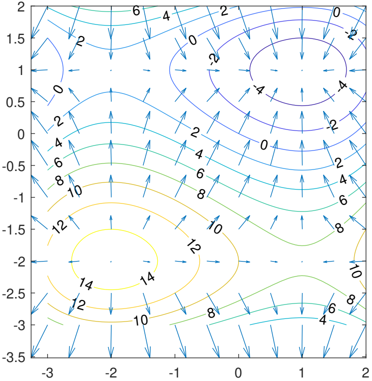


Az előző ábrán megfigyelhetjük, hogy

- a gradiensvektor merőleges az adott pontbeli szintvonalra
- a vektorok hossza a gradiens nagyságát, az iránya a gradiens irányát mutatja
- bizonyos pontokban a gradiensvektor hossza 0, vagy 0 közeli

A gradiensvektor az adott pontban a legmeredekebb emelkedés irányába mutat, a (-1)-szerese (a negatív gradiens) pedig a legmeredekebb csökkenés irányába.

Ha a gradiensmező helyett a negatív gradiensmezőt rajzoltatjuk ki, akkor a nyilak a csökkenés irányába mutatnak.



Az

$$f(x_1, x_2) = \frac{1}{4}(2x_1^3 + 3x_1^2 - 12x_1) + \frac{1}{2}(2x_2^3 + 3x_2^2 - 12x_2)$$

függvény szintvonalai és a negatív gradiens mező.

A lokális szélsőérték feltételei

Elsőrendű szükséges feltétel

Ha x^* az $f : \mathbb{R}^2 \to \mathbb{R}$ lokális minimumhelye, és f folytonosan differenciálható az x^* egy nyílt környezetében, akkor $\nabla f(x^*) = 0$.

Definíció (Stacionárius pont)

Legyen $f : \mathbb{R}^2 \to \mathbb{R}$. Az x^* pontot stacionárius pontnak hívjuk, ha $\nabla f(x^*) = 0$.

Megjegyzés

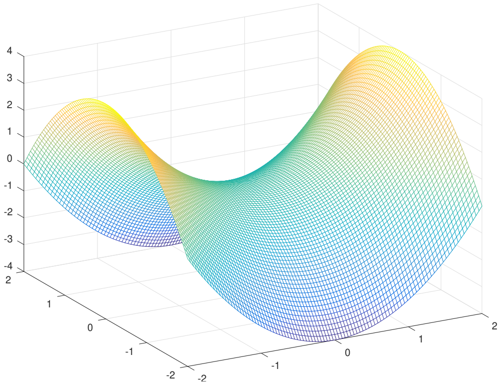
Ha x^* stacionárius pontja f-nek, akkor stacionárius pontja $-f$-nek is, azaz a stacionárius pont lokális maximum is lehet.

Definíció (Nyeregpont)

Ha x^* olyan stacionárius pontja f-nek, amely se nem lokális minimum, se nem lokális maximum, akkor nyeregpontnak hívjuk.

Példa

Legyen $f(x) = x_1^2 - x_2^2$. Ekkor $\nabla f(x) = (2x_1, -2x_2)^T$, így $x = (0,0)$ az egyetlen stacionárius pont, amely nyeregpont.



Példa

Határozzuk meg az

$$f(x_1, x_2) = \frac{1}{4}(2x_1^3 + 3x_1^2 - 12x_1) + \frac{1}{2}(2x_2^3 + 3x_2^2 - 12x_2)$$

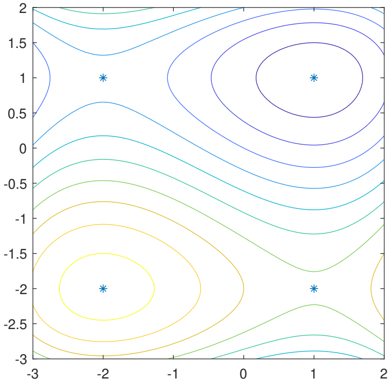
függvény stacionárius pontjait!

$$\nabla f(x) = \begin{bmatrix} \frac{3}{2}x_1^2 + \frac{3}{2}x_1 - 3 \\ 3x_2^2 + 3x_2 - 6 \end{bmatrix}$$

$$\nabla f(x) = 0 \iff x_1^2 + x_1 - 2 = 0 \text{ és } x_2^2 + x_2 - 2 = 0$$

A stacionárius pontok:

$$(1, 1), \quad (1, -2), \quad (-2, 1), \quad (-2, -2)$$



Az

$$f(x_1, x_2) = \frac{1}{4}(2x_1^3 + 3x_1^2 - 12x_1) + \frac{1}{2}(2x_2^3 + 3x_2^2 - 12x_2)$$

függvény szintvonalai és stacionárius pontjai.

A stacionárius pont típusai

Hesse-mátrix

Vezessük be a következő jelölést:

$$f_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}, \quad i, j = 1, 2.$$

Az $f : \mathbb{R}^2 \to \mathbb{R}$ függvény Hesse-mátrixa:

$$H(x) = \begin{bmatrix} f_{11}(x) & f_{12}(x) \\ f_{21}(x) & f_{22}(x) \end{bmatrix}$$

Legyen $\Delta_1 := f_{11}(x)$ és $\Delta_2 := \det(H(x))$.

A stacionárius pont típusai

Tétel

Ha az $x \in \mathbb{R}^2$ stacionárius pontban

- $\Delta_2 > 0$ és $\Delta_1 > 0$, akkor x lokális minimumhely.
- $\Delta_2 > 0$ és $\Delta_1 < 0$, akkor x lokális maximumhely.
- $\Delta_2 < 0$, akkor x nyeregpont.
- $\Delta_2 = 0$, akkor további vizsgálat szükséges.

Példa

Határozzuk meg az

$$f(x_1, x_2) = \frac{1}{4}(2x_1^3 + 3x_1^2 - 12x_1) + \frac{1}{2}(2x_2^3 + 3x_2^2 - 12x_2)$$

függvény stacionárius pontjainak típusát.

$$f(x_1, x_2) = \frac{1}{4}(2x_1^3 + 3x_1^2 - 12x_1) + \frac{1}{2}(2x_2^3 + 3x_2^2 - 12x_2)$$

A gradiens:

$$\nabla f(x) = \begin{bmatrix} \frac{3}{2}x_1^2 + \frac{3}{2}x_1 - 3 \\ 3x_2^2 + 3x_2 - 6 \end{bmatrix}$$

A stacionárius pontok:

$$(1,1), \quad (1,-2), \quad (-2,1), \quad (-2,-2)$$

A Hesse-mátrix:

$$H(x) = \begin{bmatrix} 3x_1 + \frac{3}{2} & 0 \\ 0 & 6x_2 + 3 \end{bmatrix}$$

$$H(1,1) = \begin{bmatrix} \frac{9}{2} & 0 \\ 0 & 9 \end{bmatrix}$$

$\Delta_1 > 0$ és $\Delta_2 > 0$, így az $(1,1)$ lokális minimumhely

$$H(1,-2) = \begin{bmatrix} \frac{9}{2} & 0 \\ 0 & -9 \end{bmatrix}$$

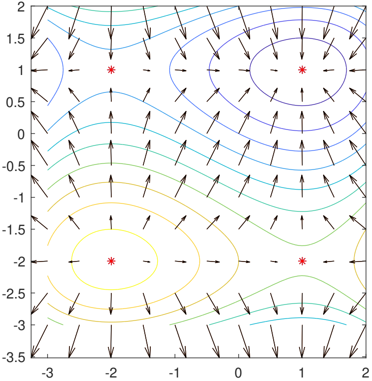
$\Delta_2 < 0$, így az $(1,-2)$ nyeregpont.

$$H(-2,1) = \begin{bmatrix} -\frac{9}{2} & 0 \\ 0 & 9 \end{bmatrix}$$

$\Delta_2 < 0$, így az $(-2,1)$ nyeregpont.

$$H(-2,-2) = \begin{bmatrix} -\frac{9}{2} & 0 \\ 0 & -9 \end{bmatrix}$$

$\Delta_1 < 0$ és $\Delta_2 > 0$, így az $(-2,-2)$ lokális maximumhely

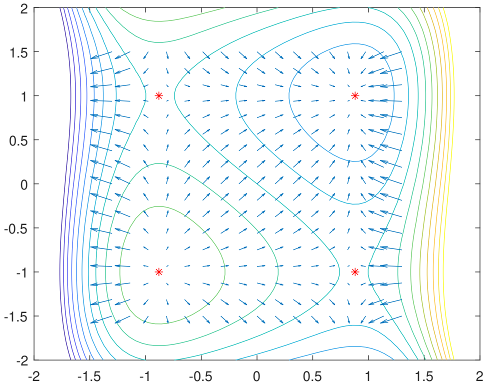


Az

$$f(x_1, x_2) = \frac{1}{4}(2x_1^3 + 3x_1^2 - 12x_1) + \frac{1}{2}(2x_2^3 + 3x_2^2 - 12x_2)$$

függvény szintvonalai, stacionárius pontjai és a negatív gradiens mező.

Az ábrán az $f(x) = x_1^5 + x_2^3 - 3x_1 - 3x_2$ függvény szintvonalai láthatók a $[-2, 2]^2$ tartományon, a negatív gradiensmezővel együtt. A függvénynek ebben a tartományban 4 stacionárius pontja van (*).
Adja meg a stacionárius pontok típusát, ha a negatív gradiensmezőt elég sűrű rácson ábrázoltuk ahhoz, hogy jól jellemezze a függvényt.



4. feladat

Határozza meg az

$$f(x) = 2x_1^2 x_2 - x_1 x_2^2 + 4x_1 x_2$$

függvény stacionárius pontjait, és azok típusát.

5. feladat

Határozza meg az

$$f(x) = x_1^3 - x_2^3 + 6x_1 x_2$$

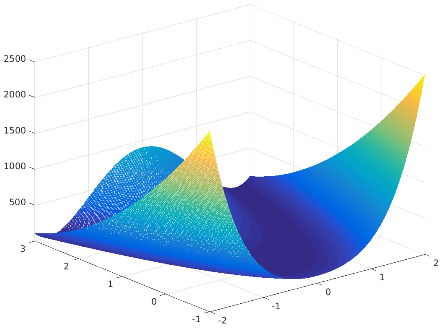
függvény stacionárius pontjait, és azok típusát.

6. feladat (Rosenbrock függvény)

Határozza meg az

$$f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

függvény stacionárius pontjait, és azok típusát.



7. feladat

Egy autókereskedés egy adott autómárkából kombi és szedán típust is értékesít. Egy piackutatás eredménye azt mutatja, hogy ha a kombik ára x_1, a szedánoké x_2, akkor a kereslet a két autótípus iránt rendre

$$k = 10000 - 2x_1 + 2.5x_2$$
$$s = 16000 + 1.5x_1 - 3x_2$$

(ha az egyik típus ára emelkedik, akkor az ezirányú kereslet csökken, viszont a másik típusé nő). Hogyan érdemes megválasztani az egyes típusok árait, ha a bevételt maximalizálni szeretnénk?

Gradiens-módszer

Az $f:\mathbb{R}^2 \to \mathbb{R}$ függvény lokális minimumhelyeit keressük.

Egy $x^{(k)}$, $k = 0, 1, \dots$ sorozatot definiálunk, mely optimális esetben közelíti a függvény egy lokális minimumhelyét.

A módszer adott, ha

- az $x^{(0)}$ kezdővektor adott,
- ismert az $x^{(k)} \mapsto x^{(k+1)}$ stratégia,
- adott a leállási feltétel.

A gradiens-módszer esetén az $x^{(k)}$ pontból a legmeredekebb leereszkedés irányában lépünk tovább.

Az $x^{(k)}$-beli legmeredekebb leereszkedés iránya: $-\nabla f(x^{(k)})$.

Gradiens módszer

- $x^{(0)}$ adott,
- ha $x^{(k)}$ adott, akkor

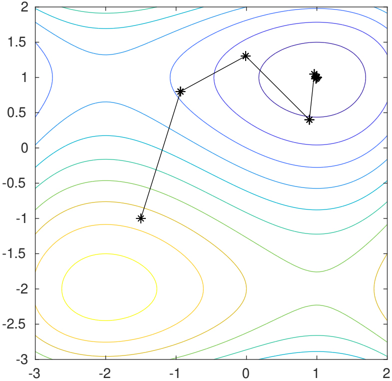
$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}), \quad k = 0, 1, \ldots,$$

  ahol $\alpha_k > 0$ a lépéshossz. α_k értékét úgy határozzuk meg, hogy $x^{(k)}$-ból indulva $p_k = -\nabla f(x^{(k)})$ irányban meghatározzuk az f minimumhelyét, vagy annak egy elég jó közelítését.
- Leállási feltétel: ha $\|\nabla f(x^{(k)})\| < \varepsilon$, ahol $\varepsilon > 0$ adott paraméter.

Megjegyzés: Az α_k lépéshossz meghatározására különféle algoritmusok léteznek.

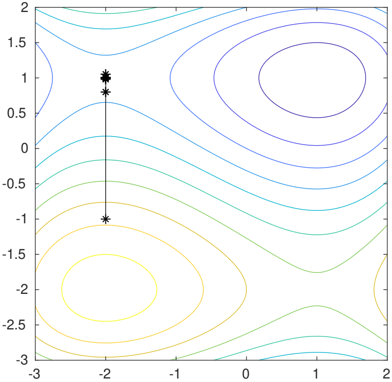
Példa

Vizsgáljuk meg a gradiens-módszer viselkedését az
$f(x_1, x_2) = \frac{1}{4}(2x_1^3 + 3x_1^2 - 12x_1) + \frac{1}{2}(2x_2^3 + 3x_2^2 - 12x_2)$ függvény esetén.



$x^{(0)} = [-1.5, -1]^T$, $\varepsilon = 0.001$, az elvégzett lépések száma 10,
$x^{(10)} = [1.0000, 0.9999]^T$.

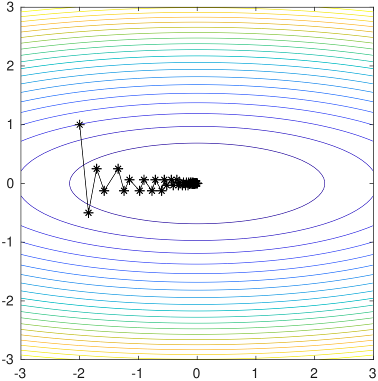
Előfordulhat, hogy a gradiens-módszer a függvény egy nyeregpontjában áll meg.



$x^{(0)} = [-2, -1]^T$, $\varepsilon = 0.001$, az elvégzett lépések száma 7, $x^{(7)} = [-2.0000, 1.0000]^T$.

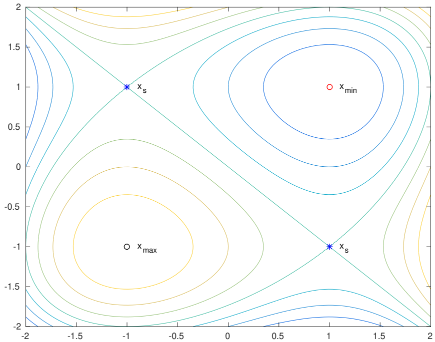
Megjegyzés

Ha a felület elnyújtott völgyeket tartalmaz, akkor a gradiens-módszer konvergenciája lassú lehet.



$f(x) = x_1^2 + 10x_2^2$,
$x^{(0)} = [-2, 1]^T$, $\varepsilon = 0.001$, az elvégzett lépések száma 78,
$x^{(78)} = [0.0000, 0.0000]^T$.

Gradiens módszer



Az

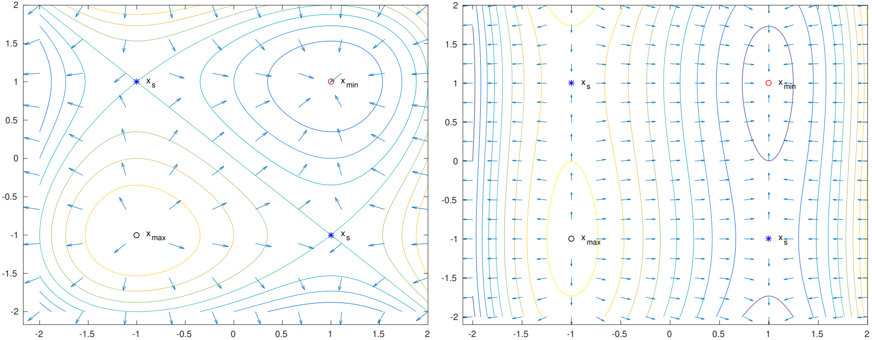
$$f(x_1, x_2) = x_1^3 + x_2^3 - 3x_1 - 3x_2$$

és a

$$f(x_1, x_2) = 10x_1^3 + x_2^3 - 30x_1 - 3x_2$$

függvény szintvonalai.

Gradiens módszer



Az

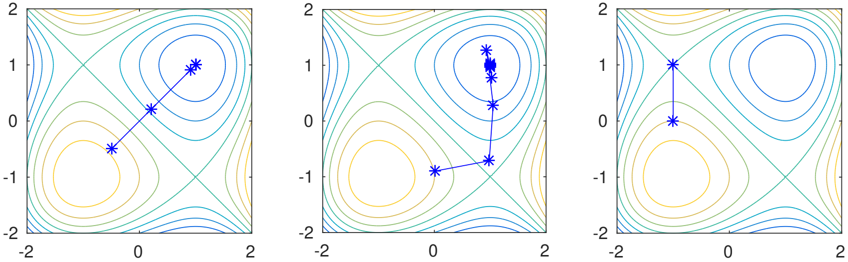
$$f(x_1, x_2) = x_1^3 + x_2^3 - 3x_1 - 3x_2$$

és a

$$f(x_1, x_2) = 10x_1^3 + x_2^3 - 30x_1 - 3x_2$$

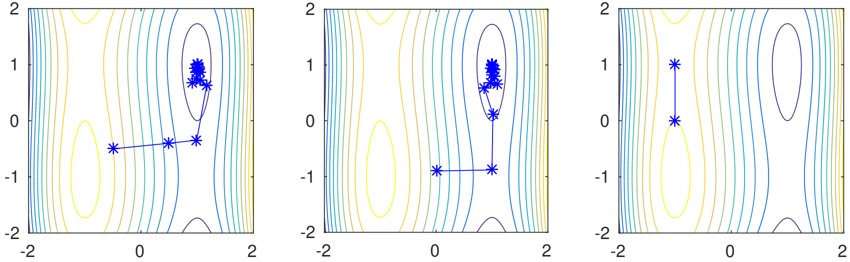
függvény szintvonalai és a negatív gradiensmezők.

Gradiens módszer



A gradiens módszer az $f(x_1, x_2) = x_1^3 + x_2^3 - 3x_1 - 3x_2$ függvény esetén.

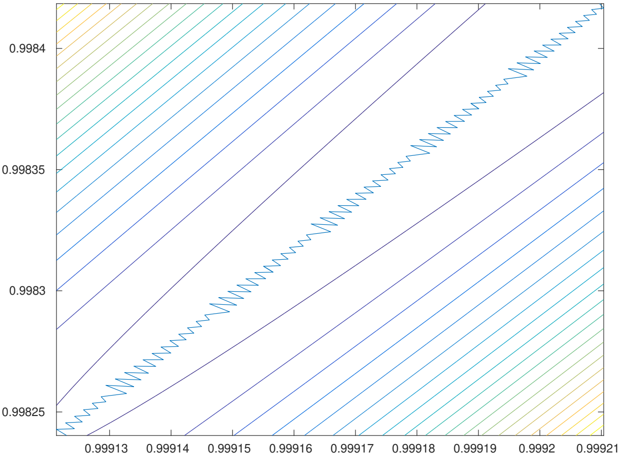
x_0	$(-0.5, -0.5)$	$(0, -0.9)$	$(-1, 0)$
lépés	6	11	2
x^*	$(1.0000, 1.0000)$	$(1.0000, 0.9999)$	$(-1, 1)$



A gradiens módszer az $f(x_1, x_2) = 10x_1^3 + x_2^3 - 30x_1 - 3x_2$ függvény esetén.

x_0	$(-0.5, -0.5)$	$(0, -0.9)$	$(-1, 0)$
lépésszám	36	33	2
x^*	$(1.0000, 0.9999)$	$(1.0000, 1.0001)$	$(-1, 1)$

Gradiens módszer



A gradiens módszer a Rosenbrock-függvényre, az utolsó 130 iterált.
$x^{(0)} = (-1.2, 1)$, $\varepsilon = 10^{-3}$. Az elvégzett lépések száma 5231.

Newton-módszer optimalizálásra

Newton-módszer nemlineáris egyenletek gyökeinek közelítésére, emlékeztető:

Az $f(x) = 0$ (ahol $f : \mathbb{R} \to \mathbb{R}$) **nemlineáris egyenlet** gyökének közelítésére szolgáló Newton-iteráció:

$$x_0 \text{ adott}, \quad x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \ldots$$

Az $F(x) = 0$ (ahol $F : \mathbb{R}^n \to \mathbb{R}^n$) **nemlineáris egyenletrendszer** gyökének közelítésére szolgáló Newton-iteráció:

$$x^{(0)} \text{ adott}, \quad F'(x^{(k)})(x^{(k+1)} - x^{(k)}) = -F(x^{(k)}), \quad k = 0, 1, 2, \ldots$$

Newton-módszer optimalizálásra

Az f függvény minimumhelye megoldása a $\nabla f(x) = 0$ egyenletnek. Mivel $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, ezért ez egy nemlineáris egyenletrendszer.

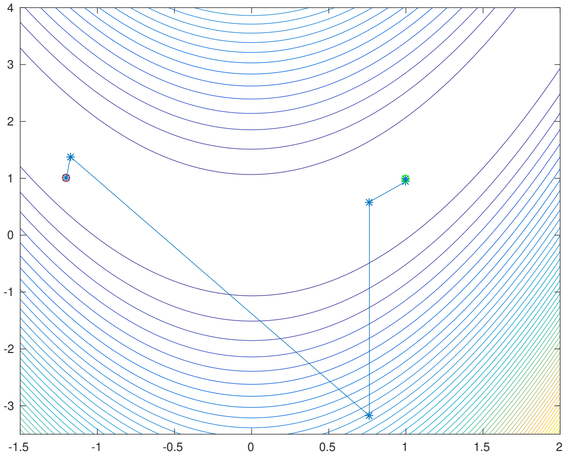
Ha f kétszer folytonosan differenciálható, akkor a Newton-módszer a $\nabla f(x) = 0$ egyenletre:

$$x^{(0)} \text{ adott}, \quad H(x^{(k)})(x^{(k+1)} - x^{(k)}) = -\nabla f(x^{(k)}), \quad k = 0, 1, 2, \ldots$$

ahol H az f függvény Hesse-mátrixa.

- $x^{(0)}$ adott,
- $H(x^{(k)})p_k = -\nabla f(x^{(k)})$, (azaz $p_k = -(H_k)^{-1}\nabla f_k$)
- $x^{(k+1)} = x^{(k)} + p_k$
- ha $\|\nabla f_k\| < \varepsilon$, akkor leállás

Newton-módszer, példa



A Newton-módszer a Rosenbrock-függvényre. $x^{(0)} = (-1.2, 1)$, $x_{opt} = (0.999996, 0.999991)$, $f(x_{opt}) = 1.8 \cdot 10^{-11}$, $k = 5$.

Numerikus matematika

Baran Ágnes

Numerikus integrálás

Integrálközelítések.

Az

$$\mathcal{I}(f) := \int_a^b f(x)dx$$

határozott integrált szeretnénk kiszámítani, ahol $f : [a, b] \to \mathbb{R}$.

Miért lehet szükség integrálközelítésre?

- f nem elemien integrálható
- f primitív függvényének felírása bonyolult
- nagyszámú integrál kiszámítására van szükségünk
- f nem explicit képlettel adott, csak bizonyos pontokban ismerjük az értékét

Az $\mathcal{I}(f)$ közelítését

$$\mathcal{I}_n(f) = \sum_{i=1}^{n} a_i f(x_i)$$

alakban keressük, ahol

$x_1, \ldots, x_n$ a közelítés alappontjai, ($x_i \in [a, b]$),

$a_1, \ldots, a_n$ súlyok (melyek az f függvénytől nem függnek).

$\mathcal{I}_n(f)$: kvadratúraképlet (szabad paraméterei: n, $x_1, \ldots, x_n$, $a_1, \ldots, a_n$)

Interpolációs kvadratúraképletek.

Legyenek adottak az $x_1, \ldots, x_n$ alappontok.

Közelítsük f-et az $x_1, \ldots, x_n$-re támaszkodó Lagrange polinomjával:

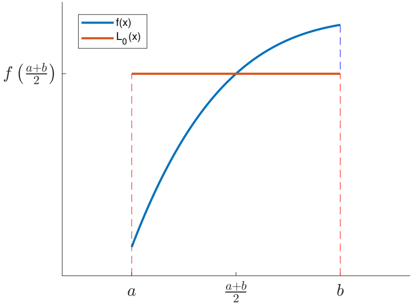
$$f(x) \approx L_{n-1}(x).$$

$$\mathcal{I}(f) = \int_a^b f(x)dx \approx \int_a^b L_{n-1}dx = \mathcal{I}_n(f)$$

Egyszerű érintőképlet.

$n = 1$, azaz 1 alappont adott, és ez az $\frac{a+b}{2}$ pont.

Ekkor a közelítő polinom egy konstansfüggvény: $L_0(x) \equiv f\left(\frac{a+b}{2}\right)$



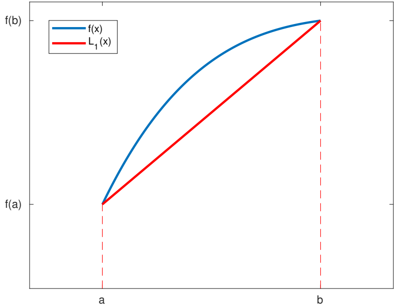
$$\boxed{\mathcal{I}_1(f) = (b-a)\cdot f\left(\frac{a+b}{2}\right)}$$

A képlet pontos minden legfeljebb elsőfokú polinom esetén.

Egyszerű trapéz-képlet.

$n = 2$, azaz 2 alappont adott, és ezek az intervallum végpontjai: a és b.

Ekkor a f-et az $(a, f(a))$ és $(b, f(b))$ adatokra illeszkedő egyenessel közelítjük.



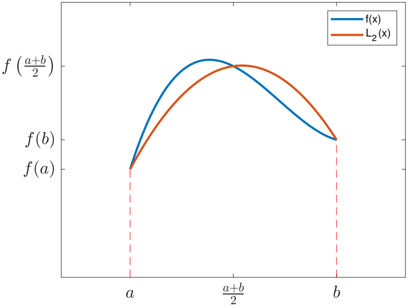
$$\boxed{\mathcal{I}_2(f) = (b-a)\frac{f(a)+f(b)}{2}}$$

A képlet pontos minden legfeljebb elsőfokú polinom esetén.

Egyszerű Simpson-képlet.

$n = 3$, azaz 3 alappont adott, és ezek az a, $\frac{a+b}{2}$ és b pontok.

Az f-et egy másodfokú polinommal közelítjük.



$$\boxed{\mathcal{I}_3(f) = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)}$$

A képlet pontos minden legfeljebb harmadfokú polinom esetén.

Összetett képletek.

Osszuk fel az $[a, b]$ intervallumot m egyforma hosszúságú részintervallumra:

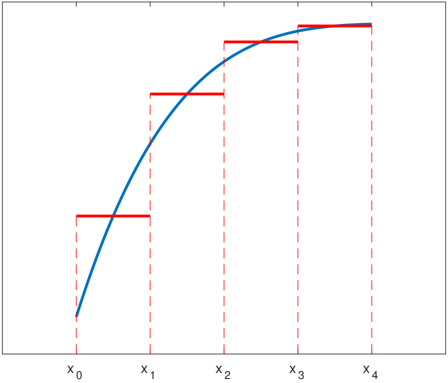
$$a = x_0 < x_1 < \cdots < x_m = b,$$

A részintervallumok hosszát jelölje h:

$$h := \frac{b-a}{m} = x_i - x_{i-1}, \quad i = 1, \ldots, m.$$

Minden részintervallumon alkalmazzuk ugyanazt az egyszerű képletet.

Összetett érintőképlet



Összetett érintőképlet

$$\mathcal{I}_{m\times 1}(f) = h\left[f\left(x_0+\frac{h}{2}\right)+f\left(x_1+\frac{h}{2}\right)+\dots f\left(x_{m-1}+\frac{h}{2}\right)\right]$$

azaz

$$\mathcal{I}_{m\times 1}(f) = h\sum_{i=0}^{m-1} f\left(x_i+\frac{h}{2}\right)$$

Az összetett érintőképlet hibája:

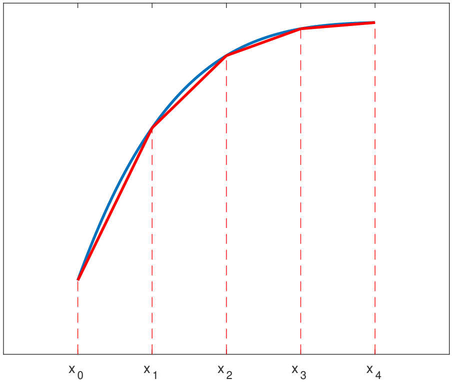
Ha f kétszer folytonosan differenciálható, akkor

$$|\mathcal{I}(f)-\mathcal{I}_{m\times 1}(f)| \le \frac{(b-a)^3}{24m^2}M_2,$$

ahol $M_2 = \max_{x\in[a,b]} |f''(x)|$

A képlet pontos minden legfeljebb elsőfokú polinom esetén.

Összetett trapéz-képlet



Összetett trapéz-képlet.

$$\mathcal{I}_{m\times 2}(f) = h\left[\frac{f(x_0)}{2} + f(x_1) + f(x_2) + \cdots + f(x_{m-1}) + \frac{f(x_m)}{2}\right].$$

azaz

$$\mathcal{I}_{m\times 2}(f) = h\left[\frac{f(x_0)}{2} + \sum_{i=1}^{m-1} f(x_i) + \frac{f(x_m)}{2}\right]$$

Az összetett trapéz-képlet hibája:

Ha f kétszer folytonosan differenciálható, akkor

$$|\mathcal{I}(f) - \mathcal{I}_{m\times 2}(f)| \leq \frac{(b-a)^3}{12m^2} M_2,$$

ahol $M_2 = \max_{x\in[a,b]} |f''(x)|$.

A képlet pontos minden legfeljebb elsőfokú polinom esetén.

Összetett Simpson-képlet.

$$\mathcal{I}_{m\times 3} = \frac{h}{6}\left[f(x_0) + 4f\left(x_0 + \frac{h}{2}\right) + 2f(x_1) + 4f\left(x_1 + \frac{h}{2}\right) + \dots \right.$$
$$\left. + 2f(x_{m-1}) + 4f\left(x_{m-1} + \frac{h}{2}\right) + f(x_m)\right].$$

azaz

$$\mathcal{I}_{m\times 3} = \frac{h}{6}\left[f(x_0) + 4\sum_{i=0}^{m-1} f\left(x_i + \frac{h}{2}\right) + 2\sum_{i=1}^{m-1} f(x_i) + f(x_m)\right]$$

Az összetett Simpson-képlet hibája:
Ha f négyszer folytonosan differenciálható, akkor

$$|\mathcal{I}(f) - \mathcal{I}_{m\times 3}(f)| \leq \frac{(b-a)^5}{2880 m^4} M_4,$$

ahol $M_4 = \max_{x\in[a,b]} |f^{(4)}(x)|$.

A képlet pontos minden legfeljebb harmadfokú polinom esetén.

Összetett képletek konvergenciája

Tétel.

Ha az n pontra épülő egyszerű képlet pontos a konstans függények esetén, akkor

$$\lim_{m\to\infty} \mathcal{I}_{m\times n}(f) = \int_a^b f(x)dx$$

minden Riemann-integrálható f függvény esetén.

Példa.

Közelítsük

$$\int_4^{5.2} \ln x dx$$

értékét összetett trapéz-képlettel úgy, hogy az intervallumot 6 részintervallumra osztjuk! Becsüljük meg a közelítés hibáját!

$m = 6$

A részintervallumok hossza: $h = (b-a)/m = 0.2$

Az alappontok:

$x_0 = 4,\ x_1 = 4.2,\ x_2 = 4.4,\ x_3 = 4.6,\ x_4 = 4.8,\ x_5 = 5,\ x_6 = 5.2$

Az integrálközelítés:

$$\begin{aligned}\mathcal{I}_{6\times 2} &= 0.2\left(\frac{\ln 4}{2} + \ln 4.2 + \ln 4.4 + \cdots + \ln 5 + \frac{\ln 5.2}{2}\right)\\ &= 1.82765.\end{aligned}$$

A közelítés hibája:

$$|\mathcal{I} - \mathcal{I}_{m\times 2}| \leq \frac{(b-a)^3}{12\cdot m^2} M_2,$$

ahol $M_2 = \max\limits_{x\in[a,b]} |f''(x)|$.

Esetünkben $f(x) = \ln x$, $f'(x) = \frac{1}{x}$, $f''(x) = -\frac{1}{x^2}$, $M_2 = \frac{1}{16}$,

$$|\mathcal{I} - \mathcal{I}_{6\times 2}| \leq \frac{1.2^3}{12\cdot 6^2} \cdot \frac{1}{16} = 0.00025.$$

Példa.

Közelítsük

$$\int_4^{5.2} \ln x dx$$

értékét összetett Simpson-képlettel úgy, hogy az intervallumot 3 részintervallumra osztjuk! Becsüljük meg a közelítés hibáját!

$m = 3,\ h = (b-a)/m = 0.4$

$x_0 = 4,\ x_1 = 4.4,\ x_2 = 4.8,\ x_3 = 5.2$

$$\begin{aligned}\mathcal{I}_{3\times 3} &= \frac{0.4}{6}\left(\ln 4 + 4\ln 4.2 + 2\ln 4.4 + \cdots + 4\ln 5 + \ln 5.2\right)\\ &= 1.82785.\end{aligned}$$

A közelítés hibája:

$$|\mathcal{I} - \mathcal{I}_{m\times 3}| \leq \frac{(b-a)^5}{2880 \cdot m^4} M_4,$$

ahol $M_4 = \max_{x\in[a,b]} |f^{(4)}(x)|$.

Esetünkben $f'''(x) = \frac{2}{x^3}$, $f^{(4)}(x) = -\frac{6}{x^4}$, $M_4 = \frac{6}{4^4} = \frac{3}{128}$,

$$|\mathcal{I} - \mathcal{I}_{3\times 3}| \leq \frac{1.2^5}{2880 \cdot 3^4} \cdot \frac{3}{128} = 0.00000025.$$

Példa.

Becsüljük meg hány részintervallumra kell osztani az alapintervallumot, ha

$$\int_0^{\pi/4} \ln(\cos x) dx$$

értékét összetett trapéz-képlettel szeretnénk közelíteni úgy, hogy a hiba kisebb legyen, mint $0.5 \cdot 10^{-2}$.

$$|\mathcal{I} - \mathcal{I}_{m\times 2}| \leq \frac{(b-a)^3}{12 \cdot m^2} M_2.$$

Itt $f(x) = \ln(\cos x)$,

$f'(x) = -\frac{\sin x}{\cos x} = -\tan x$

$f''(x) = -\frac{1}{\cos^2 x}$,

tehát $M_2 = 2$.

$$\frac{(b-a)^3}{12\cdot m^2}M_2 = \left(\frac{\pi}{4}\right)^3 \cdot \frac{1}{12m^2}\cdot 2$$

m értékét úgy határozzuk meg, hogy

$$\left(\frac{\pi}{4}\right)^3 \cdot \frac{1}{12m^2}\cdot 2 < 0.5\cdot 10^{-2}$$

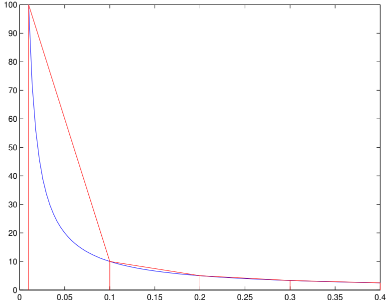
teljesüljön:

$$\frac{1}{3}\left(\frac{\pi}{4}\right)^3 \cdot 10^2 < m^2,$$

$$4.019 < m.$$

Adaptív eljárások

- a kvadratúra képlet költsége arányos a függvénykiértékelések (alappontok) számával
- az ekvidisztáns alappontrendszer időnként indokolatlanul sok számítást igényel



Adaptív eljárások

A függvény viselkedését figyelembe véve a számítás költsége csökkenthető

- Az aktuális intervallumon végezzük el az integrál közelítését két különböző módon (vagy ugyanazt a kvadratúra képletet alkalmazzuk két különböző -pl n és $2n$- alappontszám esetén, vagy ugyanarra az alappontszámra két különböző kvadratúra képletet)
- ha a két közelítés eltérése abszolútértékben nagyobb, mint $h_i\varepsilon/(b-a)$ (ahol ε adott, h_i az aktuális intervallum hossza), akkor az intervallumot osszuk fel két egyforma hosszúságú részintervallumra, és mindkettőre ismételjük meg az eljárást