Numerikus módszerek 1.

Bevezető

Krebsz Anna

ELTE IK

Elérhetőségek

Dr. Krebsz Anna
docens, ELTE IK Numerikus Analízis tanszék

e-mail:	krebsz@inf.elte.hu
honlap:	http://numanal.inf.elte.hu/$\sim$ krebsz
szoba:	2-302.

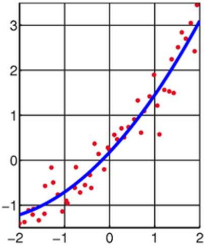
Technikai információk

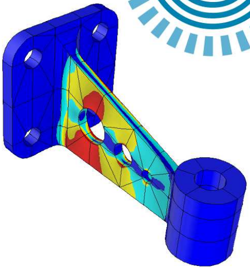
Tárgy:	Numerikus módszerek 1. előadás
	Prog. inf. BSc
Kód:	IP-18abNM1E (IK)
Félév:	2025/2026. tanév ősz
Helyszín:	Déli tömb, 0-822. Mogyoródi terem
Teams:	Csoport kód a kurzus Canvas felületén
Időpont:	Hétfő 8:10 - 9:50-ig 10 perc szünettel

A numerikus analízis célja olyan módszerek kidolgozása és elemzése, amelyek matematikai illetve **műszaki, természettudományos problémák** pontos vagy közelítő **számítógépes megoldását** célozzák meg.
Az első két félévben a **lineáris algebra** és az **analízis** numerikus módszereit tárgyaljuk. Ez egy bevezető kurzus a numerikus módszerekbe.

Alkalmazások. . .

$$\sqrt{2}$$





források: LaTeX, Prezi, http://met.hu/idojaras/elorejelzes/modellek/: AROME, WRF, INCA, WinAmp, Wikipedia

Tematika

1. I. félév.
   - Gépi számábrázolás. Hibaszámítás.
   - Lineáris egyenletrendszerek megoldása (direkt / iteratív).
     (Gauss-elimináció, LU-felbontás, LDU-felbontás, Cholesky-felbontás, QR-felbontás Gram–Schmidt-ortogonalizációval és Householder-transzformációval, mátrixnormák, Banach-féle fixponttétel, Jacobi-iteráció, Gauss–Seidel-iteráció, Richardson-iteráció)
   - Nemlineáris egyenletek megoldása.
     (intervallumfelezés, fixpont iterációk, Newton-módszer, szelőmódszer, húrmódszer)
   - Polinomok gyökeinek becslése. Horner-algoritmus a polinom és deriváltjainak helyettesítései értékeinek számítására.
2. II. félév
   - Sajátértékfeladatok megoldása (csak A szakirányon)
   - Interpoláció, approximáció
   - Numerikus integrálás

Irodalom, segédanyagok

Könyvek, jegyzetek (Numerikus analízis, Numerikus módszerek)

- Gergó Lajos
- Stoyan Gisbert
- Móricz Ferenc
- Linalg: Csörgő István

Elektronikus segédanyagok:

- **Előadás diasorai a Canvasban.**
- Példatár: Krebsz–Bozsik
- A Numerikus Analízis Tanszék, illetve oktatóinak honlapján: `http://numanal.inf.elte.hu/~{hegedus,krebsz,laszlo,soveg}`

Az előadás terve

1. Példák kézzel és Matlab-ban. (A legtöbb gépteremben legálisan hozzáférhető, lehet vele ismerkedni. Az A szakirányon a következő félévben kötelező.)

2. Előadás diasorok. (Elérhetőek lesznek a Canvasban. Definíciók, tételek, bizonyítások, példák. Néha krétás kiegészítés a táblán.)

Gyakorlati jegy:

- két évfolyam zh-ból és
- beadható HF-ból,
- részletek a gyakorlaton és a Canvasban.

Vizsga:

- „beugró kérdések”: 15 pontos beugróból legalább 8 pont elérése,
- „szóbeli vizsga”: egy tétel részletes kidolgozása.
- Az írásbeli és szóbeli együtt adja a vizsga jegyét. Részletek a kurzus Canvas felületén.)

Kérdések?

A matematikai modellezés (kör) folyamata és a hibaforrások megjelenése

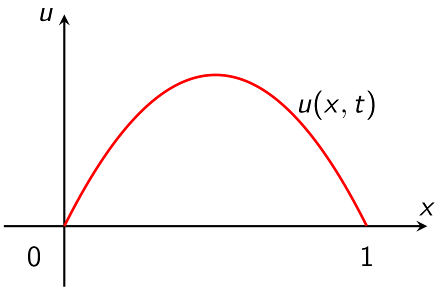
- A valóság egy részét vizsgálva igyekszünk a jelenséget fizikai, kémiai, stb. törvények alapján matematikailag leírni, egy lehetséges **matematikai modell**t megalkotni. Ez általában az adott tudományterületen dolgozó szakember feladata a rendelkezésre álló törvények, elvek felhasználásával. A valóságot csak közelíteni tudja, ezzel megjelenik a **modellhiba**.
- A modell pontos megoldása gyakran nem állítható elő véges lépésben, **közelítő módszerek**re van szükségünk. Elkészül a program, a végtelen eljárást végessel helyettesítjük, az itt megjelenő hibát **képlethibá**nak nevezzük.

A matematikai modellezés (kör) folyamata és a hibaforrások megjelenése

- A modell **bemenő paraméterei** általában mérési adatok, melyek pontatlanok, itt megjelenik a **mérési (vagy öröklött) hiba**.
- A közelítő módszer bemenő adatait véges aritmetikában ábrázoljuk, ezzel megjelenik az **input hiba**.
- A véges aritmetikában történő számolás során kerekítés, túl- illetve alulcsordulás léphet fel. Ezek **a műveleti (kerekítési) hibák**.
- A megvalósított **közelítő módszert teszteljük** és összehasonlítjuk a várt eredménnyel. Ha a kapott eredmény rossz, akkor elölről kezdjük az egyes lépések finomításával.

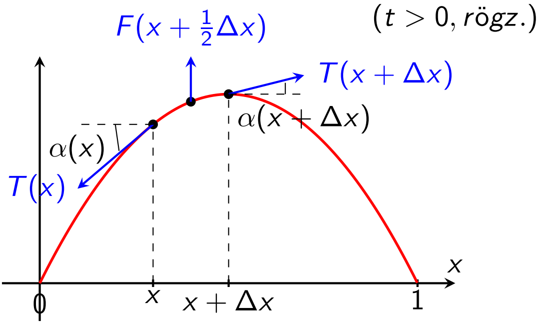
Húzzunk ki egy egységnyi hosszú rugalmas fémszálat és rögzítsük a végpontjait. Feszítsük meg, $t = 0$ időpontban engedjük el és hagyjuk rezegni.

Feladat: a rugalmas szál rezgésének meghatározása, vagyis az $u(x, t)$ elmozdulás meghatározása az x pontban és $t > 0$ időpontban.



Fizikai feltételek:

1. A szál tömegeloszlása homogén ($\varrho(x) \equiv \varrho$).
2. A szál tökéletesen rugalmas, azaz a szálban feszítő erő ($T(x)$) érintő irányú.
3. A nehézségi erő szálra gyakorolt hatását elhanyagoljuk.
4. A szálra ható kitérítő erő ($F(x)$) függőleges irányú és nem túl nagy.



- Az erők vízszintes komponesei kiegyenlítik egymást:

$$T(x)\cos(\alpha(x)) = T(x+\Delta x)\cos(\alpha(x+\Delta x)) = V(\text{állandó})$$

Ha V nem állandó, akkor elmozdul a szál (a 4. feltétel nem teljesül).

- Az x és $x+\Delta x$ pontokban ébredő feszítő erők függőleges komponenseinek különbsége az $x+\frac{\Delta x}{2}$ pontban ható kitérítő erőt egyenlíti ki:

$$F\left(x+\frac{\Delta x}{2}\right) = T(x+\Delta x)\sin(\alpha(x+\Delta x)) - T(x)\sin(\alpha(x)) \approx$$

$$\approx \underbrace{(\Delta x \cdot \varrho)}_{tömeg} \underbrace{\frac{\partial^2 u}{\partial t^2}(x,t)}_{gyorsulás} = m \cdot a$$

A kapott egyenletet osszuk le V-vel:

$$\frac{T(x+\Delta x)\sin(\alpha(x+\Delta x))}{T(x+\Delta x)\cos(\alpha(x+\Delta x))}-\frac{T(x)\sin(\alpha(x))}{T(x)\cos(\alpha(x))} \approx \frac{(\Delta x\cdot\varrho)}{V}\frac{\partial^2 u}{\partial t^2}(x,t)$$

$$\tan(\alpha(x+\Delta x))-\tan(\alpha(x)) \approx \frac{(\Delta x\cdot\varrho)}{V}\frac{\partial^2 u}{\partial t^2}(x,t)$$

$$\frac{\partial u}{\partial x}(x+\Delta x,t)-\frac{\partial u}{\partial x}(x,t) \approx \frac{(\Delta x\cdot\varrho)}{V}\frac{\partial^2 u}{\partial t^2}(x,t)$$

Leosztunk Δx-szel

$$\frac{\frac{\partial u}{\partial x}(x+\Delta x,t)-\frac{\partial u}{\partial x}(x,t)}{\Delta x} \approx \frac{\varrho}{V}\frac{\partial^2 u}{\partial t^2}(x,t).$$

$\Delta x \to 0$ esetén a

$$\frac{\partial^2 u}{\partial x^2}(x,t) = \frac{\varrho}{V} \cdot \frac{\partial^2 u}{\partial t^2}(x,t)$$

hiperbolikus differenciálegyeletet kapjuk.

Kiegészítjük a kezdeti feltételekkel és peremfeltételekkel:

$$\begin{aligned} u(0,t) &= 0 \\ u(1,t) &= 0 \\ u(x,0) &= s(x): \quad \text{a szál alakja kezdetben} \\ \frac{\partial u}{\partial t}(x,t_0) &= v(x): \quad \text{az elengedés pillanatában a kezdősebesség.} \end{aligned}$$

További egyszerűsítéseket teszünk, csak az időtől független speciális esettel foglalkozunk a továbbiakban.

Stacionárius eset:

$$\begin{aligned} t_0 :& \quad \text{egy adott időpillanat,} \\ U(x) :=& \; u(x, t_0) \\ A(x) :=& \; \frac{\varrho}{V} \cdot \frac{\partial^2 u}{\partial t^2}(x, t_0) \quad \text{a de. jobboldala} \end{aligned}$$

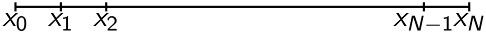
Ekkor a következő elliptikus de-t kapjuk:

$$\begin{aligned} \frac{\partial^2 U}{\partial x^2}(x) &= A(x) \\ U(0) = U(1) &= 0. \end{aligned}$$

A de. numerikus megoldása

A de. numerikus megoldása véges differencia módszerrel:
Elkészítjük a $[0;1]$ intervallum N részre történő egyenletes felosztását:

$$h=\frac{1}{N}, \quad x_i = ih \ \ (i=0,\ldots,N)$$



Ezekben a diszkrét pontokban felhasználjuk a jobboldali függvény értékét $A_i := A(x_i)$ és a megoldást is ezekben a pontokban keressük $u_i := U(x_i)$.

Tegyük fel, hogy a pontos megoldás $U \in D^3(0;1)$ és alkalmazzuk a Taylor-formulát:

$$U(x+h) = U(x) + U'(x)h + \frac{1}{2}U''(x)h^2 + \frac{1}{3!}U'''(\xi_1)h^3$$
$$U(x-h) = U(x) - U'(x)h + \frac{1}{2}U''(x)h^2 - \frac{1}{3!}U'''(\xi_2)h^3$$

Összeadva és átrendezve
$U(x+h)+U(x-h) = 2U(x)+U''(x)h^2 + \frac{1}{3!}h^3(U'''(\xi_1)-U'''(\xi_2))$

$$\frac{U(x+h)-2U(x)+U(x-h)}{h^2} = U''(x) + \frac{h}{6}(U'''(\xi_1)-U'''(\xi_2))$$

Ezzel megkaptuk az $U''(x) = \frac{\partial^2 U}{\partial x^2}(x)$ operátor 3 pontos közelítő sémáját:

$$\frac{U(x+h) - 2U(x) + U(x-h)}{h^2} \approx U''(x).$$

Ahogy láttuk a fenti képletben a közelítés hibája h-val arányos.

A diszkretizált pontokat behelyettesítve a következő LER-t kapjuk:

$$\frac{U(x_{i+1}) - 2U(x_i) + U(x_{i-1})}{h^2} = A(x_i) \quad (i = 1, \ldots, N-1)$$
$$U(x_0) = U(x_N) = 0.$$

A bevezetett jelölésekkel:

$$
\begin{aligned}
u_{i+1} - 2u_i + u_{i-1} &= h^2 A_i \quad (i = 1, \ldots, N-1) \\
u_0 = u_N &= 0.
\end{aligned}
$$

Mátrix alakban

$$
\begin{bmatrix}
2 & -1 & 0 & \ldots & 0 \\
-1 & 2 & -1 & \ldots & 0 \\
\ddots & \ddots & \ddots & & \vdots \\
 & & & & \\
0 & \ldots & 0 & -1 & 2
\end{bmatrix}
\cdot
\begin{bmatrix}
u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{N-1}
\end{bmatrix}
= -h^2 \cdot
\begin{bmatrix}
A_1 \\ A_2 \\ \vdots \\ \vdots \\ A_{N-1}
\end{bmatrix}
$$

$(N-1) \times (N-1)$ méretű LER (lineáris egyenletrendszer). Megoldása a gyors Gauss-eliminációval (progonka módszerrel) történik, lásd a félév során.

Numerikus módszerek 1.

1. előadás: Gépi számábrázolás, Hibaszámítás

Krebsz Anna

ELTE IK

Mennyi $sin(\pi)$ értéke?

`1.224646799147353e-016`

2. furcsa jelenség

Mennyi $\sum_{k=1}^{+\infty} \frac{1}{k}$ értéke?

Mennyi az n-edik részletösszeg, valamely nagy n-re? $\left(\sum_{k=1}^{n} \frac{1}{k}\right)$

Összegezhetünk oda vagy vissza ...
$n = 100000000$-re

`18.997896413852555`

`18.997896413853447`

3. furcsa jelenség

Mennyi $\sqrt{2017} - \sqrt{2016}$ értéke?
Más alakban is számolható:

$$\sqrt{2017} - \sqrt{2016} = (\sqrt{2017} - \sqrt{2016}) \cdot \frac{\sqrt{2017} + \sqrt{2016}}{\sqrt{2017} + \sqrt{2016}} =$$
$$= \frac{2017 - 2016}{\sqrt{2017} + \sqrt{2016}} = \frac{1}{\sqrt{2017} + \sqrt{2016}}.$$

Próbáljuk ki mindkét számolási módot!

```
0.011134504483941
0.016926965158418
```

A Matlab-ban

$$a = 1e - 20\,(= 10^{-20}), \;\; b = 1.$$

Mennyi lesz $a + b$ értéke?

```
1
```

Igaz-e az asszociativatás a Matlab-ban?

$$(a + b) - b, \;\; a + (b - b) =?$$

Próbáljuk ki!

```
1
1.000000000000000e-020
```

A Matlab-ban mennyi $\cosh(20) - \sinh(20)$ és $exp(-20)$ értéke?

$$\cosh(20) - \sinh(20) = \frac{\exp(20) + \exp(-20)}{2} - \frac{\exp(20) - \exp(-20)}{2} =$$
$$= \exp(-20)$$

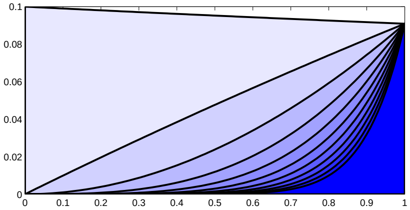
Próbáljuk ki a kétféle számítási módot!

```
0
2.061153622438558e-009
```

Mennyi a

$$T_n := \int_0^1 f_n(x) = \int_0^1 \frac{x^n}{x+10}\,dx$$

határozott integrál értéke? Analitikusan nehéz megadni az értékét. (A geometriai szemléltetésből látszik, hogy mindig pozitív és nullához tart az integrál értéke.)



$$T_n := \int_0^1 \frac{x^n}{x+10}\,dx = \int_0^1 \frac{(x+10-10)x^{n-1}}{x+10}\,dx =$$
$$= \int_0^1 x^{n-1}\,dx - 10\cdot\int_0^1 \frac{x^{n-1}}{x+10}\,dx = \frac{1}{n} - 10\cdot T_{n-1}$$

$$T_0 = \int\limits_0^1 \frac{1}{x+10}\,dx = [\ln(x+10)]_0^1 = \ln(11) - \ln(10) = \ln(1.1)$$

Tehát a rekuzió:

$$T_0 := \ln(1.1), \quad T_n := \frac{1}{n} - 10\cdot T_{n-1}\ \ (n = 1, 2\ldots).$$

Számoljuk a kapott rekurzió alapján a T_{20}. tagot Matlab-bal!

Rendezzük át a rekurziót csökkenően:

$$10T_{n-1} = \frac{1}{n} - T_n \quad \Leftrightarrow$$
$$T_{n-1} = \frac{1}{10} \cdot \left(\frac{1}{n} - T_n\right)$$

Indítsuk a rekurziót egy $M >> n$ értékből,

$$T_M := 0, \quad T_{n-1} = \frac{1}{10} \cdot \left(\frac{1}{n} - T_n\right) \ \ (n = M, \dots, m+1).$$

Számoljuk a második rekurzió alapján is a T_{20}. tagot! A két algoritmus közül melyik stabil?

```
7.483468021084803e+003
0.004347035818028
```

Definíció:

A *numerikus algoritmus* aritmetikai és logikai műveletek véges sorozata.

Definíció:

A numerikus algoritmus *stabil*, ha létezik olyan $C > 0$ konstans, hogy a kétféle B_1, B_2 bemenő adatból kapott K_1, K_2 kimenő adatokra

$$\|K_1 - K_2\| \leq C \cdot \|B_1 - B_2\|.$$

Példa

A Fibonacci sorozat rekurziója instabil. Lásd gyakorlaton.

Tartalomjegyzék

1. „Furcsa” jelenségek. . .
2. Gépi számok: a lebegőpontos számok egy modellje
3. A hibaszámítás elemei

- Gyakorlati és tudományos számításokban sokszor szükségünk van valós számok kezelésére.
- A számítógépeken csak egy véges halmaz elemei közül választhatunk.
- Ráadásul ezek több nagyságrenddel eltérhetnek.

Lebegőpontos számok, normalizált alak: 324 $\quad\rightsquigarrow\quad +0.324 \cdot 10^3$.
Kettes számrendszerben: 101000100 $\quad\rightsquigarrow\quad +0.101000100 \cdot 2^9$.
Általában: $\pm 0.\underbrace{1\ \_\ \_\ \ldots\ \_}_{t\ \text{jegy}} \cdot 2^k \qquad (k^- \leq k \leq k^+)$.

Definíció: Normalizált lebegőpontos szám

Legyen $m = \sum_{i=1}^{t} m_i \cdot 2^{-i}$, ahol $t \in \mathbb{N}$, $m_1 = 1, m_i \in \{0, 1\}$.
Ekkor az $a = \pm m \cdot 2^k$ ($k \in \mathbb{Z}$) alakú számot *normalizált lebegőpontos számnak* nevezzük.
m: a szám *mantisszája*, hossza t
k: a szám *karakterisztikája*, $k^- \leq k \leq k^+$

Lebegőpontos számok egy modellje

Jelölés: $a = \pm[m_1 \ldots m_t | k] = \pm 0.m_1 \ldots m_t \cdot 2^k$.

Jelölés: $M = M(t, k^-, k^+)$ a gépi számok halmaza, adott $k^-, k^+ \in \mathbb{Z}$ és $t \in \mathbb{N}$ esetén. (Általában $k^- < 0$ és $k^+ > 0$.)

Definíció: Gépi számok halmaza

$$M(t, k^-, k^+) =$$
$$= \left\{ a = \pm 2^k \cdot \sum_{i=1}^{t} m_i \cdot 2^{-i} \; : \; \begin{array}{l} k^- \leq k \leq k^+, \\ m_i \in \{0, 1\}, m_1 = 1 \end{array} \right\} \bigcup \{0\}$$

Gyakorlatban még hozzávesszük: ∞, $-\infty$, `NaN`,...

Gépi számok tulajdonságai, nevezetes értékei

1. $\frac{1}{2} \leq m < 1$
2. M szimmetrikus a 0-ra.
3. M legkisebb pozitív eleme:
$$\varepsilon_0 = [100\dots0|k^-] = \frac{1}{2} \cdot 2^{k^-} = 2^{k^- -1}$$
4. M-ben az 1 után következő gépi szám és 1 különbsége:
$$\varepsilon_1 = [100\dots01|1] - [100\dots00|1] = 2^{-t} \cdot 2^1 = 2^{1-t}$$
5. M legnagyobb eleme:
$$M_\infty = [111\dots11|k^+] = 1.00\dots00 \cdot 2^{k^+} - 0.00\dots01 \cdot 2^{k^+} = \\ = (1-2^{-t}) \cdot 2^{k^+}$$
6. M elemeinek száma (számossága):
$$|M| = 2 \cdot 2^{t-1} \cdot (k^+ - k^- + 1) + 1$$

Példa

$M(3,-1,2)$ gépi számainak alakja: $\pm 0.1\_\,\_\cdot 2^k, \quad (-1 \leq k \leq 2)$

Elemei $k=0$ esetén: $0.100, 0.101, 0.110, 0.111,$ azaz $\frac{1}{2}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}$.

Valamint $k=-1$ esetén ezek fele, $k=1$ esetén ezek kétszerese, $k=2$ esetén ezek négyszerese. (Továbbá negatív előjellel...)

$\varepsilon_0 = [100|-1] = 0.100 \cdot 2^{-1} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0.25$

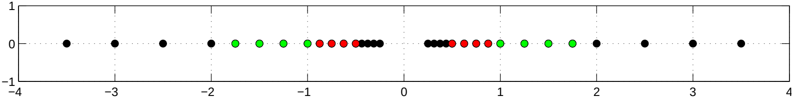
$\varepsilon_1 = [101|1] - 1 = 0.101 \cdot 2^1 - 1 = \frac{1}{8} \cdot 2 = \frac{1}{4} = 0.25$

$M_\infty = [111|2] = 0.111 \cdot 2^2 = \frac{7}{8} \cdot 4 = \frac{7}{2} = 3.5$

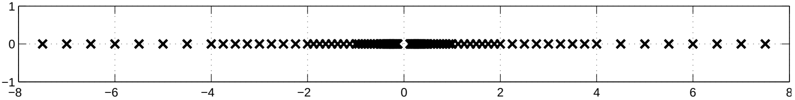
$|M| = 2 \cdot 2^2 \cdot 4 + 1 = 33$

Példa gépi számhalmazra

$M(3,-1,2)$



$M(4,-2,3)$



float $\sim M(23,-128,127)$, double $\sim M(52,-1024,1023)$

bitek, nevezetes értékek?

Hogyan feleltetünk meg egy $\mathbb{R}$-beli számnak egy gépi számot?
Jelöljük $\mathbb{R}_M$-mel az ábrázolható számok tartományát, azaz
$\mathbb{R}_M := \{x \in \mathbb{R} : \ |x| \le M_\infty\}$.

Definíció: Input függvény

Az $fl \colon \mathbb{R}_M \to M$ függvényt *input függvénynek* nevezzük, ha

$$fl(x) = \begin{cases} 0 & \text{ha } |x| < \varepsilon_0, \\ \tilde{x} & \text{ha } \varepsilon_0 \le |x| \le M_\infty, \end{cases}$$

ahol $\tilde{x}$ az x-hez legközelebbi gépi szám (a kerekítés szabályai szerint).

Tehát már az is egyfajta hibát okoz számításkor, hogy valós számokat számítógépre viszünk. . . de mekkorát?

Tétel: Input hiba

Minden $x \in \mathbb{R}_M$ esetén

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

Következmény: Input hiba

Ha $\varepsilon_0 \leq |x| \leq M_\infty$, akkor

$$\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2} \cdot \varepsilon_1 = 2^{-t}.$$

A hiba tehát lényegében ε_1-től, azaz t-től függ.

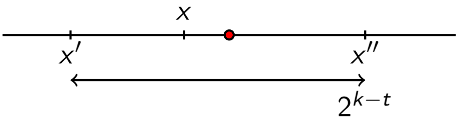
Mennyi a hiba, ha $|x| > M_\infty$?

Bizonyítás:

1. Ha $|x| < \varepsilon_0$, akkor $fl(x) = 0$, így $|x - fl(x)| = |x| < \varepsilon_0$.
2. Ha $|x| \geq \varepsilon_0$ és $x \in M$, akkor $fl(x) = x$, így $|x - fl(x)| = 0$.
3. A meggondolandó eset, amikor $|x| \geq \varepsilon_0$ és $x \notin M$.

Elegendő csak pozitív x-ekkel foglalkoznunk a 0-ra való szimmetria miatt. Keressük meg azt a két szomszédos gépi számot: $x' < x < x''$ és $x', x'' \in M$, amelyek közrefogják x-et.
Legyen $x' = [1\_ \ldots \_|k]$ alakú. Mennyi x' és x'' távolsága?

Ha x'-ben az utolsó helyiértékhez 1-et adunk, akkor x''-t kapjuk.
Tehát $x'' - x' = 2^{-t} \cdot 2^k = 2^{k-t}$.



Ha x az intervallum első felében van, akkor $fl(x) = x'$, ha a második felében, akkor $fl(x) = x''$. Ezért x és $fl(x)$ eltérése legfeljebb az intervallum fele, azaz $\frac{1}{2} \cdot 2^k \cdot 2^{-t}$. Vagyis

$$|x - fl(x)| \leq \frac{1}{2} \cdot 2^k \cdot 2^{-t}.$$

Viszont x abszolút értékére, fenti alakját figyelembe véve $0.1 \cdot 2^k = \frac{1}{2} \cdot 2^k \leq |x|$ is teljesül, ezért a becslést így folytathatjuk:

$$|x - fl(x)| \leq |x| \cdot 2^{-t} = \frac{1}{2} \cdot |x| \cdot \underbrace{2^{1-t}}_{\varepsilon_1} = \frac{1}{2} \cdot |x| \cdot \varepsilon_1.$$

$\square$

Definíció: Hibák jellemzése

Legyen A egy pontos érték, a pedig egy közelítő értéke. Ekkor:

$\Delta a := A - a$	a közelítő érték (pontos) hibája,
$\lvert\Delta a\rvert := \lvert A - a\rvert$	a közelítő érték abszolút hibája,
$\Delta_a \geq \lvert\Delta a\rvert$	az a egy abszolút hibakorlátja,
$\delta a := \frac{\Delta a}{A} \approx \frac{\Delta a}{a}$	az a relatív hibája,
$\delta_a \geq \lvert\delta a\rvert$	az a egy relatív hibakorlátja.

Példa

Vizsgáljuk meg a 3.14 számot mint a π két tizedesjegyre kerekített értékét!

Tétel: az alapműveletek hibakorlátai

$$\Delta_{a\pm b} = \Delta_a + \Delta_b \qquad \delta_{a\pm b} = \frac{|a|\cdot\delta_a + |b|\cdot\delta_b}{|a\pm b|}$$

$$\Delta_{a\cdot b} = |b|\cdot\Delta_a + |a|\cdot\Delta_b \qquad \delta_{a\cdot b} = \delta_a + \delta_b$$

$$\Delta_{a/b} = \frac{|b|\cdot\Delta_a + |a|\cdot\Delta_b}{b^2} \qquad \delta_{a/b} = \delta_a + \delta_b$$

Megjegyzés: a kapott korlátok két esetben lehetnek nagyságrendileg nagyobbak, mint a kiindulási értékek hibái:

1. $\delta_{a\pm b}$ esetén, amikor közeli számokat vonunk ki egymásból.
2. $\Delta_{a/b}$ esetén, amikor kicsi számmal osztunk.

Ezeket az eseteket az algoritmusok implementálásakor el kell kerülni.

Biz.: az összeadást és kivonást azonos előjelű számok között értjük. Az $a \pm b$ hibája

$$\Delta(a \pm b) = (A \pm B) - (a \pm b) = (A - a) \pm (B - b) = \Delta a \pm \Delta b$$

$$|\Delta(a \pm b)| = |\Delta a \pm \Delta b| \leq |\Delta a| + |\Delta b| \leq \Delta_a + \Delta_b = \Delta_{a \pm b}.$$

Biz.: az összeadás, kivonás hibakorlátai

Nézzük a relatív hibát

$$\frac{\Delta(a \pm b)}{a \pm b} = \frac{\Delta a \pm \Delta b}{a \pm b} = \frac{a \cdot \delta a \pm b \cdot \delta b}{a \pm b}$$

$$\frac{|\Delta(a \pm b)|}{|a \pm b|} = \frac{|a \cdot \delta a \pm b \cdot \delta b|}{|a \pm b|} \leq \frac{|a| \cdot |\delta a| + |b| \cdot |\delta b|}{|a \pm b|} \leq$$

$$\leq \frac{|a| \cdot \delta_a + |b| \cdot \delta_b}{|a \pm b|} = \delta_{a \pm b}$$

A szorzás hibája

$$\begin{aligned}\Delta(a\cdot b) &= A\cdot B - a\cdot b = A\cdot B - A\cdot b + A\cdot b - a\cdot b = \\ &= A(B-b) + b(A-a) = A\cdot\Delta b + b\cdot\Delta a = \\ &= (a+\Delta a)\cdot\Delta b + b\cdot\Delta a \approx a\cdot\Delta b + b\cdot\Delta a \\ &\ (\Delta a\cdot\Delta b \text{ elhanyagolható})\end{aligned}$$

$$|\Delta(a\cdot b)| \leq |a|\cdot|\Delta b| + |b|\cdot|\Delta a| \leq |a|\cdot\Delta_b + |b|\cdot\Delta_a = \Delta_{a\cdot b}$$

A relatív hiba

$$\delta(a\cdot b) = \frac{\Delta(a\cdot b)}{a\cdot b} \approx \frac{a\cdot\Delta b + b\cdot\Delta a}{a\cdot b} = \frac{\Delta b}{b} + \frac{\Delta a}{a} = \delta b + \delta a$$

$$|\delta(a\cdot b)| \leq |\delta a| + |\delta b| \leq \delta_a + \delta_b = \delta_{a\cdot b}$$

Az osztás hibája

$$\Delta\left(\frac{a}{b}\right) = \frac{A}{B} - \frac{a}{b} = \frac{A \cdot b - a \cdot B}{Bb} =$$
$$= \frac{A \cdot b - a \cdot b + a \cdot b - a \cdot B}{Bb} = \frac{b \cdot (A - a) - a \cdot (B - b)}{Bb} =$$
$$= \frac{b \cdot \Delta a - a \cdot \Delta b}{(b + \Delta b) \cdot b} \approx \frac{b \cdot \Delta a - a \cdot \Delta b}{b^2}$$

($\Delta b \cdot b$ elhanyagolható)

$$\left|\Delta\left(\frac{a}{b}\right)\right| \leq \frac{|b| \cdot |\Delta a| + |a| \cdot |\Delta b|}{b^2} \leq \frac{|b| \cdot \Delta_a + |a| \cdot \Delta_b}{b^2} = \Delta_{a/b}$$

Az osztás relatív hibája

$$\begin{aligned}\delta\left(\frac{a}{b}\right) &= \frac{\Delta(\frac{a}{b})}{\frac{a}{b}} \approx \frac{b\cdot\Delta a - a\cdot\Delta b}{b^2}\cdot\frac{b}{a} = \\ &= \frac{b\cdot\Delta a - a\cdot\Delta b}{b\cdot a} = \frac{\Delta a}{a} - \frac{\Delta b}{b} = \\ &= \delta a - \delta b = \delta\left(\frac{a}{b}\right)\end{aligned}$$

$\left|\delta\left(\frac{a}{b}\right)\right| \le |\delta a| + |\delta b| \le \delta_a + \delta_b = \delta_{a/b}$ $\square$

1. Tétel: a függvényérték hibája

Ha $f \in C^1(k_{\Delta_a}(a))$ és $k_{\Delta_a}(a) = [a - \Delta_a; a + \Delta_a]$, akkor

$$\Delta_{f(a)} = M_1 \cdot \Delta_a,$$

ahol $M_1 = \max \{ |f'(\xi)| : \xi \in k_{\Delta_a}(a) \}$.

Biz.: a Lagrange-féle középértéktétel felhasználásával.

$$\Delta f(a) = f(A) - f(a) = f'(\xi) \cdot (A - a) = f'(\xi) \cdot \Delta a,$$

valamely $\xi \in k_{\Delta_a}(a)$ értékre. Vizsgáljuk az abszolút hibát.

Jó felső becslést adva nyerjük az abszolút hibakorlátot:

$$|\Delta f(a)| = |f'(\xi)| \cdot |\Delta a| \leq M_1 \cdot \Delta_a = \Delta_{f(a)},$$

$\square$

2. Tétel: a függvényérték hibája

Ha $f \in C^2(k_{\Delta_a}(a))$ és $k_{\Delta_a}(a) = [a - \Delta_a; a + \Delta_a]$, akkor

$$\Delta_{f(a)} = |f'(a)|\, \Delta_a + \frac{M_2}{2} \cdot \Delta_a^2,$$

ahol $M_2 = \max\{\, |f''(\xi)| \,:\, \xi \in k_{\Delta_a}(a)\,\}$.

Biz.: a Taylor-formula felhasználásával.

$$\Delta f(a) = f(A) - f(a) = f'(a) \cdot (A - a) + \frac{f''(\xi)}{2} \cdot (A - a)^2,$$

valamely $\xi \in k_{\Delta_a}(a)$ értékre. Vizsgáljuk az abszolút hibát.

Jó felső becslést adva nyerjük az abszolút hibakorlátot:

$$|\Delta f(a)| = |f'(a)| \cdot |\Delta a| + \frac{|f''(\xi)|}{2} \cdot |\Delta a|^2 \leq$$
$$\leq |f'(a)| \cdot \Delta_a + \frac{M_2}{2} \cdot \Delta_a^2 = \Delta_{f(a)},$$

$\square$

Következmény: függvényérték relatív hibája

Ha Δ_a kicsi, akkor $\delta_{f(a)} = \frac{|a||f'(a)|}{|f(a)|} \cdot \delta_a$.

Definíció: Az f függvény a-beli kondíciószáma

A $c(f,a) = \frac{|a||f'(a)|}{|f(a)|}$ mennyiséget az f függvény a-beli kondíciószámának nevezzük.

Biz.: Ha Δ_a kicsi, akkor a 2. tételben szereplő eredményben a Δ_a^2-es tagot elhanyagolhatjuk, így felhasználva, hogy $\Delta_a = |a| \cdot \delta_a$

$$|\delta f(a)| \approx \frac{|f'(a)| \cdot \Delta_a}{|f(a)|} = \frac{|a|\, \delta_a \cdot |f'(a)|}{|f(a)|} = \frac{|a|\, |f'(a)|}{|f(a)|} \cdot \delta_a.$$

$\square$

Numerikus módszerek 1.

2. előadás: Lineáris egyenletrendszerek megoldása, Gauss-elimináció

Krebsz Anna

ELTE IK

Tartalomjegyzék

- **Általános iskolában:**
  Matematikai versenyfeladat 3. osztály

  A MATEK szó minden betűje egy-egy számjegyet jelöl. A számjegyekre igazak a következő állítások:

$$\begin{aligned} M + A + T + E + K &= 25 \\ M + A &= 11 \\ A + T &= 10 \\ T + E &= 12 \\ E + K &= 10 \end{aligned}$$

  Melyik betű melyik számjegyet jelöli, ha az öt betű öt különböző számjegyet jelöl?

- **Gazdasági számítások:**

  Tegyük fel, hogy egy üzem kétféle végterméket állít elő négyféle alkatrész felhasználásával. Jelölje $A_1, A2$ a végtermékeket, az A_3, A_4 a félkész termékeket és A_5, A_6 az alapanyagokat. Az egyes alapanyagok és félkész termékek egymásba és a végtermékbe való beépülését a **közvetlen ráfordítás mátrix (K)** adja meg. A mátrix k_{ij} eleme azt mutatja, hogy az i. termékből közvetlenül (nem más terméken keresztül) mennyi épül be a j. termékbe.

  A **teljes ráfordítások mátrixában (T)** a t_{ij} elem azt mutatja, hogy egy darab A_j termék összesen hány darab A_i elemet tartalmaz. Ennek meghatározása a $T = (I - K)^{-1}$ képletből történik. A kétféle mátrix alkalmazása:
  x alapanyagból $y = (I - K) \cdot x$ végtermék lesz és
  y végtermékhez $x = T \cdot y = (I - K)^{-1} \cdot y$ alapanyag kell.

3-5. Példa

- Mérnöki feladatok numerikus megoldása (lásd a bevezető példát)
- Interpolációs spline-ok megadása (lásd 2. félév)
- Approximációs feladatok megoldása (lásd 2. félév)

- **Hálózatok stacionárius modellezése:** villamos hálózatok, áramkörök, víz- és gázellátó csőrendszerek irányított gráffal történő leírása után. Az él iránya megfelel a várt áramlási iránynak. Minden élhez tartozik egy szám, az ott szállított áram (víz stb.) mennyiségét adja. Egyes csomópontokhoz is tartozhat áram, ezek a külső pontok. Ilyen áram a ponton keresztül be ill. kifolyó áram, amely ugyancsak ismeretlen lehet.

  A gráf minden csomópontjában felírjuk az első Kirchhoff–féle törvényt, amely szerint - figyelembe véve az élek irányát - a csomópontban találkozó élek áramainak összege nulla. Ez az anyag-megmaradási törvény egy lineáris reláció, és a minden csomóponthoz tartozó relációk összessége adja a lineáris egyenletrendszert.

Lineáris egyenletrendszer (LER)

Hagyományos alak:

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

n egyenlet, n ismeretlen

Mátrix alak:

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

vagyis

$$Ax = b \qquad A \in \mathbb{R}^{n\times n},\ b, x \in \mathbb{R}^n.$$

Feladat:

A és b adottak, keressük x-et.

Tétel: emlékeztető lin. alg.-ból

- LER megoldható $\iff$ b felírható az A oszlopvektorainak lineáris kombinációjaként.
- Egyértelműen létezik megoldás $\iff$ A oszlopai lineárisan függetlenek $\iff$ $\mathrm{rang}(A) = n \iff \det(A) \neq 0 \iff A$ invertálható ($x = A^{-1}b$).

Megj.:

- Ha A speciális alakú (pl. diagonális vagy háromszög alakú), akkor egyszerűen megkapható a megoldás.
- Cramer-szabályt max. 3×3-as mátrixokra alkalmazunk.

Lineáris egyenletrendszerek megoldási módszerei

- Direkt módszerek, felbontások (véges lépésszám, „pontos" megoldás)
  - Gauss-elimináció, progonka módszer
  - LU-felbontás, LDU, $LL^\top$, Cholesky
  - QR-felbontás (Gram–Schmidt ort., Householder trf.)
  - ILU-felbontás
- Iterációs módszerek (vektor sorozat, mely a megoldáshoz „tart")
  - mátrixnormák, Banach-féle fixponttétel
  - Jacobi-iteráció
  - Gauss–Seidel-iteráció
  - Richardson-iteráció
  - ILU-algoritmus
- Variációs módszerek (egy „célfüggvény" minimalizálása által)
  - Gradiens-módszer
  - Konjugált gradiens-módszer

Legyen $a_{in+1} := b_i$, azaz $[A|b]$ a tárolási forma.
GE := Gauss-elimináció.

$$A^{(0)} := \left[\begin{array}{cccc|r} a_{11} & a_{12} & \cdots & a_{1n} & a_{1n+1} = b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & a_{2n+1} = b_2 \\ \vdots & & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & a_{nn+1} = b_n \end{array}\right]$$

Célunk: A LER-t egyszerűbb alakra hozni:

1. balról jobbra: a főátló alatt kinullázzuk az elemeket, „előre", GE
2. jobbról balra: a főátló fölött nullázunk, „vissza", visszahelyettesítés

Gauss-elimináció: 1. lépés

Az 1. egyenletet változatlanul hagyjuk.

Ha $a_{11}^{(0)} \neq 0$, akkor az i-edik egyenletből ($i = 2, 3, \ldots, n$) kivonjuk az 1. egyenlet $\left(\frac{a_{i1}^{(0)}}{a_{11}^{(0)}}\right)$-szeresét: hogy $a_{i1}^{(0)}$ kinullázódjon.
($\rightsquigarrow$ elimináció, kiküszöbölés)

$$A^{(1)} = \left[\begin{array}{ccccc|c} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1n}^{(0)} & a_{1n+1}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} & a_{2n+1}^{(1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \cdots & a_{nn}^{(1)} & a_{nn+1}^{(1)} \end{array}\right],$$

ahol

$$a_{ij}^{(1)} = a_{ij}^{(0)} - \frac{a_{i1}^{(0)}}{a_{11}^{(0)}} \cdot a_{1j}^{(0)} \qquad (i = 2, \ldots, n;\ j = 2, \ldots, n, n+1).$$

Gauss-elimináció: 2. lépés

Az 1. és 2. egyenletet változatlanul hagyjuk.

Ha $a_{22}^{(1)} \neq 0$, akkor az i-edik egyenletből $(i = 3, 4, \ldots, n)$ kivonjuk a 2. egyenlet $\left(\frac{a_{i2}^{(1)}}{a_{22}^{(1)}}\right)$-szeresét: hogy $a_{i2}^{(1)}$ kinullázódjon.

$$A^{(2)} = \left[\begin{array}{cccccc|c} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1n}^{(0)} & a_{1n+1}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} & a_{2n+1}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} & a_{3n+1}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} & a_{nn+1}^{(2)} \end{array}\right],$$

ahol

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}} \cdot a_{2j}^{(1)} \qquad (i = 3, \ldots, n;\ j = 3, \ldots, n, n+1).$$

Gauss-elimináció: k. lépés

Az $1., 2., \ldots, k$. egyenleteket változatlanul hagyjuk.

Ha $a_{kk}^{(k-1)} \neq 0$, akkor az i-edik egyenletből $(i = k+1, \ldots, n)$ kivonjuk a k-adik egyenlet $\left(\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}\right)$-szeresét: hogy $a_{ik}^{(k-1)}$ kinullázódjon. Ezt a lépést láttuk, amikor a 2. lépésben az 1. lépés eredményét felhasználtuk. Ha 2 helyére k-t írunk, akkor megkapjuk az általános képleteket.

Tétel: A Gauss-elimináció általános lépése

Ha $a_{k,k}^{(k-1)} \neq 0$, akkor a k. lépés képletei

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \cdot a_{kj}^{(k-1)} \qquad \begin{array}{l} k = 1, \ldots, n-1; \\ i = k+1, \ldots, n; \\ j = k+1, \ldots, n, n+1. \end{array}$$

Így $n-1$ lépés után felső háromszögmátrix alakú LER-t kapunk:

$$A^{(n-1)} = \left[\begin{array}{ccccc|c} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1n-1}^{(0)} & a_{1n}^{(0)} & a_{1n+1}^{(0)} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n-1}^{(1)} & a_{2n}^{(1)} & a_{2n+1}^{(1)} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ddots & a_{n-1n-1}^{(n-2)} & a_{n-1n}^{(n-2)} & a_{n-1n+1}^{(n-2)} \\ 0 & 0 & \ldots & 0 & a_{nn}^{(n-1)} & a_{nn+1}^{(n-1)} \end{array}\right].$$

Ezután visszafelé haladva: az aktuális egyenletet osztjuk a főátlóbeli elemmel, majd a főátló fölött kinullázzuk az elemeket, az eddigiekel analóg „sorműveletek” alkalmazásával.
Végül $[I|x]$ alakot nyerünk. ($I \in \mathbb{R}^{n\times n}$ egységmátrix.)

Gauss-elimináció: visszahelyettesítés

Az algoritmus második része („jobbról-balra”), a felső háromszög alakú LER megoldása képlettel is kifejezhető. Figyeljük meg, hogy a felső-háromszögmátrixú alaknál soronként azonos felső indexek vannak.

A visszahelyettesítés

$$x_n = \frac{a_{nn+1}^{(n-1)}}{a_{nn}^{(n-1)}},$$

$$x_i = \frac{1}{a_{ii}^{(i-1)}} \left(a_{in+1}^{(i-1)} - \sum_{j=i+1}^{n} a_{ij}^{(i-1)} \cdot x_j \right) \qquad (i = n-1, \ldots, 1).$$

Példa: LER megoldása GE-val

Oldjuk meg a következő lineáris egyenletrendszert Gauss-elimináció alkalmazásával!

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

Az elimináció: Kézi számolásnál függőleges vonalat húzunk a jobboldali vektor elé, számítógéppel ezt programozással oldjuk meg.

1. lépés:

$$2.\ \text{sor} \underbrace{-\left(\frac{-4}{2}\right)}_{+2} * 1.\ \text{sor}$$

$$3.\ \text{sor} - \underbrace{\left(\frac{6}{2}\right)}_{+3} * 1.\ \text{sor}$$

$$\left[\begin{array}{rrr|r} 2 & 0 & 3 & -1 \\ -4 & 5 & -2 & 3 \\ 6 & -5 & 4 & -3 \end{array}\right] \rightarrow \left[\begin{array}{rrr|r} 2 & 0 & 3 & -1 \\ 0 & 5 & 4 & 1 \\ 0 & -5 & -5 & 0 \end{array}\right] \rightarrow$$

2. lépés:

$$3.\ \text{sor}\ \underbrace{-\ \left(\frac{-5}{5}\right)}_{+1}\ *\ 2.\ \text{sor}$$

$$\left[\begin{array}{ccc|c} 2 & 0 & 3 & -1 \\ 0 & 5 & 4 & 1 \\ 0 & -5 & -5 & 0 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 2 & 0 & 3 & -1 \\ 0 & 5 & 4 & 1 \\ 0 & 0 & -1 & 1 \end{array}\right] \rightarrow$$

A visszahelyettesítés:

3. sor $/(-1)$
2. sor $-$ 4 $*$ új 3. sor.
1. sor $-$ 3 $*$ új 3. sor.

$$\left[\begin{array}{ccr|r} 2 & 0 & 3 & -1 \\ 0 & 5 & 4 & 1 \\ 0 & 0 & -1 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|r} 2 & 0 & 0 & 2 \\ 0 & 5 & 0 & 5 \\ 0 & 0 & 1 & -1 \end{array}\right]$$

2. sor $/5$
1. sor $/2.$

$$\left[\begin{array}{ccc|r} 2 & 0 & 0 & 2 \\ 0 & 5 & 0 & 5 \\ 0 & 0 & 1 & -1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|r} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{array}\right]$$

Tehát a lineáris egyenletrendszer megoldása az $\mathbf{x} = [1, 1, -1]^T$ vektor.

- LER megoldása (láttuk példán is)
- Determináns meghatározása: mivel a GE lépései determináns tartók, ezért

$$\det(A) = \det(\Delta\text{alak}) = \prod_{k=1}^{n} a_{kk}^{(k-1)}$$

  Vigyázzunk : ha sort vagy oszlopot cserélünk, a determináns értéke változik.

- Több jobb oldallal (b) megoldás: lehet egyszerre, így a mátrixon csak egyszer eliminálunk.

$$[A|b_1|b_2|b_3] \rightarrow \text{GE} \rightarrow \text{visszahely} \rightarrow [I|x_1|x_2|x_3]$$

- Mátrix inverzének meghatározása az $A \cdot X = I$ mátrixegyenlet megoldását jelenti.

$$A \cdot [x_1 | \dots | x_n] = [e_1 | \dots | e_n] \quad \Leftrightarrow \quad \begin{array}{rl} Ax_1 & = e_1 \\ \dots & \\ Ax_n & = e_n \end{array}$$

Visszavezettük az előző pontra. A GE-t kiterjesztett mátrixon hajtjuk végre

$$[A \,|\, I] \rightarrow \mathrm{GE} \rightarrow \text{visszahely} \rightarrow [I \,|\, A^{-1}],$$

visszahelyettesítés után jobb oldalon kapjuk az inverz mátrixot. Sor csere esetén az inverz nem változik, oszlopcsere esetén változik (lásd gyak.).

Példa: mátrix determinánsának és inverzének számítása GE-val

Mi az előző példa mátrixának determinánsa és inverze?

$$\det(A) = \det(\Delta\text{alak}) = \begin{vmatrix} 2 & 0 & 3 \\ 0 & 5 & 4 \\ 0 & 0 & -1 \end{vmatrix} = 2 \cdot 5 \cdot (-1) = -10$$

Az elimináció:
1. lépés:

$$2.\ \text{sor} \underbrace{-\left(\frac{-4}{2}\right)}_{+2} * 1.\ \text{sor}$$

$$3.\ \text{sor} - \underbrace{\left(\frac{6}{2}\right)}_{3} * 1.\ \text{sor}$$

$$\left[\begin{array}{rrr|rrr} 2 & 0 & 3 & 1 & 0 & 0 \\ -4 & 5 & -2 & 0 & 1 & 0 \\ 6 & -5 & 4 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{rrr|rrr} 2 & 0 & 3 & 1 & 0 & 0 \\ 0 & 5 & 4 & 2 & 1 & 0 \\ 0 & -5 & -5 & -3 & 0 & 1 \end{array}\right] \to$$

2. lépés:

3. sor $\underbrace{-\left(\frac{-5}{5}\right)}_{+1}$ $*$ 2. sor

$$\left[\begin{array}{ccc|ccc} 2 & 0 & 3 & 1 & 0 & 0 \\ 0 & 5 & 4 & 2 & 1 & 0 \\ 0 & -5 & -5 & -3 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccc|ccc} 2 & 0 & 3 & 1 & 0 & 0 \\ 0 & 5 & 4 & 2 & 1 & 0 \\ 0 & 0 & -1 & -1 & 1 & 1 \end{array}\right] \to$$

A visszahelyettesítés:

3. sor $/(-1)$
2. sor $-$ 4 $*$ új 3. sor.
1. sor $-$ 3 $*$ új 3. sor.

$$\left[\begin{array}{ccc|ccc} 2 & 0 & 3 & 1 & 0 & 0 \\ 0 & 5 & 4 & 2 & 1 & 0 \\ 0 & 0 & -1 & -1 & 1 & 1 \end{array}\right] \to \left[\begin{array}{ccc|ccc} 2 & 0 & 0 & -2 & 3 & 3 \\ 0 & 5 & 0 & -2 & 5 & 4 \\ 0 & 0 & 1 & 1 & -1 & -1 \end{array}\right]$$

2. sor $/5$
1. sor $/2$.

$$\left[\begin{array}{ccc|ccc} 2 & 0 & 0 & -2 & 3 & 3 \\ 0 & 5 & 0 & -2 & 5 & 4 \\ 0 & 0 & 1 & 1 & -1 & -1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -1 & \frac{3}{2} & \frac{3}{2} \\ 0 & 1 & 0 & -\frac{2}{5} & 1 & \frac{4}{5} \\ 0 & 0 & 1 & 1 & -1 & -1 \end{array}\right] = [I|A^{-1}]$$

Az inverz a jobb oldalon álló mátrix.

Gauss-elimináció: megjegyzések

Megoldható-e egyáltalán a LER? Vizsgáljuk?

Majd GE közben kiderül.

Megoldható, de mégsem tudjuk a GE-t végigcsinálni?

Előfordulhat... $\rightsquigarrow$ *sort cserélünk* $\rightsquigarrow$ *nem változik a megoldás. Ha oszlopot cserélünk, akkor a megoldás komponensei a cserének megfelelően változnak.*

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot x = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

Biztos és stabil megoldás a főelemkiválasztás.

Definíció: részleges főelemkiválasztás

A k-adik lépésben válasszunk egy olyan m indexet, melyre $\left|a_{mk}^{(k-1)}\right|$ maximális ($m \in \{ k, k+1, \ldots, n \}$), majd cseréljük ki a k-adik és m-edik sort.

Definíció: teljes főelemkiválasztás

A k-adik lépésben válasszunk egy olyan (m_1, m_2) indexpárt, melyre $\left|a_{m_1 m_2}^{(k-1)}\right|$ maximális ($m_1, m_2 \in \{ k, k+1, \ldots, n \}$), majd cseréljük ki a k-adik és m_1-edik sort, valamint a k-adik és m_2-edik oszlopot.

Tétel:

A GE elvégezhető sor és oszlopcsere nélkül
$\Leftrightarrow \quad a_{kk}^{(k-1)} \neq 0 \;\; (k = 1, 2, \ldots, n-1)$.

Biz.: trivi a rekurzióból. $\square$

Definíció: főminorok

Az A főminorai a

$$D_k = \det\left(\begin{bmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \ldots & a_{kk} \end{bmatrix}\right), \quad (k = 1, 2, \ldots, n)$$

determinánsok. Ezek az A bal felső $k \times k$-s részmátrixaimak determinánsai.

Tétel:

$D_k \neq 0 \;\; (k = 1, 2, \ldots, n-1) \quad \Leftrightarrow \quad a_{kk}^{(k-1)} \neq 0 \;\; (k = 1, 2, \ldots, n-1)$.

Biz.: A GE átalakításai determináns tartók, ezért

$$D_k = a_{11} \cdot a_{22}^{(1)} \cdot \ldots \cdot a_{kk}^{(k-1)} = D_{k-1} \cdot a_{kk}^{(k-1)},$$

amiből az állítás adódik. A $D_n \neq 0$ illetve az $a_{nn}^{(n-1)} \neq 0$ feltétel nem szükséges a GE-hoz, csak a LER megoldhatóságához. $\square$

Megj.:

- Numerikus szempontból jobb, ha alkalmazunk főelemkiválasztást. Ezzel a GE-s hányadosaink pontosabbak lesznek.
- Determináns számításakor a cserékkel vigyázni kell!

Tétel: A Gauss-elimináció műveletigénye

$$\frac{2}{3}n^3 + \mathcal{O}(n^2)$$

Biz.: Rögzített k-ra: a k. lépés képletéből számolva

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \cdot a_{kj}^{(k-1)} \qquad \begin{array}{l} k = 1, \dots, n-1; \\ i = k+1, \dots, n; \\ j = k+1, \dots, n, n+1. \end{array}$$

$(n-k)$ osztás, $(n-k)(n-k+1)$ szorzás és $(n-k)(n-k+1)$ összeadás kell.
Összesen $(n-k)(2(n-k)+3)$ művelet. $(n-k =: s)$

$$\sum_{k=1}^{n-1}(n-k)(2(n-k)+3)=\sum_{s=1}^{n-1}s(2s+3)=2\sum_{s=1}^{n-1}s^2+3\sum_{s=1}^{n-1}s=$$

$$=2\frac{(n-1)n(2n-1)}{6}+3\frac{(n-1)n}{2}=\frac{2}{3}n^3+\mathcal{O}(n^2). \quad \square$$

Definíció: $\mathcal{O}(n^2)$ függvény

Az $f(n)$ függvényt $\mathcal{O}(n^2)$-es nagyságrendűnek nevezzük, ha $\frac{f(n)}{n^2}$ korlátos minden $n \in \mathbb{N}$-re.

A felső háromszögmátrixú LER megoldásának műveletigénye.

Tétel: A visszahelyettesítés műveletigénye

$$n^2 + \mathcal{O}(n)$$

Biz.:

$$x_n = \frac{a_{nn+1}^{(n-1)}}{a_{nn}^{(n-1)}}, \quad x_i = \frac{1}{a_{ii}^{(i-1)}} \left(a_{in+1}^{(i-1)} - \sum_{j=i+1}^{n} a_{ij}^{(i-1)} \cdot x_j \right) \quad (i = n-1, \ldots, 1).$$

Rögzített i. sorra 1 db osztás, $(n-i)$ szorzás és $(n-i)$ összeadás.
Összesen: $2(n-i)+1$ művelet $(n-i =: s)$.

$$1 + \sum_{s=1}^{n-1} (2s+1) = 1 + 2 \cdot \frac{n(n-1)}{2} + (n-1) = n^2 + \mathcal{O}(n). \quad \square$$

Példák Matlab-ban

1. A Gauss-elimináció működése „kisebb” ($n \approx 7$) LER-ekre
2. A beépített megoldó rutin persze sokkal gyorsabb
3. Egyre nagyobb méretű ($n = 10, 20, 30, \ldots, 200$) mátrixokra a GE futási idejének viselkedése tényleg n^3-szerű

Numerikus módszerek 1.

3. előadás: Mátrixok LU-felbontása

Krebsz Anna

ELTE IK

Tartalomjegyzék

Balról szorzás alsó háromszögmátrixokkal

Mi történik, ha az alábbi $L \in \mathbb{R}^{3\times 3}$ mátrixszal megszorzunk egy $A \in \mathbb{R}^{3\times 3}$ mátrixot balról?

$$\begin{array}{cc} & \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{bmatrix} \end{array}$$

Az 1. sor kétszeresét hozzáadjuk a 2. sorhoz.

Balról szorzás alsó háromszögmátrixokkal

Mi történik, ha az alábbi $L \in \mathbb{R}^{3\times 3}$ mátrixszal megszorzunk egy $A \in \mathbb{R}^{3\times 3}$ mátrixot balról?

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix}$$

Az 1. sor kétszeresét hozzáadjuk a 2. sorhoz, valamint az 1. sor háromszorosát levonjuk a 3. sorból. ($\sim$ GE 1. lépése volt)

A Gauss-elimináció lépései mátrixszorzással

Írjuk fel a GE k-adik lépését ugyanilyen módszerrel! ($A \in \mathbb{R}^{n\times n}$)

$$L_k = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{k+1k} & 1 & & \\ & & \vdots & & \ddots & \\ & & -l_{nk} & & & 1 \end{pmatrix} = I - \ell_k e_k^\top, \quad \ell_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ l_{k+1k} \\ \vdots \\ l_{nk} \end{pmatrix}.$$

(A zérus elemek nincsenek feltüntetve L_k-ban.)

Tehát ha $l_{ik} = \dfrac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \quad (k = 1, \ldots, n-1;\;\; i = k+1, \ldots, n)$,

akkor $L_k \cdot A^{(k-1)} = A^{(k)}$, vagyis megkaptuk a GE k-adik lépését.

Példa: GE az L_k mátrixokkal

Írjuk fel a Gauss-elimináció lépéseit mátrixszorzások segítségével a következő mátrix esetén (ua. mint az előző előadáson)!

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

Megoldás: **1. lépés**

$$A^{(1)} = L_1 \cdot A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 & 3 \\ -4 & 5 & -2 \\ 6 & -5 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 3 \\ 0 & 5 & 4 \\ 0 & -5 & -5 \end{bmatrix}$$

2. lépés

$$A^{(2)} = L_2 \cdot A^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 & 3 \\ 0 & 5 & 4 \\ 0 & -5 & -5 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 3 \\ 0 & 5 & 4 \\ 0 & 0 & -1 \end{bmatrix} =: U$$

Tehát $A^{(2)} = L_2 \cdot L_1 \cdot A =: U$, a kapott felsőháromszög alakot U-val jelöljük.
Fejezzük ki A-t a képletből:

$$A = \underbrace{L_1^{-1} \cdot L_2^{-1}}_{=:L} \cdot U = L \cdot U.$$

Ezzel megkaptuk az A mátrix LU-felbontását. Ennek az elméletét tárgyaljuk a következőkben.

Definíció: alsó háromszögmátrix

Az $L \in \mathbb{R}^{n\times n}$ mátrixot *alsó háromszögmátrixnak* nevezzük, ha $i < j$ esetén $l_{ij} = 0$. (A főátló felett csupa nulla.)

$$\mathcal{L} := \{\, L \in \mathbb{R}^{n\times n} \,:\, l_{ij} = 0 \;(i < j)\,\},$$
$$\mathcal{L}_1 := \{\, L \in \mathbb{R}^{n\times n} \,:\, l_{ij} = 0 \;(i < j),\;\; l_{ii} = 1\,\}.$$

Definíció: felső háromszögmátrix

Az $U \in \mathbb{R}^{n\times n}$ mátrixot *felső háromszögmátrixnak* nevezzük, ha $i > j$ esetén $u_{ij} = 0$. (A főátló alatt csupa nulla.)

$$\mathcal{U} := \{\, U \in \mathbb{R}^{n\times n} \,:\, u_{ij} = 0 \;(i > j)\,\},$$
$$\mathcal{U}_1 := \{\, U \in \mathbb{R}^{n\times n} \,:\, u_{ij} = 0 \;(i > j),\;\; u_{ii} = 1\,\}.$$

Állítás: háromszögmátrixról

1. Ha $L', L'' \in \mathcal{L}$, akkor $L' \cdot L'' \in \mathcal{L}$.
2. Ha $U', U'' \in \mathcal{U}$, akkor $U' \cdot U'' \in \mathcal{U}$.
3. Ha $L', L'' \in \mathcal{L}_1$, akkor $L' \cdot L'' \in \mathcal{L}_1$.
4. Ha $U', U'' \in \mathcal{U}_1$, akkor $U' \cdot U'' \in \mathcal{U}_1$.
5. Ha $L \in \mathcal{L}$ és $\exists L^{-1}$, akkor $L^{-1} \in \mathcal{L}$.
6. Ha $U \in \mathcal{U}$ és $\exists U^{-1}$, akkor $U^{-1} \in \mathcal{U}$.
7. Ha $L \in \mathcal{L}_1$, akkor $\exists L^{-1}$ és $L^{-1} \in \mathcal{L}_1$.
8. Ha $U \in \mathcal{U}_1$, akkor $\exists U^{-1}$ és $U^{-1} \in \mathcal{U}_1$.

Biz.: házi feladat (beadható). $\square$

Definíció: L_k

$L_k := I - \ell_k e_k^\top \in \mathbb{R}^{n\times n}$, ahol $\ell_k \in \mathbb{R}^n$, $(\ell_k)_i = 0$ $(i \le k)$ és $e_k \in \mathbb{R}^n$ a k-adik egységvektor.

Állítás: L_k inverze

$$L_k^{-1} = I + \ell_k e_k^\top .$$

Biz.:

$$L_k \cdot L_k^{-1} = (I - \ell_k e_k^\top)(I + \ell_k e_k^\top) = I \underbrace{-\ell_k e_k^\top + \ell_k e_k^\top}_{0} - \ell_k \underbrace{e_k^\top \ell_k}_{0} e_k^\top = I. \quad \square$$

Szemléletesen?

Az L_k mátrixokról

Hogyan szorzunk össze két ilyen mátrixot?

$$
\begin{array}{cc}
 & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix} \\
\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} & \begin{pmatrix} 1 & 0 & 0 \\ \color{red}{2} & 1 & 0 \\ \color{red}{1} & \color{red}{3} & 1 \end{pmatrix}
\end{array}
\qquad\qquad
\begin{array}{cc}
 & \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \\
\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix} & \begin{pmatrix} 1 & 0 & 0 \\ \color{red}{2} & 1 & 0 \\ \color{red}{7} & \color{red}{3} & 1 \end{pmatrix}
\end{array}
$$

A bal oldali sorrendben „szépen” szorzódik. Általában is.

Állítás: L_k mátrixok szorzata

$$L_1^{-1} \cdot L_2^{-1} \cdots L_{n-1}^{-1} = I + \ell_1 e_1^\top + \ell_2 e_2^\top + \ldots + \ell_{n-1} e_{n-1}^\top.$$

Szemléletesen?

Biz.: Indukcióval.

- $$\begin{aligned} L_1^{-1} \cdot L_2^{-1} &= (I + \ell_1 e_1^\top)(I + \ell_2 e_2^\top) = \\ &= I + \ell_1 e_1^\top + \ell_2 e_2^\top + \ell_1 \underbrace{(e_1^\top \ell_2)}_{0} e_2^\top = \\ &= I + \ell_1 e_1^\top + \ell_2 e_2^\top \end{aligned}$$

- Tegyük fel, hogy $k+1 \leq n-1$ és

$$L_1^{-1} \cdot L_2^{-1} \cdots L_k^{-1} = I + \ell_1 e_1^\top + \ell_2 e_2^\top + \ldots + \ell_k e_k^\top.$$

- $L_1^{-1} \cdot L_2^{-1} \cdots L_k^{-1} \cdot L_{k+1}^{-1} =$

$$= (I + \ell_1 e_1^\top + \ell_2 e_2^\top + \ldots + \ell_k e_k^\top)(I + \ell_{k+1} e_{k+1}^\top) =$$

$$= I + \ell_1 e_1^\top + \ell_2 e_2^\top + \ldots + \ell_k e_k^\top + \ell_{k+1} e_{k+1}^\top +$$

$$+ \underbrace{\ell_1 e_1^\top \ell_{k+1} e_{k+1}^\top + \ldots + \ell_k e_k^\top \ell_{k+1} e_{k+1}^\top}_{\text{kiesnek}} =$$

$$= I + \ell_1 e_1^\top + \ell_2 e_2^\top + \ldots + \ell_k e_k^\top + \ell_{k+1} e_{k+1}^\top = \checkmark. \qquad \square$$

Tartalomjegyzék

Definíció: LU-felbontás

Az A mátrix LU-felbontásának nevezzük az $L \cdot U$ szorzatot, ha

$$A = LU, \quad L \in \mathcal{L}_1, \; U \in \mathcal{U}.$$

A Gauss-eliminációt felírhatjuk alsó háromszögmátrixok segítségével:

$$L_{n-1} \cdots L_2 \cdot L_1 \cdot A = U,$$

majd az inverzekkel egyesével átszorozva:

$$A = \underbrace{L_1^{-1} \cdot L_2^{-1} \cdots L_{n-1}^{-1}}_{L} \cdot U = LU.$$

A fenti szorzat is alsó háromszögmátrix. Láttuk az előző tételből, hogy az L mátrix elemeit egy egységmátrixból kapjuk úgy, hogy minden oszlopba ez egyesek alá beletesszük a neki megfelelő ℓ_k vektor nem nulla elemeit (ezek a GE-s hányadosok). Tehát ennek előállításához nem kell több művelet, mint amit a GE-val végzünk.

Példa: LU-felbontás GE-val

Készítsük el a példamátrixunk LU-felbontását

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

a) részletezve az L_k mátrixokat, a számítás menetét,

b) majd „tömör” írásmóddal!

Megoldás: (a) 1. lépés

$$A^{(1)} = L_1 \cdot A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 & 3 \\ -4 & 5 & -2 \\ 6 & -5 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 3 \\ 0 & 5 & 4 \\ 0 & -5 & -5 \end{bmatrix}$$

L_1^{-1}-et úgy kapjuk, hogy L_1 1. oszlopában az átló alatti elemeket (-1)-szeresére változtatjuk. Megfigyelhetjük, hogy ezek a tényleges GE-s hányadosok. Láttuk, hogy L meghatározáshoz csak ℓ_1-re van szükségünk.

$$L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}$$

2. lépés

$$A^{(2)} = L_2 \cdot A^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 & 3 \\ 0 & 5 & 4 \\ 0 & -5 & -5 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 3 \\ 0 & 5 & 4 \\ 0 & 0 & -1 \end{bmatrix} =: U$$

L_2^{-1}-et úgy kapjuk, hogy L_2 2. oszlopában az átló alatti elemeket (-1)-szeresére változtatjuk. Megfigyelhetjük, hogy ez a tényleges GE-s hányados. Láttuk, hogy L meghatározáshoz csak ℓ_2-re van szükségünk.

$$L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

Példa

Tehát $A^{(2)} = L_2 \cdot L_1 \cdot A =: U$

Fejezzük ki A-t a képletből:

$$A = \underbrace{L_1^{-1} \cdot L_2^{-1}}_{=:L} \cdot U = L \cdot U.$$

Tehát $L = L_1^{-1} \cdot L_2^{-1}$. Az L_k mátixok szorzatára felírt tétel alapján ehhez nem kell mátrixot szoroznunk, csak az ℓ_k vektorokból kell összeraknunk L-et.

$$L = L_1^{-1} \cdot L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & -1 & 1 \end{bmatrix}$$

A kapott eredményt szorzással is ellenőrizhetjük.

(b) Tömör írásmódban: 1. lépés

A GE-s hányadosokat minden lépésben az eliminált pozíciókon tudjuk tárolni (éppen ennyi nulla van az oszlopban). Könnyen megjegyezhető ezek képzése: az eliminálandó mátrix rész 1. oszlopában az első elemmel leosztjuk az alatta levőket. Ezzel minden a helyére került. Vonalakkal jelezzük, hogy itt már tárolásról is szó van. A jobb alsó 2×2-es mátrix részen elvégezzük az eliminációt.

$$\begin{bmatrix} 2 & 0 & 3 \\ -4 & 5 & -2 \\ 6 & -5 & 4 \end{bmatrix} \to \left[\begin{array}{c|cc} 2 & 0 & 3 \\ \hline \frac{-4}{2} & 5 & 4 \\ \frac{6}{2} & -5 & -5 \end{array}\right] \to$$

Példa

2. lépés:

Ugyanúgy dolgozunk tovább, de most már csak a jobb alsó 2×2-es mátrix részen, a többit változatlanul leírjuk.

$$
\left[\begin{array}{r|rr}
\multicolumn{1}{r}{2} & 0 & 3 \\
\hline
-2 & 5 & 4 \\
3 & -5 & -5
\end{array}\right]
\to
\left[\begin{array}{r|r|r}
\multicolumn{1}{r}{2} & \multicolumn{1}{r}{0} & 3 \\
\hline
-2 & \multicolumn{1}{r}{5} & 4 \\
\cline{2-3}
3 & \frac{-5}{5} & -1
\end{array}\right]
$$

Olvassuk ki a keresett mátrixokat!

$$
A = \begin{bmatrix} 2 & 0 & 3 \\ -4 & 5 & -2 \\ 6 & -5 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 & 3 \\ 0 & 5 & 4 \\ 0 & 0 & -1 \end{bmatrix} = L \cdot U
$$

Tétel: LU-felbontás létezése

Ha a Gauss-elimináció végrehajtható sor és oszlopcsere nélkül (azaz $a_{kk}^{(k-1)} \neq 0$ $(k = 1, \ldots, n-1)$), akkor az A mátrix LU-felbontása létezik.

Biz.: Ha a GE végrehajtható sor és oszlopcsere nélkül, akkor az L_k mátrixok felírhatók és L, U előállítható. $\square$

Megj.:

- $u_{kk} = a_{kk}^{(k-1)}$ és $D_k = a_{11} \cdot a_{22}^{(1)} \cdots a_{kk}^{(k-1)}$
- Ha van A-nak LU-felbontása, ahol U átlójában nem nullák állnak, akkor $u_{kk} = a_{kk}^{(k-1)} \neq 0$.
- $a_{nn}^{(n-1)} \neq 0 \Leftrightarrow \det(A) = D_n \neq 0$.
- Ha a GE végrehajtható, de $a_{nn}^{(n-1)} = 0$, akkor létezik LU-felbontás, de $\det(A) = \det(L) \cdot \det(U) = \det(U) = 0$-ból $u_{nn} = 0$. Ebben az esetben a LER vagy nem oldható meg vagy nem egyértelműen.

Tétel: LU-felbontás létezése és egyértelműsége (főminorokkal)

- Ha $D_k \neq 0 \ (k = 1, \ldots, n-1)$, akkor létezik az A mátrix LU-felbontása és $u_{kk} \neq 0 \ (k = 1, \ldots, n-1)$.
- Ha $\det(A) \neq 0$, akkor a felbontás egyértelmű.

Biz.: létezés: az LU-felbontás létezése a GE-nál tanult tételünkből következik. $D_k \neq 0 \Leftrightarrow a_{kk}^{(k-1)} \neq 0$ a megadott indexekre, ezért a GE végrehajtható és az L, U mátrixok előállíthatóak.

Egyértelműség: indirekt tegyük fel, hogy az A invertálható mátrix LU-felbontása nem egyértelmű, azaz legalább két különböző felbontás létezik:

$$A = L_1 \cdot U_1 = L_2 \cdot U_2.$$

$$A = L_1 \cdot U_1 = L_2 \cdot U_2.$$

Az egyenlőséget U_2^{-1}-zel jobbról, majd L_1^{-1}-zel balról szorozva kapjuk, hogy

$$U_1 \cdot U_2^{-1} = L_1^{-1} \cdot L_2.$$

A szóban forgó inverzek léteznek, hiszen
$\det(A) = \det(L_i) \cdot \det(U_i) = \det(U_i) \neq 0$, $i = 1, 2$-re.
Az egyenlőség bal oldalán egy felső háromszögmátrix, jobb oldalán pedig egy 1 főátlójú alsó háromszögmátrix áll. Ez csak úgy lehet, ha az egységmátrixról van szó. Tehát

$$\begin{aligned} U_1 \cdot U_2^{-1} = I \quad &\Longrightarrow \quad U_1 = U_2, \\ L_1^{-1} \cdot L_2 = I \quad &\Longrightarrow \quad L_1 = L_2. \end{aligned}$$

Ellentmondásra jutottunk, vagyis az LU-felbontás egyértelmű. $\square$

L és U megadása GE-val

Az eddigieket összefoglalva felírhatjuk az $A = LU$ felbontást:

$$L \in \mathcal{L}_1 \text{ és } l_{ij} = \frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}} \ (i > j), \qquad U \in \mathcal{U} \quad \text{és} \quad u_{ij} = a_{ij}^{(i-1)} \ (i \leq j).$$

Miért jó az LU-felbontás?

Tegyük fel, hogy

- az $Ax = b$ LER megoldható, és
- rendelkezésünkre áll az $A = LU$ felbontás.

Ekkor $Ax = L \cdot \underbrace{U \cdot x}_{y} = b$ helyett $\quad (\frac{2}{3}n^3 + \mathcal{O}(n^2))$

1. oldjuk meg az $Ly = b$ alsó háromszögű, $\quad (n^2 + \mathcal{O}(n))$
2. majd az $Ux = y$ felső háromszögű LER-t. $\quad (n^2 + \mathcal{O}(n))$

Összehasonlításul: egy mátrix-vektor szorzás műveletigénye:
$n \cdot (2n - 1) = 2n^2 + \mathcal{O}(n)$.

Persze valamikor elő kell állítani az LU-felbontást. $\quad (\frac{2}{3}n^3 + \mathcal{O}(n^2))$
Előnyös, ha sokszor ugyanaz A: az ILU-algoritmusnál illetve az inverz iterációnál látjuk majd alkalmazását.

Az LU-felbontás „közvetlen” kiszámítása

- Nem ismerjük L-t és U-t: ismeretlenek a mátrixokban.
- Viszont szorzatukat ismerjük: $LU = A$.
- A egyes elemeit a mátrixszorzás alapján felírva egyenleteket kapunk L és U elemeire.
- *Jó sorrendben* felírva az egyenleteket, mindig megkapjuk egy-egy új ismeretlen értékét.
- A GE-nál láttuk, hogy U 1. sora azonos A 1. sorával (a GE az 1.sort nem változtatja).
- L 1. oszlopát úgy kapjuk, hogy A 1. oszlopát leosztjuk a_{11}-gyel.

$$\begin{pmatrix} 1. & 1. & 1. & 1. \\ 2. & 3. & 3. & 3. \\ 4. & 4. & 5. & 5. \\ 6. & 6. & 6. & 7. \end{pmatrix} \qquad \begin{pmatrix} 1. & 3. & 5. & 7. \\ 2. & 3. & 5. & 7. \\ 2. & 4. & 5. & 7. \\ 2. & 4. & 6. & 7. \end{pmatrix} \qquad \begin{pmatrix} 1. & 1. & 1. & 1. \\ 2. & 3. & 3. & 3. \\ 2. & 4. & 5. & 5. \\ 2. & 4. & 6. & 7. \end{pmatrix}$$

sorfolytonosan — oszlopfolytonosan — parkettaszerűen

Példa: *LU*-felbontás közvetlenül

- (a) Készítsük el a példamátrixunk LU-felbontását közvetlenül a mátrixszorzás alapján.
- (b) Nézzünk egy újabb példát is. (Vigyázat, $\det(B_2) = 0$.)

$$A = \begin{bmatrix} 2 & 0 & 3 \\ -4 & 5 & -2 \\ 6 & -5 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 2 & -2 & 3 \\ -4 & 4 & -2 \\ 6 & -5 & 4 \end{bmatrix}$$

Sorfolytonosan: U 1. sorát ismerjük. A 2. sor számítása:

$$\begin{bmatrix} 2 & 0 & 3 \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \qquad \begin{aligned} l_{21} \cdot 2 &= -4 \\ l_{21} \cdot 0 + 1 \cdot u_{22} &= 5 \\ l_{21} \cdot 3 + 1 \cdot u_{23} &= -2 \end{aligned}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \quad \begin{bmatrix} 2 & 0 & 3 \\ -4 & 5 & -2 \\ 6 & -5 & 4 \end{bmatrix} \qquad \begin{aligned} l_{21} &= -2 \\ u_{22} &= 5 \\ u_{23} &= -2 - (-2) \cdot 3 = 4 \end{aligned}$$

A 3. sor számítása:

$$\begin{aligned} l_{31} \cdot 2 &= 6 \\ l_{31} \cdot 0 + l_{32} \cdot u_{22} &= -5 \\ l_{31} \cdot 3 + l_{32} \cdot u_{23} + 1 \cdot u_{33} &= 4 \end{aligned} \qquad \begin{aligned} l_{31} &= 3 \\ l_{32} &= \frac{-5}{5} = -1 \\ u_{33} &= 4 - 3 \cdot 3 - (-1) \cdot 4 = -1 \end{aligned}$$

Sorfolytonosan: U 1. sorát ismerjük. A 2. sor számítása:

$$\begin{bmatrix} 2 & -2 & 3 \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \qquad \begin{aligned} l_{21} \cdot 2 &= -4 \\ l_{21} \cdot (-2) + 1 \cdot u_{22} &= 4 \\ l_{21} \cdot 3 + 1 \cdot u_{23} &= -2 \end{aligned}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \quad \begin{bmatrix} 2 & -2 & 3 \\ -4 & 4 & -2 \\ 6 & -5 & 4 \end{bmatrix} \quad \begin{aligned} l_{21} &= -2 \\ u_{22} &= 4 - (-2) \cdot (-2) = 0 \\ u_{23} &= -2 - (-2) \cdot 3 = 4 \end{aligned}$$

A 3. sor számítása:

$$l_{31} \cdot 2 = 6 \qquad\qquad l_{31} = 3$$

$$l_{31} \cdot (-2) + l_{32} \cdot u_{22} = -5 \qquad \rightsquigarrow \text{ellentmondásos egyenlet}$$

Mivel $D_2 = \det(B_2) = 0$, így $u_{22} = 0$ lesz. Az LU-felbontás nem készíthető el. GE-t alkalmazva $a_{22}^{(1)} = 0$ lenne, emiatt sort kéne cserélni.

Tétel: az LU-felbontás „közvetlen” kiszámítása

Az L és U mátrixok elemei a következő képletekkel számolhatók:

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

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

Ha jó sorrendben számolunk, mindig ismert az egész jobb oldal.

Az LU-felbontás „közvetlen” kiszámítása

Biz.: Írjuk fel az $A \in \mathbb{R}^{n \times n}$ mátrix, mint mátrixszorzat i-edik sorának j-edik elemét feltéve, hogy $A = L \cdot U$. Használjuk ki, hogy háromszögmátrixokról van szó, majd válasszunk le egy tagot.

Ha $i \leq j$, azaz egy főátló feletti (vagy főátlóbeli) elemről van szó, akkor $k > i \Rightarrow l_{ik} = 0$, valamint $l_{ii} = 1$, és így

$$a_{ij} = \sum_{k=1}^{n} l_{ik} \cdot u_{kj} = \sum_{k=1}^{i} l_{ik} \cdot u_{kj} = u_{ij} + \sum_{k=1}^{i-1} l_{ik} \cdot u_{kj}.$$

Ebből u_{ij} kifejezhető

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} \cdot u_{kj}.$$

Biz. folyt. Ha $i > j$, azaz egy főátló alatti elemről van szó, akkor $k > j \Rightarrow u_{kj} = 0$, és így

$$a_{ij} = \sum_{k=1}^{n} l_{ik} \cdot u_{kj} = \sum_{k=1}^{j} l_{ik} \cdot u_{kj} = l_{ij} \cdot u_{jj} + \sum_{k=1}^{j-1} l_{ik} \cdot u_{kj}.$$

Ha $u_{jj} \neq 0$ (találkoztunk már ezzel a feltétellel), akkor l_{ij} kifejezhető

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

Figyeljük meg, hogy ha valamely „jó sorrendben” (lásd az előadás diasorát) megyünk végig az (i, j) indexekkel A elemein, akkor az l_{ij} illetve u_{ij} értékét megadó egyenlőségek jobb oldalán minden mennyiség ismert. $\square$

Tartalomjegyzék

Tétel: Az LU-felbontás műveletigénye

$$\frac{2}{3}n^3 + \mathcal{O}(n^2)$$

Biz.: A GE-ból trivi, mert vele az LU-felbontás is előállítható.
A képletekből: Rögzített j-re:

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} \cdot u_{kj},$$

u_{ij}-hez $(i-1)$ szorzás és $(i-1)$ összeadás kell. Összesen $2(i-1)$ művelet. Rögzített i-re:

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

l_{ij}-hez 1 osztás, $(j-1)$ szorzás és $(j-1)$ összeadás kell. Összesen $2j-1$ művelet.

$$\sum_{j=1}^{n}\sum_{i=1}^{j}2(i-1)+\sum_{i=2}^{n}\sum_{j=1}^{i-1}(2j-1)=$$

$$\sum_{j=1}^{n}2\cdot\frac{(j-1)j}{2}+\sum_{i=2}^{n}\left(2\cdot\frac{(i-1)i}{2}-(i-1)\right)=$$

$$\sum_{j=1}^{n}j^2-\sum_{j=1}^{n}j+\sum_{i=2}^{n}(i-1)^2=\sum_{j=1}^{n}j^2-\sum_{j=1}^{n}j+\sum_{i=1}^{n-1}s^2$$

$$=\frac{n(n+1)(2n+1)}{6}-\frac{n(n+1)}{2}+\frac{(n-1)n(2n-1)}{6}=\frac{2}{3}n^3+\mathcal{O}(n^2). \quad \square$$

A háromszögmátrixú LER megoldás műveletigénye

Tétel: Az $Ux = y$ megoldásának műveletigénye

$$n^2 + \mathcal{O}(n)$$

Biz.: lásd GE visszahelyettesítés.

Tétel: Az $Ly = b$ megoldásának műveletigénye

$$n^2 + \mathcal{O}(n)$$

Biz.: Rögzített i. sorra $(i-1)$ szorzás és $(i-1)$ összeadás.
Összesen: $2(i-1)$ művelet.

$$\sum_{i=2}^{n} 2(i-1) = \sum_{s=1}^{n-1} 2s = 2 \cdot \frac{n(n-1)}{2} = n^2 + \mathcal{O}(n). \quad \square$$

Példák Matlab-ban

1. Az LU-felbontás működése „kisebb” ($n \approx 7$) mátrixokra,
2. valamint „nagyobb” mátrixokra ($n \approx 50$) színkóddal.
3. LER megoldása LU-felbontás segítségével.
4. Sok LER ($m \approx 10, 100$) megoldása futási idejének összevetése nagyobb mátrixok ($n \approx 50, 100, 200$) esetén: GE-val valamint az LU-felbontás kihasználásával.

Numerikus módszerek 1.

4. előadás: Megmaradási tételek, progonka módszer, LDU-felbontás, Cholesky-felbontás

Krebsz Anna

ELTE IK

Definíció: szimmetrikus mátrixok

Az A mátrix szimmetrikus, ha $A = A^\top$.

Definíció: pozitív definit mátrixok

Az $A \in \mathbb{R}^{n\times n}$ szimmetrikus mátrix *pozitív definit*, ha

1. $\langle Ax, x\rangle = x^\top Ax > 0$ bármely $0 \neq x \in \mathbb{R}^n$ esetén; vagy
2. minden főminorára $D_k = \det(A_k) > 0$; vagy
3. minden sajátértéke pozitív.

Állítás: pozitív definit mátrixok ekvivalens jellemzése

Az előző **1. 2. 3.** feltételek ekvivalensek.

Biz.: nélkül. $\square$

Definíció:

Az A mátrix **szigorúan diagonálisan domináns a soraira**, ha $|a_{ii}| > \sum_{j=1, j\neq i} |a_{ij}| \quad (i = 1, \ldots, n)$.

Definíció:

Az A mátrix **szigorúan diagonálisan domináns az oszlopaira**, ha $|a_{ii}| > \sum_{j=1, j\neq i} |a_{ji}| \quad (i = 1, \ldots, n)$.

Példa:

A következő mátrix szigorúan diagonálisan domináns a soraira és oszlopaira is.

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

Definíció:

Az A mátrix **fél sávszélessége** $s \in \mathbb{N}$, ha

$$\forall\ i,j : |i-j| > s : \ \ a_{ij} = 0 \text{ és}$$
$$\exists\ k,l : |k-l| = s : a_{kl} \neq 0.$$

Példa:

A következő mátrix szimmetrikus, pozitív definit és fél sávszélessége 1.

$$\begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix}$$

Definíció:

Az A mátrix **profilja** sorokra a $(k_1, \ldots, k_n)$, oszlopokra az $(l_1, \ldots, l_n)$ szám n-sek, melyekre

$$\forall\ j = 1, \ldots, k_i: \quad a_{ij} = 0 \text{ és } a_{i,k_i+1} \neq 0,$$
$$\forall\ i = 1, \ldots, l_j: \quad a_{ij} = 0 \text{ és } a_{l_j+1,j} \neq 0.$$

Soronként és oszloponként az első nem nulla elemig a nullák száma.

Példa:

A mátrix profilja sorokra $(0, 0, 2, 1)$, oszlopokra $(0, 1, 1, 2)$.

$$\begin{bmatrix} 4 & 0 & 0 & 0 \\ 2 & 4 & 1 & 0 \\ 0 & 0 & 4 & 3 \\ 0 & 1 & 2 & 4 \end{bmatrix}$$

Készítsük el az $Ax = b$ LER k. sor utáni particionálását ($k < n$, $k \in \mathbb{N}$) és tegyük fel, hogy $A_{11} \in \mathbb{R}^{k \times k}$ invertálható.

$$\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \cdot \left[\begin{array}{c} x_1 \\ \hline x_2 \end{array}\right] = \left[\begin{array}{c} b_1 \\ \hline b_2 \end{array}\right]$$

Particionált alakban a LER:

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 &= b_1 \\ A_{21}x_1 + A_{22}x_2 &= b_2 \end{aligned}$$

Végezzünk el egy blokkos GE-s lépést:
2. egyenlet $-$ $(A_{21} \cdot A_{11}^{-1})$ 1. egyenlet

$$\underbrace{(A_{21} - A_{21}A_{11}^{-1}A_{11})}_{0}x_1 + (A_{22} - A_{21}A_{11}^{-1}A_{12})\,x_2 = b_2 - A_{21}A_{11}^{-1}b_1$$

A GE blokkos lépése után a 2. sor alakja:

$$\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) x_2 = b_2 - A_{21}A_{11}^{-1}b_1.$$

Particionálva a LER:

$$\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{array}\right] \cdot \left[\begin{array}{c} x_1 \\ \hline x_2 \end{array}\right] = \left[\begin{array}{c} b_1 \\ \hline b_2 - A_{21}A_{11}^{-1}b_1 \end{array}\right]$$

- Most már csak az $(n-k) \times (n-k)$-s jobb alsó mátrix részen kell folytatnunk a GE-t.
- $k = 1$ esetén $A_{11} = (a_{11})$. Feltéve, hogy $a_{11} \neq 0$, akkor a fenti lépés a (blokk nélküli) 1. GE-s lépést írja le.

Definíció: Schur-komplementer

Tegyük fel, hogy $A_{11} \in \mathbb{R}^{k \times k}$ invertálható mátrix. Az A mátrix A_{11}**-re vonatkozó Schur-komplementere** az

$$[A|A_{11}] := A_{22} - A_{21}A_{11}^{-1}A_{12}$$

$(n-k) \times (n-k)$-s mátrix.

A Schur komplementer azt mutatja, hogy az A_{11}-gyel végzett GE után mely mátrixon kell folytatni az eliminációt. Az új fogalom segítségével könnyebben fogalmazhatjuk meg, hogy a GE mely tulajdonságokat örökíti tovább.

Tétel: megmaradási tételek a GE-ra

A GE során a következő tulajdonságok öröklődnek A-ról a Schur-komplementerre:

1. $\det(A) \neq 0 \quad \Rightarrow \quad \det([A|A_{11}]) \neq 0$
2. A szimmetrikus $\Rightarrow [A|A_{11}]$ szimmetrikus
3. A pozitív definit $\Rightarrow [A|A_{11}]$ pozitív definit
4. A szig. diag. dom. $\Rightarrow [A|A_{11}]$ szig. diag. dom.
5. $[A|A_{11}]$ fél sávszélessége $\leq A$ fél sávszélessége
6. A GE során a profilnál a soronkénti és oszloponkénti nullák az első nem nulla elemig megmaradnak.

Gondoljuk végig az LU-felbontás L, U mátrixára a megfelelő tulajdonságokat.

Biz.: 1.) Determináns:

Mivel a GE determináns tartó, így $\det(A) = \det(A^{(1)}) \neq 0$.

$$A^{(1)} = \left[\begin{array}{r|c} A_{11} & A_{12} \\ \hline 0 & [A|A_{11}] \end{array}\right]$$

$$0 \neq \det(A^{(1)}) = \underbrace{\det(A_{11})}_{\neq 0} \cdot \det([A|A_{11}]) \quad \Leftrightarrow \quad \det([A|A_{11}]) \neq 0$$

$\square$

2.) Szimmetria:

Ha A szimmetrikus, akkor A_{11} és A_{22} is az, továbbá $A_{21}^\top = A_{12}$.

$$\begin{aligned}[A|A_{11}]^\top &= (A_{22} - A_{21}A_{11}^{-1}A_{12})^\top = A_{22}^\top - A_{12}^\top (A_{11}^{-1})^\top A_{21}^\top = \\ &= A_{22}^\top - A_{12}^\top (A_{11}^\top)^{-1} A_{21}^\top = A_{22} - A_{21}A_{11}^{-1}A_{12} = [A|A_{11}]\end{aligned}$$

$\square$

Biz.: 3.) Pozitív definitség:
Tudjuk, hogy $\langle Ax, x\rangle > 0$ minden $x \neq 0$ vektorra.
Be kell látnunk, hogy $\langle [A|A_{11}]x_2, x_2\rangle > 0$ minden $x_2 \neq 0$ vektorra.
Vegyük észre, hogy $x \in \mathbb{R}^n$ és $x_2 \in \mathbb{R}^{n-k}$.

$$Ax = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \cdot \left[\begin{array}{c} x_1 \\ \hline x_2 \end{array}\right] = \left[\begin{array}{c} A_{11}x_1 + A_{12}x_2 \\ \hline A_{21}x_1 + A_{22}x_2 \end{array}\right]$$

Legyen $x_2 \in \mathbb{R}^{n-k}$ tetszőleges, válasszuk meg $x_1 \in \mathbb{R}^k$ vektort úgy, hogy Ax első k komponense 0 legyen:

$$A_{11}x_1 + A_{12}x_2 = 0 \quad \Rightarrow \quad x_1 := -A_{11}^{-1}A_{12}x_2.$$

$$x_1 := -A_{11}^{-1} A_{12} x_2$$

Helyettesítsük be a skaláris szorzatba:

$$\begin{aligned} 0 < \langle Ax, x\rangle &= \langle \underbrace{A_{11}x_1 + A_{12}x_2}_{0}, x_1\rangle + \langle A_{21}x_1 + A_{22}x_2, x_2\rangle = \\ &= \langle A_{21}(-A_{11}^{-1}A_{12}x_2) + A_{22}x_2, x_2\rangle = \\ &= \langle (-A_{21}A_{11}^{-1}A_{12} + A_{22})x_2, x_2\rangle = \\ &= \langle (A_{22} - A_{21}A_{11}^{-1}A_{12})x_2, x_2\rangle = \langle [A|A_{11}]x_2, x_2\rangle \end{aligned}$$

$\square$

Biz.: 4.) Szigorúan diagonálisan domináns a soraira $k = 1$ esetén:
A GE az első sort nem változtatja, ezen a szig. diag. dom. megmarad. Be kellene látnunk, hogy $i = 2, \ldots, n$-re

$$\left|a_{ii}^{(1)}\right| > \sum_{j=2, j\neq i}^{n} \left|a_{ij}^{(1)}\right|.$$

A GE képleteit behelyettesítve

$$\left|a_{ii} - \frac{a_{i1}}{a_{11}} a_{1i}\right| > \sum_{j=2, j\neq i}^{n} \left|a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j}\right|.$$

Szorozzuk be mindkét oldalt $|a_{11}| \neq 0$-val

$$|a_{ii} a_{11} - a_{i1} a_{1i}| > \sum_{j=2, j\neq i}^{n} |a_{ij} a_{11} - a_{i1} a_{1j}| \quad (i = 2, \ldots, n).$$

A kapott egyenlőtlenség bal oldalát lefelé, jobb oldalát felfelé becsüljük

$$|a_{ii}a_{11}| - |a_{i1}a_{1i}| > \sum_{j=2, j\neq i}^{n} \left(|a_{ij}a_{11}| + |a_{i1}a_{1j}|\right) \quad (i = 2, \dots, n).$$

A továbbiakban ezt fogjuk belátni. Az 1. sort a GE helyben hagyja, ezért itt továbbra is igaz, hogy $|a_{11}| > \sum_{j=2}^{n} |a_{1j}|$
Szorozzuk $|a_{i1}| \neq 0$-val és vegyük külön az i. tagot:

$$|a_{11}a_{i1}| > |a_{1i}a_{i1}| + \sum_{j=2, j\neq i}^{n} |a_{1j}a_{i1}|.$$

Írjuk fel a szigorúan diagonálisan dominanciát az $i = 2, \dots, n$-re
$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| = |a_{i1}| + \sum_{j=2, j\neq i}^{n} |a_{ij}|$.
Szorozzuk $|a_{11}|$-gyel mindkét oldalt:

$$|a_{ii}a_{11}| > |a_{i1}a_{11}| + \sum_{j=2, j\neq i}^{n} |a_{ij}a_{11}|.$$

Becsüljük $|a_{ii}a_{11}|$-t alulról

$$|a_{ii}a_{11}| > |a_{1i}a_{i1}| + \sum_{j=2, j\neq i}^{n} \left(|a_{1j}a_{i1}| + |a_{ij}a_{11}|\right).$$

Átrendezve a bizonyítandó állítást kapjuk

$$|a_{ii}a_{11}| - |a_{1i}a_{i1}| > \sum_{j=2, j\neq i}^{n} \left(|a_{1j}a_{i1}| + |a_{ij}a_{11}|\right).$$

Nézzük meg, hogy korábban mivel szoroztunk:

- Ha $a_{i1} = 0$, akkor ezen a soron nem változtat a GE, tehát a diag. dominancia nem változik.
- $a_{11} \neq 0$, mivel ez feltétele a GE-nak.

Az oszlopokra vonatkozó bizonyítás analóg módon elvégezhető. $\square$

Rövidített GE (progonka módszer)

A gyakorlatban megszokott, hogy tridiagonális (háromátlós) LER-t kell megoldanunk. Az év eleji példában is láttuk, de köbös spline-ok meghatározása esetén is ilyen alakú LER-t kapunk. A speciális alakot felhasználva hatékonyabb alakot algoritmust készítünk.

- Tárolás: n^2 helyett $3n-2$ elem.
- Műveletigény: $\frac{2}{3}n^3 + \mathcal{O}(n^2)$ helyett $8n + \mathcal{O}(1)$.

Mivel a GE a sávszélességet megtartja, tridiagonális esetben a három átlón kívül mindig nulla lesz. A GE végén kapott U mátrix is csak két átlót tartalmaz, ezért a visszahelyettesítés i. egyenlete

$$a_{ii}^{(i-1)} x_i + a_{ii+1}^{(i-1)} x_{i+1} = a_{in+1}^{(i-1)}.$$

Ebből x_i-t kifejezve, új jelölésrendszerrel $x_i = f_i x_{i+1} + g_i$ $(i = 1, \ldots, n)$ alakú.

Rövidített GE (progonka módszer)

Jelölések: $A = \text{tridiag}\,(\beta_{i-1}, \alpha_i, \gamma_i)$,

$$A = \begin{bmatrix} \alpha_1 & \gamma_1 & 0 & \cdots & 0 \\ \beta_1 & \alpha_2 & \gamma_2 & & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \beta_{n-2} & \alpha_{n-1} & \gamma_{n-1} \\ 0 & & 0 & \beta_{n-1} & \alpha_n \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}.$$

A LER 1. egyenlete:

$$\alpha_1 x_1 + \gamma_1 x_2 = b_1 \quad \to \quad \alpha_1 x_1 = -\gamma_1 x_2 + b_1 \quad \to \quad x_1 = -\frac{\gamma_1}{\alpha_1} x_2 + \frac{b_1}{\alpha_1}$$

Az $x_1 = f_1 x_2 + g_1$ alakot keresve $f_1 = -\frac{\gamma_1}{\alpha_1}$ és $g_1 = \frac{b_1}{\alpha_1}$.

Rövidített GE (progonka módszer)

Tegyük fel, hogy $f_1, \ldots, f_{i-1}$ és $g_1, \ldots, g_{i-1}$, továbbá az $x_k = f_k x_{k+1} + g_k$ $(k = 1, \ldots, i-1)$ rekurzió ismert. Az $x_i = f_i x_{i+1} + g_i$ rekurzió képleteit szeretnénk meghatározni. Írjuk fel az i. egyenletet és helyettesítsük be x_{i-1} helyére a rekurziót:

$$\begin{aligned}
\beta_{i-1}x_{i-1} + \alpha_i x_i + \gamma_i x_{i+1} &= b_i \\
\beta_{i-1}(f_{i-1}x_i + g_{i-1}) + \alpha_i x_i + \gamma_i x_{i+1} &= b_i \\
(\beta_{i-1}f_{i-1} + \alpha_i)x_i + \gamma_i x_{i+1} &= b_i - \beta_{i-1}g_{i-1}
\end{aligned}$$

$$\begin{aligned}
(\alpha_i + \beta_{i-1}f_{i-1})x_i &= -\gamma_i x_{i+1} + (b_i - \beta_{i-1}g_{i-1}) \\
x_i &= -\frac{\gamma_i}{\alpha_i + \beta_{i-1}f_{i-1}}x_{i+1} + \frac{b_i - \beta_{i-1}g_{i-1}}{\alpha_i + \beta_{i-1}f_{i-1}}.
\end{aligned}$$

Innen $f_i = -\frac{\gamma_i}{\alpha_i+\beta_{i-1}f_{i-1}}$ és $g_i = \frac{b_i-\beta_{i-1}g_{i-1}}{\alpha_i+\beta_{i-1}f_{i-1}}$.

Írjuk fel az n. egyenletet és helyettesítsük be x_{n-1} helyére a rekurziót:

$$\begin{aligned} \beta_{n-1}x_{n-1} + \alpha_n x_n &= b_n \\ \beta_{n-1}(f_{n-1}x_n + g_{n-1}) + \alpha_n x_n &= b_n \\ (\beta_{n-1}f_{n-1} + \alpha_n)x_n &= b_n - \beta_{n-1}g_{n-1} \end{aligned}$$

$$x_n = \frac{b_n - \beta_{n-1}g_{n-1}}{\alpha_n + \beta_{n-1}f_{n-1}} =: g_n$$

$\square$

Algoritmus: progonka módszer

1. lépés: $f_1 := -\frac{\gamma_1}{\alpha_1}, \quad g_1 := \frac{b_1}{\alpha_1}$

$$i = 2, \ldots, n-1: \quad f_i := -\frac{\gamma_i}{\alpha_i + \beta_{i-1} f_{i-1}}$$

$$g_i := \frac{b_i - \beta_{i-1} g_{i-1}}{\alpha_i + \beta_{i-1} f_{i-1}}$$

$$g_n := \frac{b_i - \beta_{n-1} g_{n-1}}{\alpha_n + \beta_{n-1} f_{n-1}}$$

2. lépés: $x_n := g_n$

$i = n-1, n-2, \ldots, 1: \quad x_i = f_i x_{i+1} + g_i$

Megj.: 3 művelettel több, de könnyebben megjegyezhető az algoritmus, ha f_n értékét is meghatározzuk. Ekkor $x_{n+1} := 0$-val indítjuk a 2. lépést.

Műveletigény:

1. lépés (előre):
f_1, g_1 : 2 művelet.

A ciklus i. lépésében: a közös nevezőben 2 db, f_i-ben 1 db, g_i-ben 3 db, tehát $i = 2, \ldots, n-1$-re összesen $6(n-2)$ db.

g_n-ben 5 db művelet.

2. lépés (vissza):
$i = n-1, n-2, \ldots, 1$-re $2(n-1)$ db művelet.

Összesen:
$2 + 6(n-2) + 5 + 2(n-1) = 8n - 7 = 8n + \mathcal{O}(1)$ művelet. $\square$

Definíció: LDU-felbontás

Az $A \in \mathbb{R}^{n \times n}$ mátrix LDU-felbontásának nevezzük az $A = L \cdot D \cdot U$ szorzatot, ha $L \in \mathcal{L}_1$ alsó háromszögmátrix, D diagonális mátrix és $U \in \mathcal{U}_1$ felső háromszögmátrix.

Előállítás LU-felbontásból:

Az $A = L \cdot \widetilde{U}$ felbontásban $L \in \mathcal{L}_1$ jó, $D = \operatorname{diag}(\widetilde{u}_{11}, \ldots, \widetilde{u}_{nn})$.
A keresett $U \in \mathcal{U}_1$ mátrixot úgy kapjuk, hogy $U = D^{-1}\widetilde{U}$, azaz minden i-re $\widetilde{U}$ i. sorát $\widetilde{u}_{ii}$-vel osztjuk. Ekkor

$$A = L\widetilde{U} = LD \cdot \underbrace{(D^{-1}\widetilde{U})}_{U} = LDU.$$

Példa: *LDU*-felbontás *LU*-felbontásból

Készítsük el példamátrixunk LDU-felbontását az LU-felbontás segítségével.

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

Korábban láttuk, hogy

$$A = \begin{bmatrix} 2 & 0 & 3 \\ -4 & 5 & -2 \\ 6 & -5 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 & 3 \\ 0 & 5 & 4 \\ 0 & 0 & -1 \end{bmatrix} = L \cdot \widetilde{U}.$$

Legyen $D := \operatorname{diag}(2, 5, -1)$, $U := D^{-1}\widetilde{U}$. Tehát $A = LDU$, ahol

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

Balról D^{-1}-zel úgy szorzunk, hogy D megfelelő átlóbeli elemeivel osztjuk a megfelelő sorokat. $\square$

Tétel: az LDU-felbontás „közvetlen” kiszámítása

Az L, D és U mátrixok elemeit jó sorrendben (lásd LU-felbontás) számolva a jobboldalon mindig ismert értékek lesznek:

$$i < j \text{ (felső)} \qquad u_{ij} = \frac{1}{d_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} l_{ik} \cdot d_{kk} \cdot u_{kj} \right),$$

$$i = j \text{ (diag)} \qquad d_{ii} = a_{ii} - \sum_{k=1}^{i-1} l_{ik} \cdot d_{kk} \cdot u_{ki},$$

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

A képleteket az $A = L\widetilde{U}$ felbontás „közvetlen” képleteiből kapjuk:

$$\widetilde{u}_{ii} \mapsto d_{ii}, \quad \widetilde{u}_{kj} \mapsto d_{kk} u_{kj}.$$

Tétel: Szimmetrikus mátrix LDU-felbontása

Ha A szimmetrikus mátrix, akkor az LDU-felbontásában $U = L^\top$.

Biz.: az $A = LDU$ felbontás bal oldalát szorozzuk L^{-1}-zel, jobb oldalát $(L^{-1})^\top$-tal:

$$L^{-1}A(L^{-1})^\top = L^{-1} \cdot (LDU) \cdot (L^{-1})^\top = DU(L^{-1})^\top.$$

A bal oldali mátrixról tudjuk, hogy szimmetrikus, a jobboldali felső háromszögmátrix. Ebből következik, hogy a jobboldali mátrix diagonális mátrix. $U(L^{-1})^\top \in \mathcal{U}_1$, így $U(L^{-1})^\top = I$.

$$U(L^{-1})^\top = I \quad \Leftrightarrow \quad U(L^\top)^{-1} = I \quad \Leftrightarrow \quad U = L^\top$$

$\square$

Következmény:

- Szimmetrikus mátrix esetén az LDU-felbontás megtartja a szimmetriát. A teljes mátrix helyett elég pl. az alsó háromszög részét tárolni. Az $A = LDU$ felbontás valójában $LDL^\top$-felbontás lesz, ahol szintén elég L, D-t tárolni. Ezzel a tárolás- és műveletigény kb. a felére csökken $(\frac{1}{3}n^3 + \mathcal{O}(n^2))$.
- Szimmetrikus mátrix esetén az $LDL^\top$-felbontás GE-val közvetlenül is elkészíthető.

Példa: LDU-felbontás LU-felbontásból

Készítsük el szimmetrikus példamátrixunk $LDL^\top$-felbontását a GE segítségével.

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 8 & 6 \\ 1 & 6 & 6 \end{bmatrix}$$

Példa

A GE-s hányadosokat minden lépésben az eliminált pozíciókon tudjuk tárolni: az eliminálandó mátrix rész 1. oszlopában az első elemmel leosztjuk az alatta levőket. Vonalakkal jelezzük, hogy itt már tárolásról is szó van. A jobb alsó 2×2-es mátrix részen elvégezzük az eliminációt.

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 8 & 6 \\ 1 & 6 & 6 \end{bmatrix} \to \left[\begin{array}{c|cc} 1 & & \\ \hline 2 & 4 & 4 \\ 1 & 4 & 5 \end{array}\right] \to$$

Ugyanúgy dolgozunk tovább, de most már csak a jobb alsó 2×2-es mátrix részen, a többit változatlanul leírjuk.

$$\left[\begin{array}{c|cc} 1 & & \\ \hline 2 & 4 & 4 \\ 1 & 4 & 5 \end{array}\right] \to \left[\begin{array}{c|c|c} 1 & \multicolumn{2}{c}{} \\ \hline 2 & \multicolumn{2}{c}{4} \\ \cline{2-3} 1 & 1 & 1 \end{array}\right]$$

Készen vagyunk, csak le kell olvasnunk a felbontást: $A = LDL^\top$, ahol

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad L^\top = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

$\square$

Az $LDL^\top$-felbontás „közvetlen" kiszámítása

Tétel: az $LDL^\top$-felbontás „közvetlen" kiszámítása

Az L és U mátrixok elemei a következő képletekkel számolhatók:

$$i = j \text{ (diag)} \qquad d_{ii} = a_{ii} - \sum_{k=1}^{i-1} l_{ik} \cdot d_{kk} \cdot l_{ik},$$

$$i > j \text{ (alsó)} \qquad l_{ij} = \frac{1}{d_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} \cdot d_{kk} \cdot l_{jk} \right).$$

Ha jó sorrendben számolunk, mindig ismert az egész jobb oldal.

Tartalomjegyzék

Definíció: Cholesky-felbontás, avagy $LL^\top$-felbontás

Az $A \in \mathbb{R}^{n \times n}$ szimmetrikus mátrix Cholesky-felbontásának nevezzük az $L \cdot L^\top$ szorzatot, ha $A = LL^\top$, ahol $L \in \mathbb{R}^{n \times n}$ alsó háromszögmátrix és $l_{ii} > 0 \ \ (i = 1, \ldots, n)$.

Tétel: Cholesky-felbontás $\exists!$

Ha A szimmetrikus és pozitív definit mátrix, akkor egyértelműen létezik Cholesky-felbontása.

Cholesky-felbontás egyértelműség bizonyítása

Biz.: Egyértelműség: Tegyük fel indirekt, hogy létezik legalább két különböző felbontás,

$$A = L_1 L_1^\top = L_2 L_2^\top,$$

ahol $L_1, L_2 \in \mathcal{L}$, melyek diagonális elemei pozitívak.
Legyen $D_1 = \operatorname{diag}((L_1)_{ii})$ és $D_2 = \operatorname{diag}((L_2)_{ii})$.

$$\underbrace{(L_1 D_1^{-1})}_{\in \mathcal{L}_1} \cdot \underbrace{(D_1 L_1^\top)}_{\in \mathcal{U}} = \underbrace{(L_2 D_2^{-1})}_{\in \mathcal{L}_1} \cdot \underbrace{(D_2 L_2^\top)}_{\in \mathcal{U}}$$

A két oldalon egy-egy LU-felbontást látunk. Mivel az LU-felbontás egyértelmű (a főminorok nem nullák): $D_1 L_1^\top = D_2 L_2^\top$.

A főátlókban lévő elemek egyeznek, ezért $(L_1)_{ii}^2 = (L_2)_{ii}^2 \ \forall \ i$-re.
A diagonális elemek pozitivitása miatt

$$\forall i : \ (L_1)_{ii} = (L_2)_{ii} \quad \Rightarrow \quad L_1 = L_2 \quad \Rightarrow \quad D_1 = D_2.$$

Ezzel ellentmondásra jutottunk.

Létezés: Mivel A szimmetrikus és pozitív definit, ezért $D_k = \det(A_k) > 0$ $(k = 1, \ldots, n)$. A főminorok pozitivitásából következik, hogy $\exists\,!\ A = \widetilde{L}\widetilde{U}$ LU-felbontás és $\widetilde{u}_{ii} > 0\ \forall\ i$-re.
Legyen $D = \operatorname{diag}\left(\sqrt{\widetilde{u}_{11}}, \ldots, \sqrt{\widetilde{u}_{nn}}\right)$, így

$$A = \underbrace{(\widetilde{L}D)}_{B} \cdot \underbrace{(D^{-1}\widetilde{U})}_{C} = B \cdot C.$$

$B, C \in \mathcal{L}$, átlójuk egyaránt a $\widetilde{u}_{ii}$ elemekből áll. Be kell még látnunk, hogy $C^\top = B$.
A szimmetria miatt $A = A^\top$, azaz $BC = C^T B^\top$.
Bal oldalról szorozzunk B^{-1}-zel, jobbról $(B^\top)^{-1}$-zel:

$$B^{-1}(BC)(B^\top)^{-1} = B^{-1}(C^T B^\top)(B^\top)^{-1}$$
$$\mathcal{U}_1 \in C(B^\top)^{-1} = B^{-1}C^\top \in \mathcal{L}_1$$
$$B^{-1}C^\top = I \quad \Leftrightarrow \quad C^\top = B$$

□

Miért jó az $LL^\top$-felbontás?

Tegyük fel, hogy

- az $Ax = b$ LER megoldható,
- A szimmetrikus és
- rendelkezésünkre áll az $A = LL^\top$ felbontás.

Ekkor $Ax = L \cdot \underbrace{L^\top \cdot x}_{y} = b$ helyett $\quad (\frac{1}{3}n^3 + \mathcal{O}(n^2))$

1. oldjuk meg az $Ly = b$ alsó háromszögű, $\quad (n^2 + \mathcal{O}(n))$
2. majd az $L^T x = y$ felső háromszögű LER-t. $\quad (n^2 + \mathcal{O}(n))$

Persze valamikor elő kell állítani az $LL^\top$-felbontást, de csak L-et kell tárolni hozzá. $\quad (\frac{1}{3}n^3 + \mathcal{O}(n^2))$
Előnyös, ha sokszor ugyanaz A.

1. előállítási módszer: LU-felbontásból LDU-n keresztül.

- Legyen az A mátrix LU-felbontása: $A = \widetilde{L}\widetilde{U}$.
- Ha A poz. def., akkor $\widetilde{U}$ főátlóbeli elemei mind pozitívak. (!)
(Látjuk, hogy elkészíthető-e a Cholesky-felbontás.)
- Legyen $D := \operatorname{diag}(\widetilde{u}_{1,1}, \ldots, \widetilde{u}_{n,n})$, valamint $U = D^{-1}\widetilde{U}$.
- Kiderül, hogy szimmetrikus A esetén $U = \widetilde{L}^\top$. $(A = \widetilde{L}D\widetilde{L}^\top)$
- $\sqrt{D} := \operatorname{diag}\left(\sqrt{\widetilde{u}_{1,1}}, \ldots, \sqrt{\widetilde{u}_{n,n}}\right)$ jelöléssel most
$A = \underbrace{\widetilde{L} \cdot \sqrt{D}}_{L} \cdot \underbrace{\sqrt{D} \cdot \widetilde{L}^\top}_{L^\top} = L \cdot L^\top$.

Megj.: Nem szükséges az $LDL^\top$-felbontást előállítani, $\widetilde{U}$ elemeit felhasználva egyből az utolsó pontra térhetünk.

2. előállítási módszer: „mechanikusan" a GE-n keresztül.

- Az a_{11} helyére $\sqrt{a_{11}}$-et írunk.
- Végigosztjuk az 1. oszlopot $\sqrt{a_{11}}$-gyel.
- Eliminálunk a maradék $(n-1) \times (n-1)$-es mátrixban.
- Megyünk tovább...
- A végén csak az alsó háromszögmátrixot olvassuk ki.

3. előállítási módszer: mátrixszorzás alapján.

Tétel: az $LL^\top$-felbontás „közvetlen" kiszámítása

Az L mátrix elemei az A alsóháromszögbeli elemeiből a következő képletekkel számolhatók:

$$i = j \text{ (átló)} \qquad l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2},$$

$$i > j \text{ (alsó)} \qquad l_{ij} = \frac{1}{l_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} \cdot l_{jk} \right).$$

Ha jó sorrendben számolunk, mindig ismert az egész jobb oldal.

Biz.: Az LU-felbontáshoz hasonlóan. Írjuk fel az $A \in \mathbb{R}^{n\times n}$ mátrix, mint mátrixszorzat i-edik sorának j-edik elemét feltéve, hogy $A = L \cdot L^\top$. Használjuk ki, hogy háromszögmátrixokról van szó, majd válasszunk le egy tagot.

Ha $i = j$, azaz egy főátlóbeli elemről van szó, akkor $k > j \Rightarrow l_{j,k} = 0$, valamint $(L^\top)_{kj} = l_{jk}$, és így

$$a_{jj} = \sum_{k=1}^{n} l_{jk} \cdot (L^\top)_{kj} = \sum_{k=1}^{j} l_{jk}^2 = l_{jj}^2 + \sum_{k=1}^{j-1} l_{jk}^2.$$

Ebből l_{jj} kifejezhető

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}.$$

Biz. folyt. Ha $i > j$, azaz egy főátló alatti elemről van szó, akkor

$$a_{ij} = \sum_{k=1}^{n} l_{ik} \cdot (L^\top)_{kj} = \sum_{k=1}^{j} l_{ik} \cdot l_{jk} = l_{ij} \cdot l_{jj} + \sum_{k=1}^{j-1} l_{ik} \cdot l_{jk}.$$

Ha $l_{jj} \neq 0$ (találkoztunk már ezzel a feltétellel), akkor l_{ij} kifejezhető

$$l_{ij} = \frac{1}{l_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} \cdot l_{jk} \right).$$

Figyeljük meg, hogy ha valamely „jó sorrendben” (lásd LU-felbontásnál a sorrendek) megyünk végig az (i, j) indexekkel A alsóháromszögbeli elemein, akkor az l_{ij} illetve l_{jj} értékét megadó egyenlőségek jobb oldalán minden mennyiség ismert. $\square$

Tétel: A Cholesky-felbontás előállításának műveletigénye

A szorzások és osztások száma

$$\frac{1}{3}n^3 + \mathcal{O}(n^2),$$

valamint n darab négyzetgyökvonás is szükséges.

Biz.: A képletekből:

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}, \quad l_{ij} = \frac{1}{l_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} \cdot l_{jk} \right).$$

Rögzített j-re: l_{jj}-hez $2(j-1)$ szorzás és összeadás kell.

Rögzített i, j-re: l_{ij}-hez 1 osztás, $(j-1)$ szorzás és $(j-1)$ összeadás kell. Összesen $2j-1$ művelet.

$$\sum_{j=1}^{n} 2(j-1) + \sum_{i=2}^{n} \sum_{j=1}^{i-1} (2j-1) =$$

$$\sum_{s=1}^{n-1} 2s + \sum_{i=2}^{n} \left(2 \cdot \frac{(i-1)i}{2} - (i-1) \right) =$$

$$\sum_{s=1}^{n-1} 2s + \sum_{i=2}^{n} (i-1)^2 = \sum_{s=1}^{n-1} 2s + \sum_{t=1}^{n-1} t^2$$

$$= 2 \cdot \frac{(n-1)n}{2} + \frac{(n-1)n(2n-1)}{6} = \frac{1}{3} n^3 + \mathcal{O}(n^2). \quad \square$$

Példa

Készítsük el a következő (szimmetrikus, pozitív definit) mátrix Cholesky-felbontását

(a) az LU-felbontás alapján,

(b) „mechanikusan".

$$A = \begin{pmatrix} 4 & 2 & 4 \\ 2 & 10 & 5 \\ 4 & 5 & 6 \end{pmatrix}.$$

Példa (a)

LU-**felbontásból:** A mátrixon elvégezzük a GE lépéseit:

1. lépés:

$$\begin{bmatrix} 4 & 2 & 4 \\ 2 & 10 & 5 \\ 4 & 5 & 6 \end{bmatrix} \to \left[\begin{array}{c|cc} 4 & 2 & 4 \\ \hline {\color{red}\frac{1}{2}} & 9 & 3 \\ {\color{red}1} & 3 & 2 \end{array}\right] \to$$

2. lépés:

$$\left[\begin{array}{c|cc} 4 & 2 & 4 \\ \hline {\color{red}\frac{1}{2}} & 9 & 3 \\ {\color{red}1} & 3 & 2 \end{array}\right] \to \left[\begin{array}{c|cc} 4 & 2 & 4 \\ \hline {\color{red}\frac{1}{2}} & 9 & 3 \\ \cline{2-3} {\color{red}1} & \multicolumn{1}{c|}{{\color{red}\frac{1}{3}}} & 1 \end{array}\right]$$

Készen vagyunk az eliminációval, csak le kell olvasnunk $\widetilde{L}$, $\widetilde{U}$-ot.

$$A = \widetilde{L} \cdot \widetilde{U} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 1 & \frac{1}{3} & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 & 2 & 4 \\ 0 & 9 & 3 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 4 \\ 2 & 10 & 5 \\ 4 & 5 & 6 \end{bmatrix}.$$

$D = \operatorname{diag}(\sqrt{4}, \sqrt{9}, \sqrt{1}) = \operatorname{diag}(2, 3, 1).$

$$L = \widetilde{L} \cdot D = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 1 & \frac{1}{3} & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 3 & 0 \\ 2 & 1 & 1 \end{bmatrix}$$

A diagonális mátrix-szal jobbról szorzás az $\widetilde{L}$ megfelelő oszlopait szorozza az átlóbeli elemekkel. $\square$

Példa (b)

„Mechanikusan” közvetlenül a GE-ból: Az a_{11} helyére $\sqrt{a_{11}}$-et írunk. Végigosztjuk az 1. oszlopot $\sqrt{a_{11}}$-gyel. A jobb alsó 2×2-es mátrix részen elvégezzük az 1. sor segítségével az eliminációt.

$$\begin{bmatrix} 4 & 2 & 4 \\ 2 & 10 & 5 \\ 4 & 5 & 6 \end{bmatrix} \to \left[\begin{array}{c|cc} 2 & & \\ \hline 1 & 9 & 3 \\ 2 & 3 & 2 \end{array}\right] \to$$

Ugyanúgy dolgozunk tovább, de most már csak a jobb alsó 2×2-es mátrix részen, a többit változatlanul leírjuk.

$$\left[\begin{array}{c|cc} 2 & & \\ \hline 1 & 9 & 3 \\ 2 & 3 & 2 \end{array}\right] \to \left[\begin{array}{c|c|c} 2 & & \\ \hline 1 & 3 & \\ \cline{2-3} 2 & 1 & 1 \end{array}\right] \to \left[\begin{array}{c|c|c} 2 & & \\ \hline 1 & 3 & \\ \cline{2-3} 2 & 1 & \sqrt{1} \end{array}\right]$$

Az utolsó átlóbeli elemből ne felejtsünk el gyököt vonni.

Készen vagyunk, ellenőrizhetjük a Cholesky-felbontást:

$$A = L \cdot L^\top = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 3 & 0 \\ 2 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 1 & 2 \\ 0 & 3 & 1 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 4 \\ 2 & 10 & 5 \\ 4 & 5 & 6 \end{bmatrix}.$$

$\square$

Példák Matlab-ban

1. Példák pozitív definit mátrixokra,

Numerikus módszerek 1.

5. előadás: QR-felbontás: Gram–Schmidt ortogonalizáció, Householder-transzformációk és alkalmazásaik

Krebsz Anna

ELTE IK

Tartalomjegyzék

Definíció: ortogonális mátrix

Egy $Q \in \mathbb{R}^{n\times n}$ mátrix *ortogonális*, ha az inverze a transzponáltja, azaz

$$Q^\top Q = I.$$

Megj.: Ekkor $QQ^\top = I$ is teljesül. ($Q^{-1} = Q^\top$)

Definíció: skaláris szorzat

Az $x, y \in \mathbb{R}^n$ vektorok *skaláris szorzata*

$$\langle x, y\rangle := y^\top x = \sum_{k=1}^{n} x_k \cdot y_k.$$

Definíció: ortonormált rendszer

A $q_1, \dots, q_n \in \mathbb{R}^n$ vektorok *ortonormált rendszert* alkotnak, ha

$$\langle q_i, q_j \rangle = \begin{cases} 0 & \text{ha } i \neq j, \\ 1 & \text{ha } i = j. \end{cases}$$

Állítás: ortogonális mátrixok oszlopvektorairól

A $Q \in \mathbb{R}^{n \times n}$ ortogonális mátrix oszlopai, mint vektorok ortonormált rendszert alkotnak.

Biz.: Gondoljunk bele: $Q^\top Q = I$. $\square$

Definíció: ortogonális rendszer

A $q_1, \ldots, q_n \in \mathbb{R}^n$ vektorok *ortogonális rendszert* alkotnak, ha

$$\langle q_i, q_j \rangle = 0 \qquad (i \neq j).$$

Állítás: ortogonális rendszerekből álló mátrixokról

Ha a $q_1, \ldots, q_n \in \mathbb{R}^n$ vektorok ortogonális rendszert alkotnak, akkor a $Q := (q_1, \ldots, q_n) \in \mathbb{R}^{n \times n}$ mátrix esetén a $Q^\top Q$ szorzatmátrix diagonális. ($QQ^\top$ általában nem.)

Biz.: Gondoljunk bele: $Q^\top Q = D$ diagonális mátrix. $\square$

Elnevezések:

- $\langle q_i, q_j \rangle = \delta_{ij}$ (Kronecker-féle delta).
- $q_i \perp q_j \Leftrightarrow \langle q_i, q_j \rangle = 0 \ \ (i \neq j)$: az oszlopok merőlegesek, avagy *ortogonálisak* egymásra
- $\langle q_i, q_i \rangle = 1$: minden oszlopvektor hossza 1, avagy *normált* $\|q_i\|_2 := \sqrt{\langle q_i, q_i \rangle}$: „hossz”, avagy „kettes norma”

Példa: ortogonális mátrixok

Az alábbi mátrixok ortogonálisak:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}.$$

Állítás: ortogonális mátrixok szorzata

Ha $Q_1, Q_2 \in \mathbb{R}^{n\times n}$ ortogonális mátrixok, akkor a szorzatuk, $Q_1 Q_2$ is ortogonális.

Biz.: Tudjuk, hogy $Q_1^\top Q_1 = I$ és $Q_2^\top Q_2 = I$.

Kell, hogy $Q_1 Q_2$ is ortogonális.

Vizsgáljuk:

$$(Q_1 Q_2)^\top (Q_1 Q_2) = Q_2^\top \underbrace{Q_1^\top Q_1}_{I} Q_2 = Q_2^\top Q_2 = I.$$

$\square$

Tartalomjegyzék

Definíció: QR-felbontás

Az $A \in \mathbb{R}^{n \times n}$ mátrix QR-felbontásának nevezzük a $Q \cdot R$ szorzatot, ha $A = QR$, ahol $Q \in \mathbb{R}^{n \times n}$ ortogonális mátrix, $R \in \mathcal{U}$ pedig felső háromszögmátrix.

Tétel: QR-felbontás létezése és egyértelműsége

Ha $\det A \neq 0$, (vagyis az A oszlopvektorai lineárisan függetlenek), akkor A-nak létezik QR-felbontása.
Ha még feltesszük, hogy $r_{ii} > 0 \ \forall \, i$-re, akkor egyértelmű is.

Biz.: Létezés: A bizonyítást a Gram–Schmidt-féle ortogonalizációs eljárás adja: az A mátrix oszlopaiból – amelyek a feltétel értelmében lineárisan függetlenek – előállítjuk a Q oszlopait és R ismeretlen elemeit.

Tekintsük a $Q \cdot R = A$ mátrixszorzást, ahol A-t és Q-t az oszlopaival adtuk meg:

$$\begin{bmatrix} q_1 & q_2 & \dots & q_n \end{bmatrix} \cdot \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & r_{nn} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}.$$

Tekintsük először A első oszlopát, a_1-et. A mátrixszorzásból

$$r_{11} \cdot q_1 = a_1, \;\Rightarrow\; q_1 = \frac{1}{r_{11}} \cdot a_1.$$

Mivel q_1-től azt várjuk el, hogy normált legyen, ezért $r_{11} := \|a_1\|_2$.

QR-felbontás létezés bizonyítás

Tegyük fel, hogy A első $k-1$ oszlopát már felhasználtuk, és így előállítottuk Q első $k-1$ oszlopát, melyek normáltak és egymásra ortogonálisak, valamint R első $k-1$ oszlopának elemeit is ismerjük.

Tekintsük most a_k-t. A mátrixszorzásból felírhatjuk a_k-t, majd kifejezhetjük q_k-t:

$$a_k = \sum_{j=1}^{k} r_{jk} \cdot q_j \quad \Longrightarrow \quad q_k = \frac{1}{r_{kk}} \left(a_k - \sum_{j=1}^{k-1} r_{jk} \cdot q_j \right)$$

Az r_{jk} értékek meghatározásához szorozzuk be skalárisan mindkét oldalt q_i-vel rögzített i értékre $(i = 1, 2, \ldots, k-1)$ és használjuk ki, hogy $\langle q_i, q_j \rangle = \delta_{ij}$, valamint q_k-tól is azt várjuk, hogy merőleges legyen az összes eddigi q_i vektorra:

QR-felbontás létezés bizonyítás

$$q_k = \frac{1}{r_{kk}} \left(a_k - \sum_{j=1}^{k-1} r_{jk} \cdot q_j \right) \quad \left| \cdot q_i \right\rangle \quad (i = 1, \ldots, k-1)$$

$$0 = \langle q_k, q_i \rangle = \frac{1}{r_{kk}} \left(\langle a_k, q_i \rangle - \sum_{j=1}^{k-1} r_{jk} \underbrace{\langle q_j, q_i \rangle}_{\delta_{ij}} \right) =$$

$$= \frac{1}{r_{kk}} \left(\langle a_k, q_i \rangle - r_{ik} \right) \quad \Rightarrow \quad r_{ik} = \langle a_k, q_i \rangle \,.$$

Továbbá q_k-tól még azt várjuk el, hogy normált legyen, ezért

$$r_{kk} = \left\| a_k - \sum_{j=1}^{k-1} r_{jk} \cdot q_j \right\|_2 \,.$$

Így megkaptuk az R mátrix k-adik oszlopának ismeretlen értékeit, az előállított q_k ortogonális az eddigi q_i-kre, valamint normált. $\square$

Biz.: Egyértelműség: Tegyük fel indirekt, hogy legalább két különböző QR-felbontásunk van

$$A = Q_1 R_1 = Q_2 R_2,$$

melyekre a R_1 és R_2 diagonális elemi pozitívak.

A-t szorozzuk balról $Q_2^{-1} = Q_2^\top$-tal és jobbról R_1^{-1}-zel

$$\underbrace{(Q_2^\top Q_1)}_{\text{ortogonális}} = \underbrace{(R_2 R_1^{-1})}_{\in\mathcal{U}}.$$

Legyen $R := R_2 R_1^{-1}$, mivel $Q := Q_2^\top Q_1$ ortogonális mátrix ($R = Q$),

$$Q^\top Q = I = R^\top R.$$

Az $R^\top R = I$ szorzatot felírva:

$$\begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & 0 & \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & r_{nn} \end{bmatrix}$$

$$\begin{bmatrix} r_{11} & 0 & \dots & \\ r_{12} & r_{22} & 0 & \\ \vdots & & \ddots & \vdots \\ r_{1n} & 0 & \dots & r_{nn} \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & \dots & \\ 0 & 1 & 0 & \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

$r_{11} \cdot r_{11} = 1$, amiből $r_{11} > 0$ miatt $r_{11} = 1$.
$j \neq 1$-re

$$r_{11} \cdot r_{1j} = 0 \quad \Rightarrow \quad r_{1j} = 0.$$

R második sorára: $r_{22} \cdot r_{22} = 1$, amiből $r_{22} > 0$ miatt $r_{22} = 1$.

A szorzat mátrix $(2, j)$-edik elemére $j \neq 2$-re

$$r_{22} \cdot r_{2j} = 0 \quad \Rightarrow \quad r_{2j} = 0.$$

A többi sorra ehhez hasonlóan ellenőrizhetjük, hogy

$$R = I \;\; \Leftrightarrow \;\; R_1 = R_2, \;\; Q_1 = Q_2.$$

Ezzel ellentmondásra jutottunk. $\square$

Megj.: Két különböző QR-felbontás esetén létezik olyan $D := \operatorname{diag}(\pm 1, \ldots, \pm 1)$ mátrix, melyre $A = \overbrace{Q \cdot D} \cdot \overbrace{D \cdot R} = \widetilde{Q} \cdot \widetilde{R}$.

Miért jó a QR-felbontás?

Tegyük fel, hogy

- az $Ax = b$ LER megoldható, és
- rendelkezésünkre áll az $A = QR$ felbontás.

Ekkor $Ax = Q \cdot \underbrace{R \cdot x}_{y} = b$ helyett $\quad (\frac{2}{3}n^3 + \mathcal{O}(n^2))$

1. a $Qy = b$ LER megoldása: $y = Q^\top b$, $\quad (2n^2 + \mathcal{O}(n))$
2. az $Rx = y$ LER-t oldjuk meg. $\quad (n^2 + \mathcal{O}(n))$

Együtt is írható: oldjuk meg az $Rx = Q^\top b$ LER-t.

Persze valamikor elő kell állítani a QR-felbontást. $\quad (2n^3 + \mathcal{O}(n^2))$
Előnyös, ha sokszor ugyanaz A, lásd QR-algoritmus (Num. mód. 2A). Így numerikusan stabilabb a LER megoldása.

Feladat: adott $a_1, \dots, a_n \in \mathbb{R}^n$ lineárisan független vektorrendszer, készítsünk belőlük egy $q_1, \dots, q_n \in \mathbb{R}^n$ ortonormált vektorrendszert úgy, hogy q_k csak $a_1, \dots, a_k$-tól függ ($k = 1, 2, \dots, n$).

Másképp, mátrixszorzás alakban: $QR = A$, avagy

$$\begin{bmatrix} q_1 & q_2 & \dots & q_n \end{bmatrix} \cdot \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ & r_{22} & \dots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}$$

Adott: A, keressük: Q, R.

Levezetés: lásd a QR-felbontás létezés bizonyítását (illetve Linalg).

Definíció: Gram–Schmidt-féle ortogonalizáció

Adott az $a_1, \ldots, a_n \in \mathbb{R}^n$ lineárisan független vektorrendszer.

1. $r_{11} := \|a_1\|_2$,
2. $q_1 := \dfrac{1}{r_{11}} a_1$ („lenormáljuk”).

   A k-adik lépésben $(k = 2, \ldots, n)$:
3. $r_{jk} := \langle a_k, q_j \rangle \quad (j = 1, \ldots, k-1)$,
4. $s_k := a_k - \sum_{j=1}^{k-1} r_{jk} \cdot q_j$,
5. $r_{kk} := \|s_k\|_2$ (s_k segédvektor hossza),
6. $q_k := \dfrac{1}{r_{kk}} s_k$ („lenormáljuk”).

Az így nyert $q_1, \ldots, q_n \in \mathbb{R}^n$ vektorrendszer ortonormált.

Definíció: Gram–Schmidt-ortogonalizáció (normálás nélkül)

Adott az $a_1, \ldots, a_n \in \mathbb{R}^n$ lineárisan független vektorrendszer.

1. $\widetilde{q_1} := a_1$,
2. $\widetilde{r_{11}} := 1$

   A k-adik lépésben ($k = 2, \ldots, n$):
3. $\widetilde{r_{jk}} := \dfrac{\langle a_k, \widetilde{q}_j \rangle}{\langle \widetilde{q}_j, \widetilde{q}_j \rangle} \quad (j = 1, \ldots, k-1)$,
4. $\widetilde{q_k} := a_k - \sum\limits_{j=1}^{k-1} \widetilde{r_{jk}} \cdot \widetilde{q}_j$,
5. $\widetilde{r_{kk}} := 1 \quad$ (nem normálunk),

Az így nyert $\widetilde{q_1}, \ldots, \widetilde{q_n} \in \mathbb{R}^n$ vektorrendszer ortogonális.

Megj.: Levezetése teljesen hasonló. Kézi számolásra alkalmasabb. Ne felejtsünk el normálni. . .

Normálás utólag:

- $A = \widetilde{Q}\widetilde{R}$,
- $D := \widetilde{Q}^\top \widetilde{Q}$, azaz $D = \operatorname{diag}\left(\langle \widetilde{q_1}, \widetilde{q_1}\rangle, \dots \langle \widetilde{q_n}, \widetilde{q_n}\rangle\right)$,
- $A = \underbrace{\widetilde{Q} \cdot \sqrt{D}^{-1}}_{Q} \cdot \underbrace{\sqrt{D} \cdot \widetilde{R}}_{R} = Q \cdot R$,

  azaz $\widetilde{Q}$ oszlopait, mint vektorokat leosztjuk azok hosszával (normáljuk őket), $\widetilde{R}$ sorait pedig szorozzuk ugyanezekkel az értékekkel.
- Közvetlenül a $\sqrt{D} = \operatorname{diag}\left(\|\widetilde{q_1}\|_2, \dots, \|\widetilde{q_n}\|_2\right)$ alakkal is dolgozhatunk.

Tétel: A Gram–Schmidt-ortogonalizáció műveletigénye

A szorzások és osztások száma

$$2n^3 + \mathcal{O}(n^2),$$

valamint n darab négyzetgyökvonás is szükséges.

Biz.: A k-adik lépésben:

skaláris szorzatok (r_{jk})	$(k-1)(2n-1)$
ortogonális vektor (s_k)	$(k-1)n + (k-1)n = (k-1)2n$
hossz (r_{kk})	$2n-1$
osztás (q_k)	n

Összesen:
$(k-1)(4n-1) + 3n - 1 = 4kn - 4n - k + 1 + 3n - 1 = 4kn - n - k,$

$$\sum_{k=1}^{n} (4kn - n - k) = 4n \sum_{k=1}^{n} k - n^2 - \sum_{k=1}^{n} k =$$

$$= 4n \cdot \frac{n(n+1)}{2} - n^2 - \frac{n(n+1)}{2} = 2n^3 + \mathcal{O}(n^2).$$

$\square$

Példa: QR, Gram–Schmidt

Készítsük el a következő mátrix QR-felbontását Gram–Schmidt-ortogonalizációval.

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

Gram–Schmidt ortogonalizációval normálással:

$$A = \begin{bmatrix} a_1 & a_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} q_1 & q_2 \end{bmatrix} \cdot \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix} = Q \cdot R.$$

1. lépés: $a_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$-ből meghatározzuk r_{11}, q_1-et:

$$r_{11} = \|a_1\|_2 = \sqrt{1^2 + 2^2} = \sqrt{5}$$

$$q_1 = \frac{1}{r_{11}} a_1 = \frac{1}{\sqrt{5}} a_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

2. lépés:

$$A = \begin{bmatrix} a_1 & a_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} q_1 & q_2 \end{bmatrix} \cdot \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix} = Q \cdot R$$

$a_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$-ből meghatározzuk r_{12}, r_{22}, q_2-t:

$$r_{12} = \langle a_2, q_1 \rangle = \left\langle \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\rangle = \frac{1}{\sqrt{5}}(2 \cdot 1 + 1 \cdot 2) = \frac{4}{\sqrt{5}}$$

$$\begin{aligned} s_2 = a_2 - r_{12} q_1 &= \begin{bmatrix} 2 \\ 1 \end{bmatrix} - \frac{4}{\sqrt{5}} \cdot \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \\ &= \frac{1}{5} \begin{bmatrix} 10 - 4 \\ 5 - 8 \end{bmatrix} = \frac{1}{5} \begin{bmatrix} 6 \\ -3 \end{bmatrix} = \frac{3}{5} \begin{bmatrix} 2 \\ -1 \end{bmatrix} \end{aligned}$$

$$r_{22} = \|s_2\|_2 = \left\| \frac{3}{5} \begin{bmatrix} 2 \\ -1 \end{bmatrix} \right\|_2 = \frac{3}{5} \cdot \sqrt{2^2 + (-1)^2} = \frac{3}{5} \cdot \sqrt{5} = \frac{3}{\sqrt{5}}$$

$$q_2 = \frac{1}{r_{22}} s_2 = \frac{3}{5} \cdot \frac{\sqrt{5}}{3} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

$\square$

Tehát a Q és R mátrixok a következők:

$$Q = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}, \quad R = \begin{bmatrix} \sqrt{5} & \frac{4}{\sqrt{5}} \\ 0 & \frac{3}{\sqrt{5}} \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 5 & 4 \\ 0 & 3 \end{bmatrix}$$

Definíció: vektorok „hossza”

Az $\mathbb{R}^n$-beli v vektorok hagyományos értelemben vett hosszát, avagy „kettes normáját” jelölje $\|.\|_2$.
A következőképpen számolható:

$$\|v\|_2 := \sqrt{\langle v, v\rangle} = \sqrt{v^\top v} = \left(\sum_{k=1}^{n} v_i^2\right)^{\frac{1}{2}}$$

Definíció: Householder-mátrix

A $H = H(v) \in \mathbb{R}^{n\times n}$ mátrixot *Householder-mátrixnak* nevezzük, ha

$$H(v) = I - 2vv^\top,$$

ahol $v \in \mathbb{R}^n$ és $\|v\|_2 = 1$.

Megjegyzés:

- A $H(v)$ transzformációs mátrixot nem kell előállítani, enélkül alkalmazzuk vektorokra, ez a Householder-transzformáció:
- $x \in \mathbb{R}^n$-re $H(v)x = (I - 2vv^\top)x = x - 2v\underbrace{(v^\top x)}_{\in\mathbb{R}}$.
- $y \in \mathbb{R}^n$-re $\quad y^\top H(v) = y^\top(I - 2vv^\top) = y^\top - 2\underbrace{(y^\top v)}_{\in\mathbb{R}} v^\top$.
- Mindkét esetben $4n$ művelet kell a mátrixszal való szorzás $2n^2 + \mathcal{O}(n)$-es műveletigénye helyett.

Állítás: Householder-mátrixok tulajdonságai

1. $H^\top = H$ (szimmetrikus),
2. $H^2 = I$, azaz $H^{-1} = H$ (ortogonális),
3. $H(v) \cdot v = -v$,
4. $\forall y \perp v : \quad H(v) \cdot y = y$.

Biz.: Használjuk ki, hogy $v^\top v = 1$ és $v^\top y = 0$.

1. $(I - 2vv^\top)^\top = I^\top - 2(v^\top)^\top v^\top = I - 2vv^\top$,
2. $(I - 2vv^\top)(I - 2vv^\top) = I - 2vv^\top - 2vv^\top + 4v\underbrace{v^\top v}v^\top = I$,
3. $(I - 2vv^\top)v = v - 2v\underbrace{v^\top v} = v - 2v = -v$,
4. $(I - 2vv^\top)y = y - 2v\underbrace{v^\top y} = y$. $\square$

Megjegyzés:

- $H(v)$ tükröző mátrix, a v-re merőleges (azaz v normálvektorú) $n-1$ dimenziós altérre (0-n átmenő egyenesre, síkra stb.) tükröz.
- Legyen $v \in \mathbb{R}^n$ és $\|v\|_2 = 1$, tetszőleges $x \in \mathbb{R}^n$ vektort bontsunk v-re merőleges és v-vel párhuzamos komponensekre: $x = a + b$, ahol $a \perp v$ és $b || v$. Ekkor az előző tétel utolsó két állítása alapján

$$H(v)x = H(v)a + H(v)b = a - b.$$

- Mivel $H(v)$ ortogonális mátrix, $\|H(v)x\|_2 = \|x\|_2$, vagyis a transzformáció a vektor hosszát nem változtatja meg.

Tétel: tetszőleges tükrözés Householder-mátrixszal

Legyen $a, b \in \mathbb{R}^n,\ a \neq b$ és $\|a\|_2 = \|b\|_2 \neq 0$. Ekkor a

$$v = \pm \frac{a - b}{\|a - b\|_2} \text{ választással } H(v) \cdot a = b.$$

Biz.: Ismerve, hogy $H(v) = I - 2vv^\top$, számoljuk végig a $H(v) \cdot a$ szorzatot. Közben használjuk ki, hogy $\|a\|_2 = \|b\|_2$, azaz $a^\top a = b^\top b$, valamint a skaláris szorzás kommutatív, azaz $a^\top b = b^\top a$.

$$\left(I - 2\frac{(a-b)(a-b)^\top}{\|a-b\|_2^2}\right) \cdot a = a - \frac{2(a-b)(a^\top a - b^\top a)}{(a-b)^\top (a-b)} =$$
$$= a - \frac{2(a-b)(a^\top a - b^\top a)}{a^\top a - a^\top b - b^\top a + b^\top b} = a - \frac{2(a-b)(a^\top a - b^\top a)}{2(a^\top a - b^\top a)} =$$
$$= a - (a-b) = b.$$

Tehát valóban, két különböző, de azonos hosszúságú vektor átvihető egymásba egy Householder-transzformáció által. □

Megjegyzés: Egyébként $H(v) \cdot b = a$ is teljesül.

Példa: Householder-féle tükrözés

Határozzuk meg azt a Householder-féle transzformációt, amely az azonos hosszúságú a, b vektorhoz előállítja azt a v vektort, melyre $H(v) \cdot a = b$. Ellenőrzésképpen végezzük is el a transzformációt.

$$a = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}.$$

$$a - b = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

$\|a - b\|_2 = \sqrt{1^2 + (-2)^2 + 1^2} = \sqrt{6}$

Tehát $v = \frac{a-b}{\|a-b\|_2} = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$ jó választás.

Ellenőrizzük végezzük el a transzformációt a-n:
$H(v) \cdot a = a - 2v \underbrace{(v^\top a)}_{\in \mathbb{R}} = a - 2(v^\top a)v.$

$$H(v) \cdot a = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} - 2 \cdot \frac{1}{\sqrt{6}} \underbrace{\begin{bmatrix} 1 & -2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}}_{3} \cdot \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} - \frac{6}{6} \cdot \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} = b \quad \checkmark$$

$\square$

Példa: Householder-féle tükrözés

Határozzuk meg azt a Householder-féle transzformációt, amely a következő a vektort $b = k \cdot e_1$ alakúra hozza. Ellenőrzésképpen végezzük is el a transzformációt.

$$a = \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}$$

A jó előjel választás σ-nak -1, mert a első eleme pozitív.

$$\sigma = -\|a\|_2 = -\sqrt{2^2 + (-2)^2 + 1^2} = -3$$

Ezzel az előjel választással stabilabb lesz az osztásunk v előállításban.

$$a - \sigma e_1 = \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix} - (-3) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \\ -2 \\ 1 \end{bmatrix}$$

Látjuk, hogy valójában egyetlen műveletet kellett elvégeznünk a vektor első elemén. Ezzel a σ előjelválasztással elérjük, hogy $\|a - \sigma e_1\|_2 \geq \|a\|_2$.

$$\|a - \sigma e_1\|_2 = \sqrt{5^2 + (-2)^2 + 1^2} = \sqrt{30}$$

$$v = \frac{a - \sigma e_1}{\|a - \sigma e_1\|_2} = \frac{1}{\sqrt{30}} \begin{bmatrix} 5 \\ -2 \\ 1 \end{bmatrix} \quad \text{jó választás.}$$

Példa

Ellenőrizzük végezzük el a transzformációt a-n:
$H(v) \cdot a = a - 2v \underbrace{(v^\top a)}_{\in \mathbb{R}} = a - 2(v^\top a)v.$

$$H(v) \cdot a = \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix} - 2 \cdot \frac{1}{\sqrt{30}} \underbrace{\begin{bmatrix} 5 & -2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}}_{15} \cdot \frac{1}{\sqrt{30}} \begin{bmatrix} 5 \\ -2 \\ 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix} - \begin{bmatrix} 5 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} -3 \\ 0 \\ 0 \end{bmatrix} = \sigma \cdot e_1 \quad \checkmark$$

$\square$

Tartalomjegyzék

Felső háromszög alakra hozás

Módszer:

- Legyen adott az $A \in \mathbb{R}^{n \times n}$ invertálható mátrix, első oszlopát jelölje a_1.
- Egy lépésben egy oszlopot kinullázunk a főátló alatt. ($\sim$ GE)
- Így $n-1$ lépésben felső háromszög alakot nyerünk.

Definíció: előjel függvény

$$\mathrm{sgn} : \mathbb{R} \to \mathbb{R}, \qquad \mathrm{sgn}\,(x) = \begin{cases} 1 & \text{ha } x > 0 \\ 0 & \text{ha } x = 0 \\ -1 & \text{ha } x < 0 \end{cases}$$

Megjegyzés: most, a Householder-transzformációknál nem engedhetjük meg a 0 értéket, helyette akár $+1$-et, akár -1-et választhatunk.

1. lépés:

$a_1 \Rightarrow \sigma_1 \cdot e_1$, ahol $\sigma_1 := -\mathrm{sgn}\,(a_{11}) \cdot \|a_1\|_2$ (tehát $|\sigma_1| = \|a_1\|_2$),

$$v_1 := \frac{a_1 - \sigma_1 e_1}{\|a_1 - \sigma_1 e_1\|_2}, \qquad H_1 := H(v_1).$$

Ekkor

$$H_1 \cdot A = H(v_1) \cdot A = \begin{pmatrix} \sigma_1 & * & \dots & * \\ 0 & & & \\ \vdots & & B & \\ 0 & & & \end{pmatrix}.$$

Megjegyzés: σ_1 megválasztásáról. . . így stabilabb.

2. lépés:

$b_1 \Rightarrow \sigma_2 \cdot e_1$, ahol $\sigma_2 := -\mathrm{sgn}\,(b_{11}) \cdot \|b_1\|_2$ (tehát $|\sigma_2| = \|b_1\|_2$),

$$\widetilde{v}_2 := \frac{b_1 - \sigma_2 e_1}{\|b_1 - \sigma_2 e_1\|_2} \in \mathbb{R}^{n-1}$$

Ekkor

$$H(\widetilde{v}_2) \cdot B = \begin{pmatrix} \sigma_2 & * & \dots & * \\ 0 & & & \\ \vdots & & C & \\ 0 & & & \end{pmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}$$

2. lépés (teljes méretben $(n \times n)$ felírva):

$$v_2 := \begin{pmatrix} 0 \\ \widetilde{v_2} \end{pmatrix}, \quad H_2 := H(v_2) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & H(\widetilde{v_2}) & \\ 0 & & & \end{pmatrix}.$$

Ekkor

$$H_2 \cdot H_1 \cdot A = \begin{pmatrix} \sigma_1 & * & * & \dots & * \\ 0 & \sigma_2 & * & \dots & * \\ 0 & 0 & & & \\ \vdots & \vdots & & C & \\ 0 & 0 & & & \end{pmatrix}.$$

Általában, k. lépés:
kinullázzuk az elemeket a főátló alatt a k. oszlopban.

Az ezt megvalósító transzformáció:

$$v_k := \begin{pmatrix} 0_{(1.)} \\ \vdots \\ 0_{(k-1).} \\ \widetilde{v_k} \end{pmatrix} \in \mathbb{R}^n, \quad H_k := H(v_k) = \begin{pmatrix} I_{k-1} & 0 \\ 0 & H(\widetilde{v_k}) \end{pmatrix}.$$

A gyakorlatban csak az $(n-k+1) \times (n-k+1)$-s mátrix részen dolgozunk a k. lépésben, mint a GE-nál. Az $(n-1)$-edik lépés után felső háromszög alakot kapunk.

A Householder-transzformáció alkalmazásai

Egyetlen LER megoldása:

$$Ax = b$$

$$H_1 \cdot A \cdot x = H_1 \cdot b$$

$$\vdots$$

$$\underbrace{H_{n-1} \cdots H_1 \cdot A}_{R} \cdot x = \underbrace{H_{n-1} \cdots H_1 \cdot b}_{d}$$

$$R \cdot x = d \to x \text{ (visszahelyettesítés)}$$

Ugyanúgy dolgozunk, mint a GE-nál. Végrehajtjuk a transzformációt az oszlopokon:

$$[A|b] \ \to \text{n-1 db H-trf.} \to \ [R|d] \ \to \text{visszahely.}$$

Mindig egyre kisebb méretű mátrixon dolgozunk a transzformációk során.

A Householder-transzformáció alkalmazásai

QR-felbontás készítése:

$$\underbrace{H_{n-1} \cdots H_2 \cdot H_1}_{Q^{-1}=Q^\top} \cdot A = R$$

$$A = \underbrace{H_1 \cdot H_2 \cdots H_{n-1}}_{Q} \cdot R = Q \cdot R$$

Megfigyelhetjük, hogy Q előállításakor mindig a jobb oldalról végezzük a transzformációt, ekkor sorokra alkalmazzuk.

Az algoritmus: Q előállítására

$$\begin{aligned} Q_0 &= I \\ k = 1, \ldots, n-1: \quad Q_k &:= Q_{k-1} H_k \\ Q &:= Q_{n-1} \end{aligned}$$

Tétel: QR-felbontás Householder-módszerrel

Invertálható mátrixok QR-felbontása elkészíthető $n-1$ db Householder-transzformáció segítségével.

Biz.: Láttuk. $\square$

Összefoglalva: A k. lépésben kinullázzuk a k. oszlop főátló alatti elemeit egy H_k ortogonális transzformáció segítségével, melyet a mátrix oszlopaira alkalmazunk a jobb alsó $(n-k+1) \times (n-k+1)$-s mátrix részen.
A Q mátrixot úgy kapjuk, hogy egy egységmátrixból indulva a k. lépésben a H_k transzformációt jobbról alkalmazzuk a sorokra csak a jobb alsó $(n-k+1) \times (n-k+1)$-s mátrix részen.
$n-1$ lépés után megkapjuk felső háromszög alakot (R) és Q-t.

Tartalomjegyzék

Tétel: A Householder-trf. műveletigénye LER-re

A LER megoldásának műveletigénye
Householder-transzformációkkal:

$$\frac{4}{3}n^3 + \mathcal{O}(n^2),$$

valamint $2(n-1)$ darab négyzetgyökvonásra is szükség van.

Biz.:
A k-adik lépésben ($n-k+1 =: h_k$ hosszú vektorokkal dolgozunk):

hossz (σ)	$2h_k - 1$,
normálvektor ($a - \sigma e_1, \|.\|_2, v$)	$1 + (2h_k - 1) + h_k = 3h_k$,
transzformáció ($(h_k - 1) + 1$ vektorra)	$h_k \cdot 4h_k$.

Összesen: $4h_k^2 + 5h_k - 1$, ($n - k =: s$)

$$\sum_{k=1}^{n-1} \left(4h_k^2 + 5h_k - 1\right) = \sum_{s=2}^{n} 4s^2 + \sum_{s=2}^{n} s + (n-1) = \frac{4}{3}n^3 + \mathcal{O}(n^2).$$

A visszahelyettesítés műveletigénye $n^2 + \mathcal{O}(n)$, belefér az előző alakba. $\square$

Tétel: A Householder-trf. műveletigénye QR-felbontásra

A QR-felbontás előállításának műveletigénye Householder-transzformációkkal:

$$\frac{8}{3}n^3 + \mathcal{O}(n^2),$$

valamint $2(n-1)$ darab négyzetgyökvonásra is szükség van.

Biz.:
A k-adik lépésben ($n - k + 1 =: h_k$ hosszú vektorokkal dolgozunk):

hossz (σ)	$2h_k - 1,$
normálvektor ($a - \sigma e_1, \|.\|_2, v$)	$1 + (2h_k - 1) + h_k = 3h_k,$
transzformáció (($h_k - 1) + h_k$ vektorra)	$(2h_k - 1) \cdot 4h_k.$

Összesen: $(8h_k^2 - 4h_k) + (5h_k - 1) = 8h_k^2 + h_k - 1$, ($n - k =: s$)

$$\sum_{k=1}^{n-1} \left(8h_k^2 + h_k - 1\right) = \sum_{s=2}^{n} 8s^2 + \sum_{s=2}^{n} s - (n-1) = \frac{8}{3}n^3 + \mathcal{O}(n^2).$$

$\square$

Megjegyzés: Ez kicsit több, mint a Gram–Schmidt-féle ortogonalizációnál, viszont ez a módszer numerikusan stabilabb.

Példák Matlab-ban

1. A Gram–Schmidt-féle ortogonalizációs eljárás működésének szemléltetése $\mathbb{R}^3$-beli vektorrendszer esetén.
2. Példák Householder-mátrixokra ($n \approx 3, 10, 20, 50$).
3. Példák Householder-transzformációra.
4. QR-felbontás készítése Householder módszerével ($n \approx 3, 7, 50, 100$).

Numerikus módszerek 1.

6. előadás: Vektor- és mátrixnormák

Krebsz Anna

ELTE IK

Emlékeztető: Vektorok „hossza"

Definíció: vektorok „hossza"

Az $x \in \mathbb{R}^n$ vektor hagyományos értelemben vett hosszát, avagy „kettes normáját" jelölje $\|.\|_2$.
A következőképpen számolható:

$$\|x\|_2 := \sqrt{\langle x, x \rangle} = \sqrt{x^\top x} = \left(\sum_{k=1}^{n} x_i^2 \right)^{\frac{1}{2}}.$$

A *(vektor)norma* a „hossz", „nagyság" általánosítása.

Definíció: vektornorma

Legyen $n \in \mathbb{N}$ rögzített. Az $\|.\| : \mathbb{R}^n \to \mathbb{R}$ leképezést vektornormának nevezzük, ha:

1. $\|x\| \geq 0 \quad (\forall x \in \mathbb{R}^n)$,
2. $\|x\| = 0 \iff x = 0$,
3. $\|\lambda \cdot x\| = |\lambda| \cdot \|x\| \quad (\forall \lambda \in \mathbb{R},\ \forall x \in \mathbb{R}^n)$,
4. $\|x + y\| \leq \|x\| + \|y\| \quad (\forall x, y \in \mathbb{R}^n)$.

Azaz a leképezés „pozitív", „pozitív homogén" és „szubadditív" (háromszög-egyenlőtlenség). Ezek a vektornormák *axiómái*.

Állítás: skaláris szorzat által generált vektornorma

Ha adott az $\langle .,. \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ skaláris szorzat, akkor az $f(x) := \sqrt{\langle x, x \rangle}$ függvény *norma*. Jele: $\|x\|_2$.

Biz.: Nem kell.
Ez a „hagyományos hossz”. $\square$

Állítás: Cauchy–Bunyakovszki–Schwarz-egyenlőtlenség (CBS)

$$|\langle x, y\rangle| \leq \|x\|_2 \cdot \|y\|_2 \quad (x, y \in \mathbb{R}^n)$$

Biz.: Bármely $\alpha \in \mathbb{R}$ esetén $\|x - \alpha y\|_2^2 \geq 0$.

$$0 \leq \|x - \alpha y\|_2^2 = \langle x - \alpha y, x - \alpha y\rangle =$$
$$= \underbrace{\langle x, x\rangle}_{\|x\|_2^2} - 2\alpha \langle x, y\rangle + \alpha^2 \underbrace{\langle y, y\rangle}_{\|y\|_2^2} \qquad (\forall \alpha \in \mathbb{R}).$$

Diszkrimináns nempozitív: $\langle x, y\rangle^2 - \|x\|_2^2 \cdot \|y\|_2^2 \leq 0$, így

$$\langle x, y\rangle^2 \leq \|x\|_2^2 \cdot \|y\|_2^2.$$

$\square$

Állítás: Gyakori vektornormák $(1, 2, \infty)$

A következő formulák vektornormákat **definiálnak** $\mathbb{R}^n$ felett:

- $\|x\|_1 := \sum_{i=1}^{n} |x_i|$ (Manhattan-norma),
- $\|x\|_2 := \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$ (Euklideszi-norma),
- $\|x\|_\infty := \max_{i=1}^{n} |x_i|$ (Csebisev-norma).

Biz.: Hf.

Példa: vektornormák

Számítsuk ki a következő vektorok $1, 2, \infty$ normáját:

$$x = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \qquad y = \begin{bmatrix} 4 \\ -8 \\ 1 \end{bmatrix}.$$

$\|x\|_1 = 3 + 4 = 7$, $\|x\|_2 = \sqrt{3^2 + 4^2} = 5$, $\|x\|_\infty = \max\{3, 4\} = 4$.

$\|y\|_1 = 4 + |-8| + 1 = 13$, $\|y\|_2 = \sqrt{4^2 + (-8)^2 + 1^2} = \sqrt{73}$,
$\|y\|_\infty = \max\{4, |-8|, 1\} = 8$.

Állítás: *p*-normák

A következő $\mathbb{R}^n \to \mathbb{R}$ függvények is vektornormákat **definiálnak**:

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \qquad (p \in \mathbb{R},\ 1 \le p < \infty).$$

Biz.: Nem kell. A háromszög-egyenlőtlenség a Minkovszki-egyenlőtlenség. $\square$

Megjegyzések:

- $0 \le p < 1$ esetén nem norma,
- $p_1 \le p_2 \Longrightarrow \|x\|_{p_1} \ge \|x\|_{p_2}$,
- Speciális esetek: $p = 1 \rightsquigarrow \|x\|_1$, $\quad p = 2 \rightsquigarrow \|x\|_2$,
- Sőt: $\lim\limits_{p\to\infty} \|x\|_p = \|x\|_\infty$.

Állítás: normák közötti egyenlőtlenségek

- $\|x\|_\infty \leq \|x\|_1 \leq n \cdot \|x\|_\infty$,
- $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \cdot \|x\|_\infty$,
- $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \cdot \|x\|_2$,
- sőt ezek alapján $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$.

Biz.: Nem kell. $\square$

(Az elsőbe könnyű belegondolni, a negyedikre láttunk példát.)

Definíció: ekvivalens normák

Az $\|.\|_a$ és $\|.\|_b$ vektornormák *ekvivalensek*, ha $\exists c_1, c_2 \in \mathbb{R}^+$, hogy

$$c_1 \cdot \|x\|_b \leq \|x\|_a \leq c_2 \cdot \|x\|_b \qquad (\forall x \in \mathbb{R}^n).$$

Állítás: végesdimenziós normák ekvivalenciája

Tetszőleges $\mathbb{R}^n$-en értelmezett vektornorma ekvivalens az Euklideszi-vektornormával. (Azaz adott végesdimenziós térben minden norma ekvivalens.)

Definíció: konvergencia vektornormában

Az $(x_k) \subset \mathbb{R}^n$ sorozat konvergens, ha létezik $x^* \in \mathbb{R}^n$ melyre

$$\lim_{k\to\infty} \|x_k - x^*\| = 0.$$

x^* a sorozat határértéke.

Megj.: Mivel $\mathbb{R}^n$-en a vektornormák ekvivalensek, ezért ha egy sorozat konvergens az egyik vektornormában, akkor mindegyikben.

Ekvivalens átfogalmazások a konvergenciára:

- Az $(x_k) \subset \mathbb{R}^n$ sorozat konvergens, ha létezik $x^* \in \mathbb{R}^n$ melyre

$$\forall\, \varepsilon > 0 \;\; \exists\, N_0 \in \mathbb{N} \;\; \forall\, k \geq N_0 : \quad \|x_k - x^*\| < \varepsilon.$$

- Az $(x_k) \subset \mathbb{R}^n$ sorozat konvergens, ha létezik $x^* \in \mathbb{R}^n$ melyre

$$\forall\, \varepsilon > 0 \;\; \exists\, N_0 \in \mathbb{N} \;\; \forall\, k \geq N_0 : \quad x_k \in K_\varepsilon(x^*).$$

Matlab példák p-normákra, egységgömbökre $(p = 1, 2, \infty, \dots)$.

Definíció: mátrixnorma

Legyen $n \in \mathbb{N}$ rögzített. Az $\|.\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ leképezést mátrixnormának nevezzük, ha:

1. $\|A\| \geq 0 \quad (\forall A \in \mathbb{R}^{n\times n})$,
2. $\|A\| = 0 \iff A = 0$,
3. $\|\lambda \cdot A\| = |\lambda| \cdot \|A\| \quad (\forall \lambda \in \mathbb{R},\ \forall A \in \mathbb{R}^{n\times n})$,
4. $\|A + B\| \leq \|A\| + \|B\| \quad (\forall A, B \in \mathbb{R}^{n\times n})$,
5. $\|A \cdot B\| \leq \|A\| \cdot \|B\| \quad (\forall A, B \in \mathbb{R}^{n\times n})$.

Ugyanaz, mint a vektornormáknál, plusz: „szubmultiplikativitás”.
Ezek a mátrixnormák axiómái.

Definíció: Frobenius-norma

A következő függvényt *Frobenius-normának* nevezzük:

$$\|.\|_F : \mathbb{R}^{n\times n} \to \mathbb{R}, \qquad \|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2}.$$

Állítás: Frobenius-norma

A $\|.\|_F$ függvény valóban mátrixnorma.

Biz.: 1–4. következik a $\|.\|_2$ vektornorma tulajdonságaiból.
Az 5. belátható CBS segítségével. $\square$

Példa: egyszerű mátrixnormák

Számítsuk ki a következő mátrixok Frobenius-normáját.

$$A = \begin{bmatrix} 1 & -4 \\ 2 & 2 \end{bmatrix}, \qquad B = \begin{bmatrix} 3 & 2 \\ 1 & 5 \end{bmatrix}.$$

$$\|A\|_F = \sqrt{1^2 + (-4)^2 + 2^2 + 2^2} = 5$$

$$\|B\|_F = \sqrt{3^2 + 2^2 + 1^2 + 5^2} = 6$$

Természetes mátrixnormák, avagy indukált normák

Definíció: indukált norma, természetes mátrixnormák

Legyen $\|.\|_v : \mathbb{R}^n \to \mathbb{R}$ tetszőleges vektornorma. Ekkor a

$$\|.\| : \mathbb{R}^{n\times n} \to \mathbb{R}, \qquad \|A\| := \sup_{x\neq 0} \frac{\|Ax\|_v}{\|x\|_v}$$

függvényt a $\|.\|_v$ *vektornorma által indukált mátrixnormának* hívjuk. Egy mátrixnormát *természetesnek* nevezünk, ha van olyan vektornorma, ami indukálja.

Tétel: indukált normák

Az „indukált mátrixnormák" valóban mátrixnormák.

Természetes mátrixnormák, avagy indukált normák

Biz.: Be kell látni, hogy a megadott alak teljesíti a mátrixnorma axiómáit.

1. Az $\|A\|$ értéke nemnegatív, hiszen vektorok normájának (nemnegatív számok) hányadosainak szuprémuma.
2. Ha $A = 0$, azaz nullmátrix, akkor $\|Ax\|_v = 0$ minden x vektorra, így a szuprémum értéke is 0. Valamint megfordítva, ha a szuprémum 0, akkor minden x-re Ax-nek nullvektornak kell lennie, ez csak úgy lehet, ha A nullmátrix.
3. $$\|\lambda A\| = \sup_{x \neq 0} \frac{\|\lambda A x\|_v}{\|x\|_v} = \sup_{x \neq 0} \frac{|\lambda| \cdot \|Ax\|_v}{\|x\|_v} = |\lambda| \cdot \sup_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v} = |\lambda| \cdot \|A\| \, .$$

Természetes mátrixnormák, avagy indukált normák

Biz. (folytatás):

4.

$$\|A+B\| = \sup_{x\neq 0} \frac{\|(A+B)x\|_v}{\|x\|_v} \leq \sup_{x\neq 0} \frac{\|Ax\|_v + \|Bx\|_v}{\|x\|_v} \leq$$

$$\leq \sup_{x\neq 0} \frac{\|Ax\|_v}{\|x\|_v} + \sup_{x\neq 0} \frac{\|Bx\|_v}{\|x\|_v} = \|A\| + \|B\|$$

5. $B = 0 \;\; \Rightarrow \;\; \|B\| = 0$, valamint
$A \cdot B = A \cdot 0 = 0 \;\; \Rightarrow \;\; \|AB\| = 0.$
Az egyenlőtlenség mindkét oldalán 0 áll, tehát igaz az állítás.

Biz. (folytatás): Ha $B \neq 0$, akkor

$$\|A \cdot B\| = \sup_{x \neq 0} \frac{\|ABx\|_v}{\|x\|_v} = \sup_{x \neq 0, Bx \neq 0} \frac{\|ABx\|_v}{\|Bx\|_v} \cdot \frac{\|Bx\|_v}{\|x\|_v} \leq$$

$$\leq \sup_{Bx \neq 0} \frac{\|ABx\|_v}{\|Bx\|_v} \cdot \sup_{x \neq 0} \frac{\|Bx\|_v}{\|x\|_v} \leq \sup_{y \neq 0} \frac{\|Ay\|_v}{\|y\|_v} \cdot \sup_{x \neq 0} \frac{\|Bx\|_v}{\|x\|_v} = \|A\| \cdot \|B\| \, .$$

Meggondolható, hogy a $Bx \neq 0$ feltétel nem változtatja meg a szuprémum értékét; közben bevezettük az $y := Bx$ jelölést. $\square$

Természetes mátrixnormák, avagy indukált normák

Megjegyzések:

- Átfogalmazás:

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v} = \sup_{\|y\|_v = 1} \|Ay\|_v \,.$$

- A sup helyett max is írható.
- Átfogalmazás:

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v} \implies \frac{\|Ax\|_v}{\|x\|_v} \leq \|A\| \implies \|Ax\|_v \leq \|A\| \cdot \|x\|_v \,.$$

Sőt: $\|A\|$ a legkisebb ilyen felső korlát.

Természetes mátrixnormák, avagy indukált normák

Definíció: illeszkedő normák

Ha egy mátrix- és egy vektornormára

$$\|Ax\|_v \leq \|A\| \cdot \|x\|_v \qquad (\forall x \in \mathbb{R}^n,\ A \in \mathbb{R}^{n\times n})$$

teljesül, akkor *illeszkedőknek* nevezzük őket.

Állítás: természetes mátrixnormák illeszkedéséről

A természetes mátrixnormák illeszkednek az őket indukáló vektornormákhoz.

Biz.: Láttuk az előbb. Az $x = 0$ eset meggondolandó. $\square$

Természetes mátrixnormák, avagy indukált normák

Milyen mátrixnormákat indukálnak az elterjedt vektornormák?

Tétel: Nevezetes mátrixnormák $(1, 2, \infty)$

A $\|.\|_p$ $(p = 1, 2, \infty)$ vektornormák által indukált mátrixnormák:

- $\|A\|_1 = \max_{j=1}^{n} \sum_{i=1}^{n} |a_{ij}|$ (oszlopnorma),
- $\|A\|_\infty = \max_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|$ (sornorma),
- $\|A\|_2 = \left(\max_{i=1}^{n} \lambda_i(A^\top A) \right)^{1/2}$ (spektrálnorma).

Jel.: $\lambda_i(M)$: az M mátrix i-edik sajátértéke $(Mv = \lambda v,\ v \neq 0)$.

Természetes mátrixnormák, avagy indukált normák

A bizonyítás „dallama":

- Az adott $f(A)$ értékre: $\|Ax\|_v \leq f(A) \cdot \|x\|_v$.
- Van olyan x vektor, hogy $\|Ax\|_v = f(A) \cdot \|x\|_v$.
- Ekkor az $f(A)$ érték, tényleg a $\|.\|_v$ vektornorma által indukált mátrixnorma, ezért jelölhetjük így: $\|A\|_v$.

Bizonyítás $\|.\|_1$ esetén:

Állítás: $\|A\|_1 = \max_{j=1}^{n} \sum_{i=1}^{n} |a_{ij}|$.

$$\|Ax\|_1 = \sum_{i=1}^{n} |(Ax)_i| = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}| \cdot |x_j| =$$

$$= \sum_{j=1}^{n} \left(|x_j| \cdot \underbrace{\sum_{i=1}^{n} |a_{ij}|}_{} \right) \leq \underbrace{\left(\max_{j=1}^{n} \sum_{i=1}^{n} |a_{ij}| \right)}_{} \cdot \|x\|_1 .$$

Legyen $x = e_k$, ahol a k-adik oszlopösszeg maximális. Ekkor

$$\|Ae_k\|_1 = \underbrace{\ldots}_{} \underbrace{\|e_k\|_1}_{1} .$$

Természetes mátrixnormák, avagy indukált normák

Bizonyítás $\|.\|_\infty$ esetén: Állítás: $\|A\|_\infty = \max_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|$.

- A becslés ugyanolyan stílusú, mint $\|.\|_1$ esetén. Gyakorlaton.
- Válasszuk az
$$x = \begin{pmatrix} \pm 1 \\ \vdots \\ \pm 1 \end{pmatrix}$$
vektort az egyenlőséghez, megfelelően választott előjelekkel...

Bizonyítás $\|.\|_2$ esetén: Állítás: $\|A\|_2 = \left(\max_{i=1}^{n} \lambda_i(A^\top A) \right)^{1/2}$.

- Előbb belátjuk, hogy a sajátértékek nemnegatívak.
- A becslés a diagonalizálás alapján adódik.
- Válasszuk a legnagyobb sajátértékhez tartozó sajátvektort az egyenlőséghez.

Természetes mátrixnormák, avagy indukált normák

Biz. (folytatás): Először belátjuk, hogy $A^\top A$ szimmetrikus és sajátértékei nemnegatívak (azaz $A^\top A$ pozitív szemidefinit).

- $(A^\top A)^\top = A^\top (A^\top)^\top = A^\top A$, azaz $A^\top A$ szimmetrikus, vagyis $A^\top A$ sajátértékei valósak.

- Legyen $y \neq 0$ az $A^\top A$ mátrix λ-hoz tartozó sajátvektora, azaz

  $$A^\top A y = \lambda \cdot y.$$

  Szorozzuk meg mindkét oldalt balról az $y^\top$ vektorral:

  $$y^\top A^\top A y = \lambda \cdot y^\top y.$$

  Innen

  $$\lambda = \frac{y^\top A^\top A y}{y^\top y} = \frac{(Ay)^\top (Ay)}{y^\top y} = \frac{\|Ay\|_2^2}{\|y\|_2^2} \geq 0.$$

Természetes mátrixnormák, avagy indukált normák

Biz. (folytatás): Ezután az indukált mátrixnormák definícióját követve Ax normáját fogjuk vizsgálni.

Kihasználjuk, hogy $A^\top A$ szimmetrikus, és így (lásd lineáris algebra) létezik U ortogonális (unitér) mátrix, amire

$$A^\top A = U^\top D U \quad \Leftrightarrow \quad U A^\top A U^\top = D$$

úgy, hogy a diagonálisban $A^\top A$ sajátértékei vannak (ezek nemnegatívak). Bevezetjük az $y = Ux$ jelölést.

$$\|Ax\|_2^2 = (Ax)^\top (Ax) = x^\top A^\top A x = x^\top U^\top D U x = (Ux)^\top D (Ux)$$

$$= y^\top D y = \sum_{i=1}^{n} \underbrace{d_{ii}}_{\geq 0} \cdot |y_i|^2 \leq \max_{i=1}^{n} d_{ii} \cdot \sum_{i=1}^{n} |y_i|^2 = \max_{i=1}^{n} \lambda_i(A^\top A) \cdot \|y\|_2^2 \,.$$

Belátjuk, hogy $\|y\|_2^2 = \|x\|_2^2$.

Természetes mátrixnormák, avagy indukált normák

$\|y\|_2^2 = y^\top y = (Ux)^\top (Ux) = x^\top U^\top U x = x^\top x = \|x\|_2^2$, ezért

$$\|Ax\|_2^2 \leq \ldots \leq \max_{i=1}^{n} \lambda_i(A^\top A) \cdot \|x\|_2^2 \,.$$

$x \neq 0$ esetén:

$$\frac{\|Ax\|_2}{\|x\|_2} \leq \left(\max_{i=1}^{n} \lambda_i(A^\top A) \right)^{1/2}$$

Még azt kell belátni, hogy van is olyan $x \neq 0$ vektor, amire a szuprémum felvétetik.

Legyen $\lambda_m = \max \lambda_i(A^\top A)$ és $v_m \neq 0$, $\|v_m\|_2 = 1$ a hozzá tartozó sajátvektor.

$$\|Av_m\|_2^2 = (Av_m)^\top (Av_m) = v_m^\top \underbrace{A^\top A\, v_m}_{\lambda_m \cdot v_m} = \lambda_m \cdot \underbrace{v_m^\top v_m}_{=1} = \lambda_m.$$

$\square$

Definíció: spektrálsugár

Egy $A \in \mathbb{R}^{n\times n}$ mátrix *spektrálsugara* $\varrho(A) := \max_{i=1}^{n} |\lambda_i(A)|$.

Megj.: A spektrálnormát a spektrálsugárral is meg tudjuk adni:

$$\|A\|_2 = \sqrt{\varrho(A^\top A)}.$$

Állítás:

Egy $A \in \mathbb{R}^{n\times n}$ szimmetrikus (önadjungált) mátrix spektrálnormája

$$\|A\|_2 = \varrho(A).$$

Biz.: Trivi.

Állítás:

Ha A normális ($A^*A = AA^*$), akkor $\|A\|_2 = \varrho(A)$.
(Spec.: ha A önadjungált, akkor normális.)

Biz.: Lineáris algebrából ismert, hogy normális mátixok esetén létezik U unitér hasonlósági transzformáció, mellyel A diagonális alakra hozható.

$$U^*AU = D = \mathrm{diag}\,(\lambda_i(A)) \quad \Leftrightarrow \quad A = UDU^*$$
$$A^*A = (UDU^*)^*UDU^* = UD^*U^*UDU^* = UD^*DU^*$$
$$\lambda_i(A^*A) = \lambda_i(D^*D) = |\lambda_i(A)|^2$$
$$\varrho(A^*A) = \varrho(A)^2$$

Innen $\|A\|_2 = \varrho(A^*A)^{1/2} = \varrho(A)$.

Példa: $\|.\|_1$ és $\|.\|_\infty$ mátrixnormára

Számítsuk ki a következő mátrix $\|.\|_1$ és $\|.\|_\infty$ mátrixnormáját.

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

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

$$\|A\|_\infty = \max\{1 + |-4|, 2 + 2\} = 5$$

Példa: $\|.\|_2$ mátrixnorma

Számítsuk ki a következő mátrix $\|.\|_2$ mátrixnormáját.

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

$$A^\top A = \begin{bmatrix} 1 & 2 \\ -4 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & -4 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 20 \end{bmatrix},$$

Szerencsénkre látjuk a sajátértékeit...

$$\|A\|_2 = \left(\max_{i=1}^{n} \lambda_i(A^\top A)\right)^{1/2} = \sqrt{\max\{5, 20\}} = \sqrt{20} \approx 4.4721.$$

Állítás

A Frobenius-norma nem természetes mátrixnorma.

Biz.: Tekintsük az $I \in \mathbb{R}^{n \times n}$ egységmátrix normáját.

- Indukált mátrixnormák esetén $\|I\| = \sup\limits_{x \neq 0} \frac{\|Ix\|_v}{\|x\|_v} = 1$.
- Másrészt $\|I\|_F = \sqrt{n}$.
- Tehát nincs olyan vektornorma, ami a Frobenius-normát indukálná (ha $n > 1$). $\square$

Állítás: spektrálsugár és norma

$$\varrho(A) \leq \|A\|$$

Biz.: Belátjuk, hogy $|\lambda| \leq \|A\|$.
(Legyen λ tetszőleges sajátérték és $v \neq 0$ a hozzátartozó sajátvektor.)

$$Av = \lambda v$$

$$Avv^\top = \lambda vv^\top$$

$$\|A\| \cdot \left\|vv^\top\right\| \geq \left\|Avv^\top\right\| = \left\|\lambda vv^\top\right\| = |\lambda| \cdot \left\|vv^\top\right\|$$

Leosztva $\left\|vv^\top\right\| \neq 0$-val $\|A\| \geq |\lambda|$. $\square$

Mátrixnormák további tulajdonságai

Feladatok gyakorlatra

Igazoljuk a következő állításokat.

(a) Ha Q ortogonális (unitér), akkor

- $\|Qx\|_2 = \|x\|_2$,
- $\|Q\|_2 = 1$,
- $\|QA\|_2 = \|AQ\|_2 = \|A\|_2$.

Mátrixnormák további tulajdonságai

Feladatok gyakorlatra

(b) $\|A\|_F^2 = \operatorname{tr}(A^\top A)$, ahol $\operatorname{tr}(B) := \sum_{k=1}^{n} b_{kk}$ a mátrix *nyoma*.

(c) Ha Q ortogonális (unitér), akkor $\|QA\|_F = \|AQ\|_F = \|A\|_F$.

(d) $\|A\|_F^2 = \sum_{i=1}^{n} \lambda_i(A^\top A)$.

(e) $\|.\|_F$ és $\|.\|_2$ ekvivalens mátrixnormák.

(f) A Frobenius-norma illeszkedik a kettes vektornormához.

Példák Matlab-ban

1. Indukált mátrixnorma szemléltetése $\mathbb{R}^2$, $p = 2$ esetén.
2. Indukált mátrixnormák közelítő számítása tetszőleges $\mathbb{R}^n$ és p esetén ($m = 100, \ldots, 1000$ vektor próbájával).

Numerikus módszerek 1.

7. előadás: LER érzékenysége

Krebsz Anna

ELTE IK

Tartalomjegyzék

Tartalomjegyzék

Definíció: mátrixok kondíciószáma

Adott $A \in \mathbb{R}^{n \times n}$ invertálható mátrix és $\|.\|$ mátrixnorma esetén a $\operatorname{cond}(A) := \|A\| \cdot \|A^{-1}\|$ mennyiséget az A mátrix *kondíciószámának* nevezzük. (Jele néha $\kappa(A)$. [kappa])

Megjegyzés:

- Csak invertálható mátrixokra értelmes.
- Értéke függ a norma választásától.
  (Pl. $\operatorname{cond}_1(A), \operatorname{cond}_2(A), \ldots$)

Állítás: a kondíciószám tulajdonságai – 1. rész

(a) Indukált mátrixnorma esetén $\operatorname{cond}(A) \geq 1$.

(b) $\operatorname{cond}(c \cdot A) = \operatorname{cond}(A), \quad (c \in \mathbb{R}, c \neq 0)$.

(c) Ha Q ortogonális, akkor $\operatorname{cond}_2(Q) = 1$.

Biz.:

(a) $1 = \|I\| = \left\|A \cdot A^{-1}\right\| \leq \|A\| \cdot \left\|A^{-1}\right\| = \operatorname{cond}(A)$.

(b) $\operatorname{cond}(cA) = \|cA\| \cdot \left\|(cA)^{-1}\right\| = \|cA\| \cdot \left\|\frac{1}{c}A^{-1}\right\| =$

$$= |c| \cdot \|A\| \cdot \frac{1}{|c|} \cdot \left\|A^{-1}\right\| = \operatorname{cond}(A).$$

(c) $\|Q\|_2 = \sup\limits_{x \neq 0} \dfrac{\|Qx\|_2}{\|x\|_2} = \sup\limits_{x \neq 0} \dfrac{\sqrt{x^\top Q^\top Q x}}{\sqrt{x^\top x}} = 1$

$$\left\|Q^{-1}\right\|_2 = \left\|Q^\top\right\|_2 = 1, \quad \operatorname{cond}_2(Q) = 1 \qquad \square$$

Állítás: a kondíciószám tulajdonságai – 2. rész

(d) Ha A szimmetrikus, akkor $\operatorname{cond}_2(A) = \dfrac{\max|\lambda_i(A)|}{\min|\lambda_i(A)|}$.

(e) Ha A szimm., pozitív definit, akkor $\operatorname{cond}_2(A) = \frac{\max \lambda_i(A)}{\min \lambda_i(A)}$.

(f) Ha A invertálható, akkor $\operatorname{cond}(A) \geq \frac{\max|\lambda_i(A)|}{\min|\lambda_i(A)|}$.

Biz.:

(d) Eml.: $\|A\|_2 = \sqrt{\max \lambda_i(A^\top A)}$.
De $\lambda_i(A^\top A) = \lambda_i(A^2) = (\lambda_i(A))^2$, így $\|A\|_2 = \max|\lambda_i(A)|$.
Az inverzre: $\|A^{-1}\|_2 = \max|\lambda_i(A^{-1})| = \dfrac{1}{\min|\lambda_i(A)|}$.

(e) A pozitiv definitség miatt nem kell abszolút érték.

(f) $\|A\| \geq \varrho(A) = \max|\lambda_i(A)|$, $\|A^{-1}\| \geq \varrho(A^{-1}) = \frac{1}{\min|\lambda_i(A)|}$. $\square$

$$A \cdot x = b$$

Vizsgáljuk meg, hogy hogyan változik meg a LER megoldása, ha a jobb oldalt, azaz a vektort *kicsit* megváltoztatjuk, „perturbáljuk''! (Mérési pontatlanság, kerekítési hiba, ...)

1. **Eredeti:**
   adott A és b, kiszámíthatjuk a megoldást: x.
   $Ax = b$
2. **Módosult:**
   adott A és $b + \Delta b$, kiszámíthatjuk a megoldást: $x + \Delta x$.
   $A(x + \Delta x) = (b + \Delta b)$

Nyilván a megoldás is *kicsit* más lesz...

Példa:

1. **Eredeti:**

$$\begin{bmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{bmatrix} \cdot x = \begin{bmatrix} 4.1 \\ 9.7 \end{bmatrix} \quad \rightarrow \quad \text{megoldás: } x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

2. **Módosult:**

$$\begin{bmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{bmatrix} \cdot (x + \Delta x) = \begin{bmatrix} 4.11 \\ 9.7 \end{bmatrix}$$

3.

$$\text{A módosult LER megoldása:} \quad x + \Delta x = \begin{bmatrix} 0.34 \\ 0.97 \end{bmatrix}$$

4. Mi történt?

LER vektorának megváltozása

Hogyan jellemezhető a megoldás megváltozása a jobb oldal megváltozásához képest?

- Mennyire változott a jobb oldal:
  $\delta b := \dfrac{\|\Delta b\|}{\|b\|} = 9.4959e - 004.$
- Emiatt mennyire változik a megoldás: $\delta x := \dfrac{\|\Delta x\|}{\|x\|} = 1.1732.$
- Vizsgáljuk a kettő hányadosát: $\dfrac{\delta x}{\delta b} = 1235.5.$
- $\mathrm{cond}\,(A) = 1623$

Tétel: LER érzékenysége a jobb oldal pontatlanságára

Ha A invertálható és $b \neq 0$, akkor illeszkedő normákban

$$\frac{1}{\|A\| \cdot \|A^{-1}\|} \cdot \frac{\|\Delta b\|}{\|b\|} \leq \frac{\|\Delta x\|}{\|x\|} \leq \|A\| \cdot \left\|A^{-1}\right\| \cdot \frac{\|\Delta b\|}{\|b\|},$$

azaz

$$\frac{1}{\text{cond}\,(A)} \cdot \delta b \leq \delta x \leq \text{cond}\,(A) \cdot \delta b.$$

Biz.:

1. $A(x + \Delta x) = (b + \Delta b)$-ből vonjuk ki az $Ax = b$ LER-t, így $A\Delta x = \Delta b$.
2. Viszont $x = A^{-1}b$ és $\Delta x = A^{-1}\Delta b$ is teljesül.

Biz. (folytatás):

3. Tehát a 4-féle alak:

$$b = Ax, \;\; x = A^{-1}b, \;\; \Delta b = A\Delta x, \;\; \Delta x = A^{-1}\Delta b.$$

4. Bármely egyenlőségnél vehetjük a normát.
(A vektornormához illeszkedő mátrixnormát használunk.)
   (a) $\|b\| = \|Ax\| \quad \Rightarrow \quad \|b\| \leq \|A\| \cdot \|x\| \quad \Rightarrow \quad \|x\| \geq \frac{\|b\|}{\|A\|},$
   (b) $\|\Delta b\| = \|A\Delta x\| \;\; \Rightarrow \;\; \|\Delta b\| \leq \|A\| \cdot \|\Delta x\| \;\; \Rightarrow \;\; \|\Delta x\| \geq \frac{\|\Delta b\|}{\|A\|},$
   (c) $\|x\| = \left\|A^{-1}b\right\| \quad \Rightarrow \quad \|x\| \leq \left\|A^{-1}\right\| \cdot \|b\|,$
   (d) $\|\Delta x\| = \left\|A^{-1}\Delta b\right\| \quad \Rightarrow \quad \|\Delta x\| \leq \left\|A^{-1}\right\| \cdot \|\Delta b\|.$

5. Az alsó becslés (b) és (c) alapján:

$$\frac{\|\Delta x\|}{\|x\|} \geq \frac{\frac{\|\Delta b\|}{\|A\|}}{\|A^{-1}\| \cdot \|b\|} = \frac{1}{\|A\| \cdot \|A^{-1}\|} \cdot \frac{\|\Delta b\|}{\|b\|}.$$

Biz. (folytatás):

6. A felső becslés (a) $\|x\| \geq \frac{\|b\|}{\|A\|}$ és (d) $\|\Delta x\| \leq \left\|A^{-1}\right\| \cdot \|\Delta b\|$ alapján:

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\left\|A^{-1}\right\| \cdot \|\Delta b\|}{\frac{\|b\|}{\|A\|}} = \|A\| \cdot \left\|A^{-1}\right\| \cdot \frac{\|\Delta b\|}{\|b\|}.$$

$\square$

$$A \cdot x = b$$

Vizsgáljuk meg, hogy hogyan változik meg a LER megoldása, ha a bal oldalt, azaz a mátrixot *kicsit* megváltoztatjuk, „perturbáljuk”! (Mérési pontatlanság, kerekítési hiba, ...)

1. **Eredeti:**
   adott A és b, kiszámíthatjuk a megoldást: x.
   $Ax = b$
2. **Módosult:**
   adott $A + \Delta A$ és b, kiszámíthatjuk a megoldást: $x + \Delta x$.
   $(A + \Delta A)(x + \Delta x) = b$

Nyilván a megoldás is *kicsit* más lesz...

Példa:

1. **Eredeti:**

$$\begin{bmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{bmatrix} \cdot x = \begin{bmatrix} 4.1 \\ 9.7 \end{bmatrix} \quad \rightarrow \quad \text{megoldás: } x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

2. **Módosult:**

$$\begin{bmatrix} 4.11 & 2.8 \\ 9.7 & 6.6 \end{bmatrix} \cdot (x + \Delta x) = \begin{bmatrix} 4.1 \\ 9.7 \end{bmatrix}$$

3. $$\text{A módosult LER megoldása: } \quad x + \Delta x = \begin{bmatrix} 2.94 \\ -2.85 \end{bmatrix}$$

4. Mi történt?

Hogyan jellemezhető a megoldás megváltozása a jobb oldal megváltozásához képest?

- Mennyire változott a mátrix: $\delta A := \frac{\|\Delta A\|}{\|A\|} = 7.8495e - 004.$
- Emiatt mennyire változik a megoldás: $\delta x := \frac{\|\Delta x\|}{\|x\|} = 3.4507.$
- Vizsgáljuk a kettő hányadosát: $\frac{\delta x}{\delta A} = 4396.1.$
- $\text{cond}\,(A) = 1623$

Tétel: LER érzékenysége a mátrix pontatlanságára

Ha A invertálható, $b \neq 0$ és $\|\Delta A\| \cdot \|A^{-1}\| < 1$, akkor indukált mátrixnormában

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A\| \cdot \|A^{-1}\|}{1 - \|\Delta A\| \cdot \|A^{-1}\|} \cdot \frac{\|\Delta A\|}{\|A\|}.$$

Lemma

Ha $\|M\| < 1$, akkor $(I + M)$ invertálható és indukált mátrixnormában

$$\left\|(I + M)^{-1}\right\| \leq \frac{1}{1 - \|M\|}.$$

Megj: A lemmához kell az indukált mátrixnorma.

Biz. lemma:

- Az $I+M$ mátrix tényleg invertálható, hiszen $\varrho(M) \leq \|M\| < 1$, azaz M sajátértékeire: $|\lambda_i(M)| < 1$, vagyis az egységsugarú körön belül helyezkednek el. Meggondolható, hogy $I+M$ sajátvektorai ugyanazok, mint M sajátvektorai, a sajátértékekre pedig $\lambda_i(I+M) = 1 + \lambda_i(M)$ teljesül, így $I+M$ minden sajátértéke pozitív, következésképpen $I+M$ invertálható.
- Vizsgáljuk most $I+M$ inverzét, majd ennek normáját.

$$(I+M)^{-1} = I \cdot (I+M)^{-1} = (I+M-M)(I+M)^{-1} =$$
$$= I - M \cdot (I+M)^{-1},$$
$$\left\|(I+M)^{-1}\right\| \leq \|I\| + \|M\| \cdot \left\|(I+M)^{-1}\right\|,$$
$$(1-\|M\|) \cdot \left\|(I+M)^{-1}\right\| \leq \|I\| = 1 \;\;\Rightarrow\;\; \left\|(I+M)^{-1}\right\| \leq \frac{1}{1-\|M\|}.$$

$\square$

Biz. tétel: Az $(A+\Delta A)(x+\Delta x) = b$ LER-ből $Ax = b$-t kivonva $(A+\Delta A)\cdot\Delta x + \Delta A\cdot x = 0$, másképp

$$\begin{aligned}(A+\Delta A)\cdot\Delta x &= -\Delta A\cdot x,\\ A\cdot(I + A^{-1}\cdot\Delta A)\cdot\Delta x &= -\Delta A\cdot x.\end{aligned}$$

Mivel feltevésünk szerint $\left\|A^{-1}\cdot\Delta A\right\| \le \left\|A^{-1}\right\|\cdot\|\Delta A\| < 1$, a lemma alapján mondhatjuk, hogy $(I + A^{-1}\cdot\Delta A)$ invertálható.

$$\Delta x = -(I + A^{-1}\cdot\Delta A)^{-1}A^{-1}\Delta A\cdot x$$

Az inverz normájára adott becslésünket is felhasználva:

$$\|\Delta x\| \le \left\|(I + A^{-1}\cdot\Delta A)^{-1}\right\|\cdot\left\|A^{-1}\right\|\cdot\|\Delta A\|\cdot\|x\|$$

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{1}{1-\|A^{-1}\cdot\Delta A\|}\cdot\left\|A^{-1}\right\|\cdot\|\Delta A\| \le \frac{\|A\|\cdot\|A^{-1}\|}{1-\|A^{-1}\|\cdot\|\Delta A\|}\cdot\frac{\|\Delta A\|}{\|A\|}.$$

$\square$

Tétel átfogalmazás:

$$\begin{aligned}
&\frac{\|A\| \cdot \left\|A^{-1}\right\|}{1 - \|\Delta A\| \cdot \|A^{-1}\|} \cdot \frac{\|\Delta A\|}{\|A\|} = \\
&= \frac{\|A\| \cdot \left\|A^{-1}\right\|}{1 - \frac{\|\Delta A\|}{\|A\|} \cdot \|A\| \cdot \|A^{-1}\|} \cdot \frac{\|\Delta A\|}{\|A\|} = \\
&= \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A) \cdot \frac{\|\Delta A\|}{\|A\|}} \cdot \frac{\|\Delta A\|}{\|A\|}.
\end{aligned}$$

Tartalomjegyzék

Megjegyzés: egyesített tétel LER érzékenységéről

Ha az

$$A \cdot x = b$$

LER esetén mind a bal oldal mátrixa, mind a jobb oldal vektora megváltozik, és az így számolt megoldásra

$$(A + \Delta A) \cdot (x + \Delta x) = b + \Delta b$$

teljesül, akkor a következő becslés igazolható:

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\mathrm{cond}\,(A)}{1 - \mathrm{cond}\,(A) \cdot \frac{\|\Delta A\|}{\|A\|}} \cdot \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right).$$

Példa

Hogyan befolyásolja az LU-felbontás a feladat kondicionáltságát? Mutassuk meg, hogy nem javul.

Biz.:

- $Ax = b \quad \Rightarrow \quad LUx = b \quad \Rightarrow \quad Ly = b, Ux = y,$
- $A = L \cdot U \quad \Rightarrow \quad \|A\| \leq \|L\| \cdot \|U\|$
- $A^{-1} = U^{-1} \cdot L^{-1} \quad \Rightarrow \quad \|A^{-1}\| \leq \|L^{-1}\| \cdot \|U^{-1}\|$
- $\mathrm{cond}\,(A) \leq \mathrm{cond}\,(L) \cdot \mathrm{cond}\,(U)$ $\square$

Sőt előfordulhat, hogy $\mathrm{cond}\,(L), \mathrm{cond}\,(U) >> \mathrm{cond}\,(A)$, azaz bizonyos mátrixok esetén előfordulhat, hogy a Gauss-elimináció nagyon pontatlan eredményt ad.

Példa gyakorlatra

Igazoljuk, hogy a QR-felbontással a feladat kondicionáltsága nem változik.

Példa gyakorlatra

Igazoljuk, hogy a Cholesky-felbontással a feladat kondicionáltsága nem változik.

Ez is mutatja a QR- és Cholesky-felbontáson alapuló módszerek stabilitását.

A kondíciószám, csak a LER megoldás (vagyis a feladat) érzékenységét jellemzi, a megoldó algoritmusét nem. A megoldó módszer jellemzésére a maradékvektort használjuk.

Definíció: reziduum- vagy maradékvektor

Legyen $\widetilde{x}$ az $Ax = b$ LER egy közelítő megoldása. Ekkor az $r := b - A\widetilde{x}$ vektort **reziduum-** vagy **maradékvektornak** nevezzük.

Látjuk, hogy a reziduum vektor könnyen számolható, alkalmazható direkt- és iterációs módszerek esetén is. Az utóbbi esetben leállási feltétel is készíthető a segítségével.

Relatív maradék

Definíció: relatív maradék

- Az $\eta := \frac{\|r\|}{\|A\| \cdot \|\widetilde{x}\|}$ ([éta]) mennyiséget **relatív maradéknak** nevezzük.
- A stabilitás inverz megfogalmazása alapján a módszer stabil, ha az $\widetilde{x}$ közelítő megoldáshoz tartozó $(A + \Delta A) \cdot \widetilde{x} = b$ LER csak kicsit perturbált az eredetihez képest, azaz $\frac{\|\Delta A\|}{\|A\|}$ kicsi.

η értéke a közelítő megoldás ismeretében könnyen számolható.
A továbbiakban ΔA ismerete nélkül szeretnénk becsléseket adni a nem ismert $\frac{\|\Delta A\|}{\|A\|}$ mennyiségre.

Tétel: becslés a relatív maradékra

Ha A invertálható, akkor illeszkedő mátrixnormában

$$\eta \leq \frac{\|\Delta A\|}{\|A\|},$$

azaz ha η nagy, akkor $\frac{\|\Delta A\|}{\|A\|}$ is nagy.

Biz.: $b = (A + \Delta A) \cdot \widetilde{x} = A \cdot \widetilde{x} + \Delta A \cdot \widetilde{x}$, innen
$b - A \cdot \widetilde{x} = r = \Delta A \cdot \widetilde{x}$, a mátrixnorma illeszkedését felhasználva

$$\|r\| \leq \|\Delta A\| \cdot \|\widetilde{x}\| \, .$$

A relatív maradékot becsülve

$$\eta = \frac{\|r\|}{\|A\| \cdot \|\widetilde{x}\|} \leq \frac{\|\Delta A\| \cdot \|\widetilde{x}\|}{\|A\| \cdot \|\widetilde{x}\|} \leq \frac{\|\Delta A\|}{\|A\|}$$

Tétel: relatív maradék 2-es normában

Ha A invertálható, akkor

$$\eta_2 = \frac{\|\Delta A\|_2}{\|A\|_2}.$$

Biz.: Belátjuk, hogy

$$\Delta A = \frac{r\widetilde{x}^\top}{\widetilde{x}^\top\widetilde{x}}$$

jó lesz perturbációnak, vagyis $\widetilde{x}$ egy ennyivel megváltoztatott mátrixú LER pontos megoldása. Végezzük el a behelyettesítést:

$$(A + \Delta A)\cdot\widetilde{x} = \left(A + \frac{r\widetilde{x}^\top}{\widetilde{x}^\top\widetilde{x}}\right)\cdot\widetilde{x} =$$

$$= A\widetilde{x} + \frac{r\widetilde{x}^\top\widetilde{x}}{\widetilde{x}^\top\widetilde{x}} = A\widetilde{x} + (b - A\widetilde{x}) = b.$$

Biz.: folyt. Felhasználjuk, hogy

$$\left\| r\widetilde{x}^\top \right\|_2 = \|r\|_2 \cdot \|\widetilde{x}\|_2 \,.$$

(Beadható HF-nak kitűzött feladat.)

A relatív maradékot becsülve

$$\frac{\|\Delta A\|_2}{\|A\|_2} = \frac{\left\| r\widetilde{x}^\top \right\|_2}{\|A\|_2 \, \|\widetilde{x}\|_2^2} = \frac{\|r\|_2 \, \|\widetilde{x}\|_2}{\|A\|_2 \, \|\widetilde{x}\|_2^2} = \frac{\|r\|_2}{\|A\|_2 \, \|\widetilde{x}\|_2} = \eta_2.$$

$\square$

Ha η_2 kicsi, akkor $\frac{\|\Delta A\|_2}{\|A\|_2}$ is kicsi.
Ha $\eta_2 < \varepsilon_1$, akkor ebben az adott aritmetikában pontosabb megoldás nem adható.

Példák Matlab-ban

1. Egy perturbált LER (jobboldala változik, mátrixa a Hilbert mátrix).
2. $\operatorname{cond}_2(H_n)$ változása a méret függvényében.
3. $\operatorname{cond}_2(V_n)$ változása a méret függvényében.
4. $\operatorname{cond}_2(\operatorname{tridiag}(-1, 2, -1))$ változása a méret függvényében.
5. $\operatorname{cond}_2(rand_n)$ változása a méret függvényében.

Példa:

Jelöljük H_5-tel az 5×5-ös Hilbert mátrixot.

$$H_5 = \left(\frac{1}{i+j-1}\right)_{i,j=1}^{5} = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 \\ 1/3 & 1/4 & 1/5 & 1/6 & 1/7 \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 \end{bmatrix}$$

1. Példa:

1. **Eredeti LER:**

$$H_5 \cdot x = \begin{bmatrix} 1/5 \\ 1/6 \\ 1/7 \\ 1/8 \\ 1/9 \end{bmatrix} \quad \rightarrow \quad \text{megoldás: } x = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

2. **Módosult LER:**

$$H_5 \cdot (x + \Delta x) = \begin{bmatrix} 1/5 \\ 1/6 \\ 1/7 \\ 1/8 \\ 1/9 + 1/1000 \end{bmatrix}$$

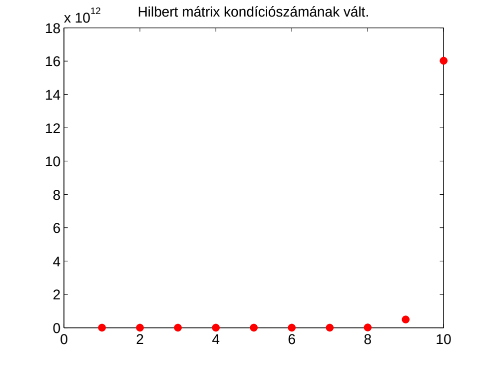
A módosult LER megoldása: $x + \Delta x = \begin{bmatrix} 0.6300 \\ -12.6000 \\ 56.7000 \\ -88.2000 \\ 45.1000 \end{bmatrix}$

Mi történt?

1. $\delta b = 0.0029$: a jobboldal relatív hibája
2. $\delta x = 114.4469$ a megoldás relatív hibája
3. a két mennyiség hányadosa: $\delta x / \delta b = 3.9006e + 004$
4. ennek becslése a tétellel: $cond_2(H_5) = 4.7661e + 005$.

2. Példa:

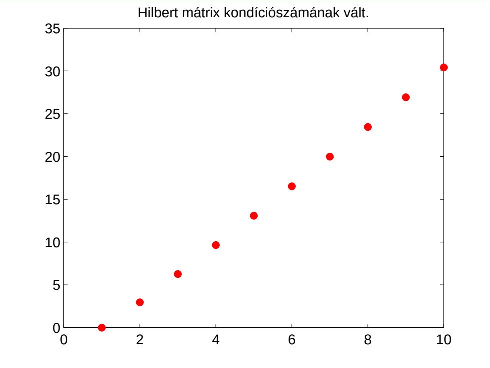
A Hilbert mátrix kondíciószámának változását vizsgáljuk:



Nem sok látszik az ábrából, mintha csak az utolsó érték lenne nagy.

2. Példa:

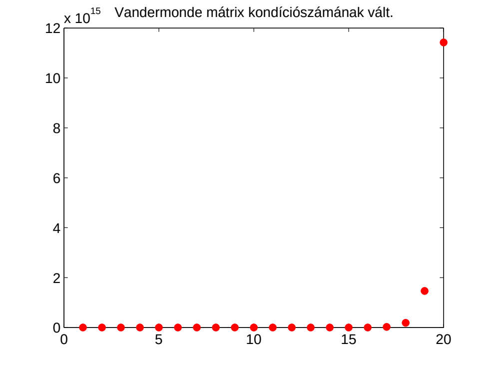
Vegyük a kondíciószámok logaritmusát!



$$\mathrm{cond}_2(H_n) \approx \exp(3.1n) \approx 22^n$$

3. Példa:

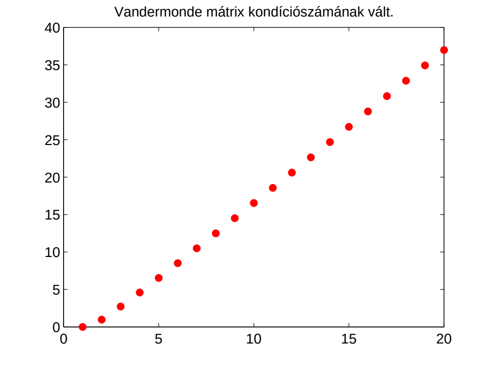
A $[0,1]$ intervallum egyenletes felosztású pontjaiból képzett Vandermonde mátrix kondíciószámának változását vizsgáljuk:



Nem sok látszik az ábrából, mintha csak az utolsó érték lenne nagy.

3. Példa:

Vegyük a kondíciószámok logaritmusát!

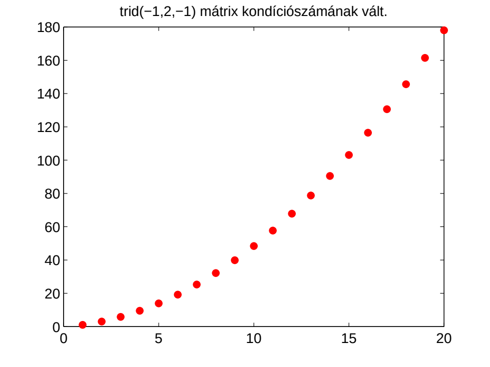


$$\mathrm{cond}_2(V_n) \approx \exp(1.85n) \approx (6.4)^n$$

A tridiag $(-1, 2, -1)$ mátrix kondíciószáma

4. Példa:

A tridiag $(-1, 2, -1)$ mátrix kondíciószámának változását vizsgáljuk:

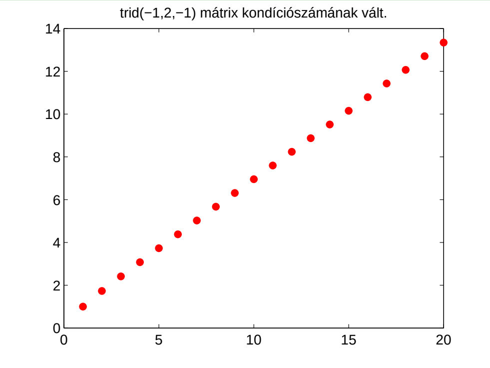


Az ábra alapján sejthető, hogy a növekedés a méret négyzetével arányos.

A tridiag $(-1, 2, -1)$ mátrix kondíciószáma

4. Példa:

Vegyük a kondíciószámok gyökét!

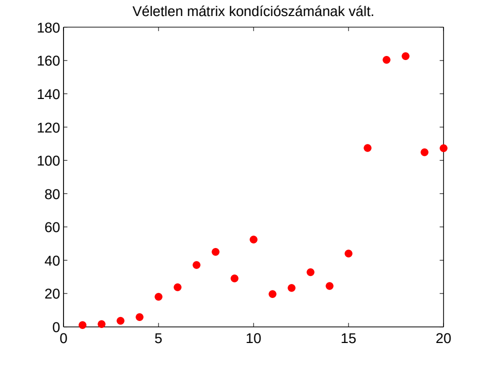


Elméletileg igazolható, hogy

$\mathrm{cond}_2(\mathrm{tridiag}\,(-1, 2, -1)) \approx \left(\frac{2(n+1)}{\pi}\right)^2.$

5. Példa:

Véletlen mátrix kondíciószámának változását vizsgáljuk:



Az előző mátrixokhoz képest egész kicsi értékeket kaptunk.

Numerikus módszerek 1.

8. előadás: Iterációs módszerek LER megoldására, Jacobi- és csillapított Jacobi-iteráció

Krebsz Anna

ELTE IK

Tartalomjegyzék

Tekintsük a következő leképezést:

$$\varphi\colon \mathbb{R}^n \to \mathbb{R}^n, \quad \varphi(x) = Bx + c,$$

ahol a $B \in \mathbb{R}^{n\times n}$ mátrixot *átmenet mátrixnak* nevezik és $c \in \mathbb{R}^n$,

majd ennek segítségével képezzük a következő (vektor)sorozatot, *iterációt*:

$$x^{(0)} \in \mathbb{R}^n \quad \text{(tetszőleges)}, \qquad x^{(k+1)} = \varphi(x^{(k)}) \quad (k = 0, 1, 2, \dots).$$

Példa

Egyszerűen számolhatók a következő sorozat elemei!

$$x^{(0)} := \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad x^{(k+1)} := \frac{1}{5}\begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix} \cdot x^{(k)} + \frac{1}{5}\begin{bmatrix} 7 \\ -1 \end{bmatrix}, \quad (k \in \mathbb{N}_0).$$

Kérdések: Mit tud ez a sorozat / iteráció? Konvergens? Milyen értelemben? Ha konvergens, mi a határértéke?
A választ majd a fixponttétel adja meg.

Eml.:

Definíció: vektorsorozat konvergenciája, határértéke

Az $\left(x^{(k)}|k\in\mathbb{N}\right)\subset\mathbb{R}^n$ vektorsorozat *konvergens* a $\|.\|$ vektornormában, ha $\exists\, x^*\in\mathbb{R}^n$, melyre

$$\forall\,\varepsilon>0:\exists\,N\in\mathbb{N}:\forall\,k>N:\left\|x^{(k)}-x^*\right\|<\varepsilon.$$

Ekkor a sorozat *határértéke* x^*, azaz $\lim\limits_{k\to\infty}x^{(k)}=x^*$.

Mi köze ennek lineáris egyenletrendszerekhez?

Ha folytonos φ függvény és $\lim\limits_{k\to\infty} x^{(k)} = x^*$, akkor a folytonosságra vonatkozó átviteli elvből

$$\varphi(x^*) = \lim_{k\to\infty} \varphi(x^{(k)}) = \lim_{k\to\infty} x^{(k+1)} = x^*.$$

A korábban megadott φ-vel $x^* = B \cdot x^* + c$.

Vagyis $(I - B) \cdot x^* = c$, azaz x^* az $(I - B) \cdot x = c$ LER megoldása.

Alkalmazzuk az $A = I - B,\ b = c,\ Ax = b$ jelölést...

Fordítva: Adott $Ax = b$ LER esetén keressünk vele ekvivalens $Bx + c = x$ egyenletet. Ebből felírhatunk egy iterációt:

$$x^{(k+1)} = Bx^{(k)} + c.$$

Hogyan írhatjuk át a megadott alakba?

Általában:

$$Ax = b, \qquad A = P + Q, \qquad (P + Q)x = b,$$

átrendezve:

$$Px = -Qx + b \quad \Longleftrightarrow \quad x = -P^{-1}Qx + P^{-1}b,$$

iterációs alakban írva:

$$x^{(k+1)} = \underbrace{-P^{-1}Q}_{B} \cdot x^{(k)} + \underbrace{P^{-1}b}_{c}.$$

Definíció: fixpont

Az $x^* \in \mathbb{R}^n$ pontot (vektort) a $\varphi\colon \mathbb{R}^n \to \mathbb{R}^n$ leképezés *fixpontjának* nevezzük, ha $x^* = \varphi(x^*)$.
Az $x = \varphi(x)$ egyenletet *fixpontegyenletnek* nevezzük.

Definíció: kontrakció

A $\varphi\colon \mathbb{R}^n \to \mathbb{R}^n$ leképezés *kontrakció*, ha $\exists\, q \in [0,1)$, hogy

$$\|\varphi(x) - \varphi(y)\| \leq q \cdot \|x - y\|\,, \qquad \forall\, x, y \in \mathbb{R}^n.$$

Megj.:

- kontrakció $\approx$ összehúzás
- q: kontrakciós együttható

Állítás

Ha $\|B\| < 1$, akkor a $\varphi\colon \mathbb{R}^n \to \mathbb{R}^n$, $\varphi(x) = B \cdot x + c$ leképezés kontrakció. (Az $\mathbb{R}^n$-en alkalmazott vektornormához illeszkedő mátrixnormát tekintve.)

Biz.:

$$\|\varphi(x) - \varphi(y)\| = \|(Bx + c) - (By + c)\| =$$
$$= \|Bx - By\| = \|B(x - y)\| \le \underbrace{\|B\|}_{:=q<1} \cdot \|x - y\| \,.$$

Tétel: Banach-féle fixponttétel $\mathbb{R}^n$-re

Ha a $\varphi\colon \mathbb{R}^n \to \mathbb{R}^n$ függvény kontrakció $\mathbb{R}^n$-en q kontrakciós együtthatóval, akkor

1. $\exists\, x^* \in \mathbb{R}^n : x^* = \varphi(x^*)$, azaz létezik fixpont,
2. a fixpont egyértelmű,
3. $\forall x^{(0)} \in \mathbb{R}^n$ esetén az $x^{(k+1)} = \varphi(x^{(k)})$, $(k \in \mathbb{N}_0)$ sorozat konvergens és $\lim\limits_{k\to\infty} x^{(k)} = x^*$,
4. továbbá a következő hibabecslések teljesülnek:
   - $\left\|x^{(k)} - x^*\right\| \leq q^k \cdot \left\|x^{(0)} - x^*\right\|$,
   - $\left\|x^{(k)} - x^*\right\| \leq \dfrac{q^k}{1-q} \cdot \left\|x^{(1)} - x^{(0)}\right\|$.

Biz.:

(a) A φ leképezés kontrakció voltából következik, hogy φ **folytonos** (sőt egyenletesen folytonos) is, ugyanis $\forall\, \varepsilon > 0$-hoz válasszuk $\delta = \varepsilon / q$-t. Ekkor ha $\|x - y\| < \delta$, akkor

$$\|\varphi(x) - \varphi(y)\| \leq q \cdot \|x - y\| < q \cdot \frac{\varepsilon}{q} = \varepsilon.$$

(b) Belátjuk, hogy a tételben definiált $(x^{(k)})$ **Cauchy-sorozat,** így konvergens. Elsőként egymást követő tagok eltérését becsüljük:

$$\begin{aligned} \left\| x^{(k+1)} - x^{(k)} \right\| &= \left\| \varphi(x^{(k)}) - \varphi(x^{(k-1)}) \right\| \leq \\ &\leq q \cdot \left\| x^{(k)} - x^{(k-1)} \right\| \leq \\ &\leq \ldots \leq q^k \cdot \left\| x^{(1)} - x^{(0)} \right\|. \end{aligned}$$

Biz. folyt.:

(c) Legyen $m \in \mathbb{N}$, $m \geq 1$, vizsgáljuk meg két m távolságra lévő tag különbségét! A háromszög-egyenlőtlenséget és a mértani sor összegképletét is felhasználva:

$$\begin{aligned}
\left\| x^{(k+m)} - x^{(k)} \right\| &= \left\| (x^{(k+m)} - x^{(k+m-1)}) + \ldots + (x^{(k+1)} - x^{(k)}) \right\| \leq \\
&\leq \left\| x^{(k+m)} - x^{(k+m-1)} \right\| + \ldots + \left\| x^{(k+1)} - x^{(k)} \right\| \leq \\
&\leq (q^{m+k-1} + \ldots + q^k) \cdot \left\| x^{(1)} - x^{(0)} \right\| = \\
&= q^k \cdot (q^{m-1} + \ldots + 1) \cdot \left\| x^{(1)} - x^{(0)} \right\| < \\
&< \frac{q^k}{1-q} \cdot \left\| x^{(1)} - x^{(0)} \right\| .
\end{aligned}$$

Mivel $k \to \infty$ esetén $(q^k) \to 0$, ezért $(x^{(k)})$ Cauchy-sorozat,

Biz. folyt.:

(d) Minden $\mathbb{R}^n$-beli Cauchy-sorozat konvergens, így $(x^{(k)})$ konvergens, $x^* := \lim(x^{(k)})$. φ folytonosságából az átviteli elv értelmében

$$\varphi(x^*) = \lim \varphi(x^{(k)}) = \lim x^{(k+1)} = x^*,$$

azaz x^* **fixpontja** φ-nek.

(e) Az **egyértelműség** belátásához indirekt tegyük fel, hogy létezik legalább két $x^* \neq x^{**}$ fixpont. Ekkor

$$\|x^* - x^{**}\| = \|\varphi(x^*) - \varphi(x^{**})\| \leq q \cdot \|x^* - x^{**}\|.$$

Átrendezve $\qquad \|x^* - x^{**}\| (1-q) \leq 0.$

Tehát $\|x^* - x^{**}\| = 0$, vagyis $x^* = x^{**}$ következik. Ellentmondás!

(f) A hibabecsléshez vizsgáljuk először a k-adik tag hibáját:

$$\left\|x^{(k)} - x^*\right\| = \left\|\varphi(x^{(k-1)}) - \varphi(x^*)\right\| \leq q \cdot \left\|x^{(k-1)} - x^*\right\| \leq \ldots \leq \\ \leq q^k \cdot \left\|x^{(0)} - x^*\right\|.$$

Valamint a korábbi képletben:

$$\left\|x^{(k+m)} - x^{(k)}\right\| < \frac{q^k}{1-q} \cdot \left\|x^{(1)} - x^{(0)}\right\|$$

$m \to \infty$ esetén felhasználva, hogy a vektornorma folytonos függvény

$$\left\|x^* - x^{(k)}\right\| \leq \frac{q^k}{1-q} \left\|x^{(1)} - x^{(0)}\right\|.$$

$\square$

Következmény: iteráció konvergenciájának elégséges feltétele

Ha $\|B\| < 1$, az $x^{(k+1)} = B \cdot x^{(k)} + c$ iteráció konvergens minden kezdőértékre.

Megj.: Attól még lehet konvergens valamely kezdőértékből indítva, ha $\|B\| \geq 1$.
(Nem szükséges feltétel.)

Lemma: spektrálsugár és az indukált normák kapcsolata

$$\varrho(B) = \inf \{ \, \|B\| \ : \ \|.\| \ \text{ indukált mátrixnorma} \, \} \, ,$$

azaz $\forall\, \varepsilon > 0 : \exists$ indukált $\|.\| : \|B\| < \varrho(B) + \varepsilon$.

Biz.: Nélkül.

Tétel: iteráció konvergenciájának ekvivalens feltétele

Az $x^{(k+1)} = B \cdot x^{(k)} + c$ iteráció akkor és csak akkor konvergens minden kezdőértékre, ha

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

Biz.:

- $\Leftarrow$: Az előző Lemma alapján trivi.
- $\Rightarrow$: Indirekt tegyük fel, hogy $\varrho(B) \geq 1$, azaz $\exists \, |\lambda| \geq 1$ sajátérték, és legyen $x^{(0)}$ olyan, hogy $x^{(0)} - x^*(\neq 0)$ kezdeti hiba a B λ-hoz tartozó sajátvektora legyen.

Ekkor:

$$B(x^{(0)} - x^*) = \lambda(x^{(0)} - x^*)$$
$$B^2(x^{(0)} - x^*) = \lambda^2(x^{(0)} - x^*) \quad \Rightarrow \ldots$$
$$B^k(x^{(0)} - x^*) = \lambda^k(x^{(0)} - x^*) \; (k \in \mathbb{N})$$
$$x^{(k)} - x^* = (Bx^{(k-1)} + c) - (Bx^* + c) = B(x^{(k-1)} - x^*) =$$
$$= B^k(x^{(0)} - x^*) = \lambda^k(x^{(0)} - x^*)$$
$$\left\| x^{(k)} - x^* \right\| = |\lambda|^k \cdot \underbrace{\left\| x^{(0)} - x^* \right\|}_{\text{konst.}} \nrightarrow 0 \quad (k \to \infty)$$

Ellentmondásra jutottunk. $\square$

Megj.: Az iteráció futtatása során nem áll rendelkezésünkre kontrakciós együttható, annak kiszámítása elméleti feladat. Ehelyett ún. tapasztalati kontrakciós együtthatóval dolgozunk.

1. Láttuk a fixponttétel bizonyításában, hogy $\left\|x^{(k+1)} - x^{(k)}\right\| \leq q \cdot \left\|x^{(k)} - x^{(k-1)}\right\|$, innen

$$q^{(k)} \approx \frac{\left\|x^{(k+1)} - x^{(k)}\right\|}{\left\|x^{(k)} - x^{(k-1)}\right\|}$$

   a k. lépésbeli tapasztalati kontrakciós együtthatónk.

2. Ennek ismeretében a hibabecslés alakja:

$$\left\|x^{(k)} - x^*\right\| \leq \frac{q^{(k)}}{1 - q^{(k)}} \left\|x^{(k)} - x^{(k-1)}\right\|.$$

   Tehát menet közben ellenőrizni tudjuk, hogy elegendő-e a pontosság.

3. Ha $|q^{(k)}| > 1$ az első néhány lépés után, akkor leállíthatjuk az iterációt divergencia miatt.
4. Vannak esetek, amikor a $(q^{(k)})$ sorozat nem monoton, ekkor érdemes $q^{(k)}$ helyett a $q \approx \sqrt{q^{(k)} q^{(k-1)}}$ mértani középpel dolgozni.
5. A fenti segítséggel „inteligens" iterációs módszer programot írhatunk, mely a sorozat elemeiből a hibabecslést elő tudja állítani és divergencia esetén sem számol feleslegesen sokat.

Példa

Mit állíthatunk a következő iteráció konvergenciájáról?

$$x^{(k+1)} := \frac{1}{5}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \cdot x^{(k)} + \frac{1}{7}\begin{bmatrix} 32.4 \\ \sqrt{\pi} \end{bmatrix}, \quad (k \in \mathbb{N}_0).$$

Mivel $\|B\|_1 = \frac{3}{5} = q$ a kontrakciós együttható, ezért az iteráció bármely $x^{(0)} \in \mathbb{R}^2$ kezdőértékre konvergens. Hibabecslést az 1-es vektornormában írhatnánk fel.

Tekintsük az $Ax = b$ lineáris egyenletrendszert, majd írjuk fel annak mátrixát

$$A = L + D + U$$

alakban, ahol L alsó háromszögmátrix, D diagonális mátrix, U pedig felső háromszögmátrix, méghozzá

- $l_{ij} = a_{ij} \quad (i < j)$,
- $d_{ij} = a_{ij} \quad (i = j)$,
- $u_{ij} = a_{ij} \quad (i > j)$.

Az elemek L, D, U mátrixokba pakolásáról van szó. A továbbiakban tegyük fel, hogy A diagonális elemei nem nullák. Ha mégis az lenne, cseréljük meg a LER-ben a sorokat, hogy teljesítse a feltételt.

Példa:

Példa $A = L + D + U$ felbontásra:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 4 & 0 & 0 \\ 7 & 8 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix} + \begin{bmatrix} 0 & 2 & 3 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{bmatrix}.$$

Megj.: Semmi köze az LU-felbontáshoz.

Tartalomjegyzék

Átalakítás:

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

Ezek alapján az iteráció a következő.

Definíció: Jacobi-iteráció

$$
x^{(k+1)} = \underbrace{-D^{-1}(L + U)}_{B_J} \cdot x^{(k)} + \underbrace{D^{-1}b}_{c_J} = B_J \cdot x^{(k)} + c_J
$$

Eml.:

$$x^{(k+1)} = -D^{-1}(L+U)\cdot x^{(k)} + D^{-1}b$$

Írjuk fel koordinátánként (komponensenként)!

Állítás: a Jacobi-iteráció komponensenkénti alakja

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

Biz.: Házi feladat meggondolni. Egyszerű. □

Írjuk fel az iteráció reziduum vektoros alakját!

$$x^{(k+1)} = -D^{-1}(L+U)\cdot x^{(k)} + D^{-1}b = D^{-1}\Big((D-A)\cdot x^{(k)} + b\Big) =$$
$$= x^{(k)} + D^{-1}\Big(-Ax^{(k)} + b\Big) = x^{(k)} + D^{-1}r^{(k)}$$

Vezessük be az $s^{(k)} := D^{-1}r^{(k)}$ segédvektort, ezzel egy lépésünk alakja:

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

Az új reziduum vektor:

$$r^{(k+1)} = b - Ax^{(k+1)} = b - A(x^{(k)} + s^{(k)}) = r^{(k)} - As^{(k)}.$$

Algoritmus: Jacobi-iteráció

$$r^{(0)} := b - Ax^{(0)}$$

$k = 1, \ldots,$ leállásig

$$s^{(k)} := D^{-1} r^{(k)} \quad \Leftrightarrow \quad Ds^{(k)} = r^{(k)} \ \text{LER}$$

$$x^{(k+1)} := x^{(k)} + s^{(k)}$$

$$r^{(k+1)} := r^{(k)} - As^{(k)}$$

Megj.: Látjuk, hogy $x^{(k+1)} - x^{(k)} = s^{(k)}$, vagyis a tapasztalati kontrakciós együtthatók számításához lépésenként egy norma értéket és egy osztást kell elvégezni.

Tétel

Ha A szig. diag. dom. a soraira, akkor az $Ax = b$ LER-re felírt Jacobi-iteráció konvergens bármely $x^{(0)}$ esetén.

Biz.: Írjuk fel a B_J mátrix elemeit: $b_{ii} = 0$ és $i \neq j$-re $b_{ij} = -\frac{a_{ij}}{a_{ii}}$.

$$\|B_J\|_\infty = \left\|-D^{-1}(L+U)\right\|_\infty = \max_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|}$$

Ha A szig. diag. dom. a soraira, akkor

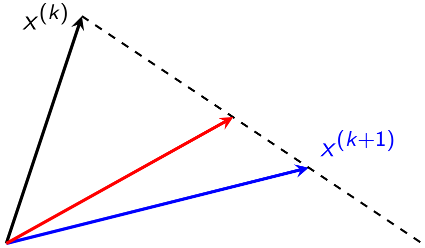
$$\forall\, i: \quad |a_{ii}| > \sum_{j=1,\, j\neq i}^{n} |a_{ij}| \quad \Leftrightarrow \quad 1 > \sum_{j=1,\, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|}.$$

Tehát minden összeg egynél kisebb, így a maximumuk is, ezzel az elégséges feltétel miatt a konvergencia teljesül.

$$\|B_J\|_\infty = \max_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1$$

A *csillapítás* avagy *tompítás* alapötlete:

$$x_J^{(k+1)} \quad \text{helyett} \quad (1-\omega)\cdot x^{(k)} + \omega \cdot x_J^{(k+1)}$$



Megj.:

- alulrelaxálás ($0 < \omega < 1$), túlrelaxálás ($\omega > 1$)
- $\omega = 1$ az eredeti módszert adja

Induljunk a Jacobi-módszerből és a „helyben hagyásból":

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

A kettő súlyozott összege:

$$x = \big[(1-\omega)I - \omega D^{-1}(L+U)\big]\cdot x + \omega D^{-1}b$$

Ezek alapján az iteráció a következő.

Definíció: csillapított Jacobi-iteráció ω paraméterrel – $J(\omega)$

$$x^{(k+1)} = \underbrace{\Big[(1-\omega)I - \omega D^{-1}(L+U)\Big]}_{B_{J(\omega)}} \cdot x^{(k)} + \underbrace{\omega D^{-1}b}_{c_{J(\omega)}}$$

Írjuk fel koordinátánként!

Állítás: $J(\omega)$ komponensenkénti alakja

$$x_i^{(k+1)} = (1-\omega) \cdot x_i^{(k)} + \omega \cdot x_{i,J}^{(k+1)},$$

ahol $x_{i,J}^{(k+1)}$ a hagyományos Jacobi-módszer $(J = J(1))$ által adott, azaz

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

Biz.: Házi feladat meggondolni. Nem nehéz. $\square$

Írjuk fel az iteráció reziduum vektoros alakját!

$$\begin{aligned} x^{(k+1)} &= (1-\omega)x^{(k)} - \omega D^{-1}(L+U)\cdot x^{(k)} + \omega D^{-1}b = \\ &= (1-\omega)x^{(k)} + \omega D^{-1}\left((D-A)\cdot x^{(k)} + b\right) = \\ &= (1-\omega)x^{(k)} + \omega x^{(k)} + \omega D^{-1}\left(-Ax^{(k)} + b\right) = \\ &= x^{(k)} + \omega D^{-1} r^{(k)} \end{aligned}$$

Vezessük be az $s^{(k)} := \omega D^{-1} r^{(k)}$ segédvektort, ezzel egy lépésünk alakja:

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

Az új reziduum vektor:

$$r^{(k+1)} = b - Ax^{(k+1)} = b - A(x^{(k)} + s^{(k)}) = r^{(k)} - As^{(k)}.$$

Algoritmus: csillapított Jacobi-iteráció $J(\omega)$

$$r^{(0)} := b - Ax^{(0)}$$

$k = 1, \ldots,$ leállásig

$$s^{(k)} := \omega D^{-1} r^{(k)} \quad \Leftrightarrow \quad Ds^{(k)} = \omega r^{(k)} \ \text{LER}$$

$$x^{(k+1)} := x^{(k)} + s^{(k)}$$

$$r^{(k+1)} := r^{(k)} - As^{(k)}$$

Megj.: Látjuk, hogy $x^{(k+1)} - x^{(k)} = s^{(k)}$, vagyis a tapasztalati kontrakciós együtthatók számításához lépésenként egy norma értéket és egy osztást kell elvégezni.

Tétel a csillapított Jacobi-iteráció ($J(\omega)$) konvergenciája

Ha az $Ax = b$ LER-re a Jacobi-iteráció konvergens minden kezdőértékre, akkor $0 < \omega < 1$-re a csillapított Jacobi-iteráció is az.

Biz.: $J(\omega)$ iteráció esetén az átmenet mátrix $(1-\omega)I + \omega B_J$.
Először belátjuk, hogy a $B_{J(\omega)}$ mátrix μ_i sajátértékeire teljesül, hogy

$$\mu_i = (1-\omega) + \omega\lambda_i,$$

ahol λ_i-k a B_J sajátértékei. A két mátrix sajátvektorai (v_i-k) azonosak.

$$\begin{aligned} B_{J(\omega)} v_i &= ((1-\omega)I + \omega B_J)v_i = (1-\omega)v_i + \omega\lambda_i v_i = \\ &= \underbrace{((1-\omega) + \omega\lambda_i)}_{\mu_i} v_i = \mu_i v_i \quad (i = 1, \ldots, n) \end{aligned}$$

Biz. folyt: A bizonyításban a konvergenciára vonatkozó szükséges és elégséges feltételt használjuk. Belátjuk, hogy

$$\varrho(B_J) < 1 \quad \Rightarrow \quad 0 < \omega < 1: \;\; \varrho(B_{J(\omega)}) < 1.$$

$\varrho(B_J) < 1$-ből következik, hogy minden i-re $|\lambda_i| < 1$.

Felhasználjuk, hogy $0 < \omega < 1$ és becsüljük $\mu_i = (1-\omega) + \omega\lambda_i$-t:

$$|\mu_i| \leq (1-\omega) + \omega\,|\lambda_i| < (1-\omega) + \omega = 1 \quad (i = 1, \ldots, n).$$

Ha minden i-re $|\mu_i| < 1$ teljesül, akkor $\varrho(B_{J(\omega)}) < 1$, vagyis a csillapított iteráció minden kezdőértékre konvergens. $\square$

Példák Matlab-ban

1. Példa iterációra, konvergens vektorsorozat számítására.
2. Konvergens és divergens iterációk tulajdonságainak szemléltetése $n = 2, 3$ dimenzióban.
3. A tapasztalati kontrakciós együtthatók szemléltetése a csillapított Jacobi iteráció esetén.

1. Példa:

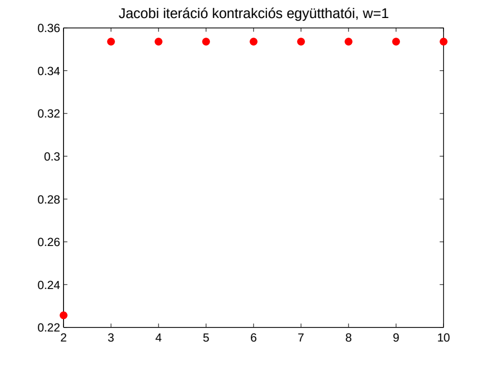
A LER alakja $Ax = b$, ahol

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

Vizsgáljuk a csillapított Jacobi iteráció tapasztalati kontrakciós együtthatóit $\omega = 1, 0.8, 0.6, 1.2, 1.8, -0.1$ esetén!

Tapasztalati kontrakciós együttható vizsgálata

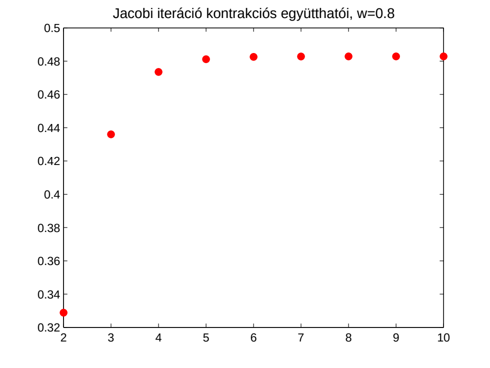
1. Példa:



$q \approx 0.3536$

Tapasztalati kontrakciós együttható vizsgálata

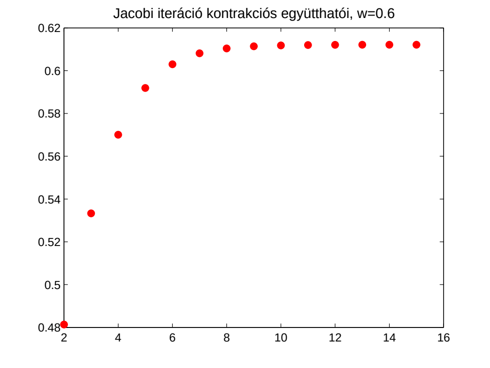
1. Példa:



$q \approx 0.4828$

Tapasztalati kontrakciós együttható vizsgálata

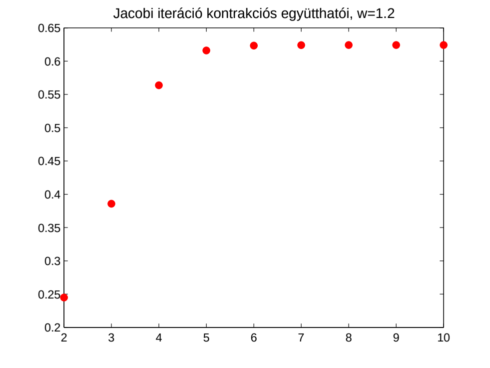
1. Példa:



$q \approx 0.6118$

Tapasztalati kontrakciós együttható vizsgálata

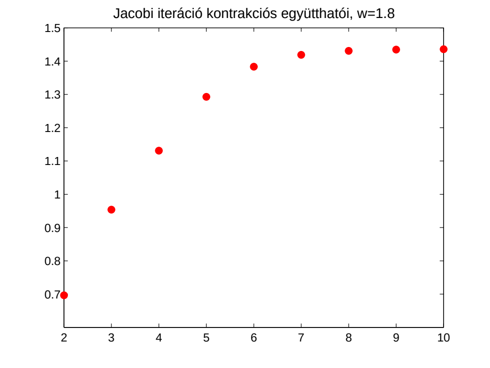
1. Példa:



$q \approx 0.6243$

Tapasztalati kontrakciós együttható vizsgálata

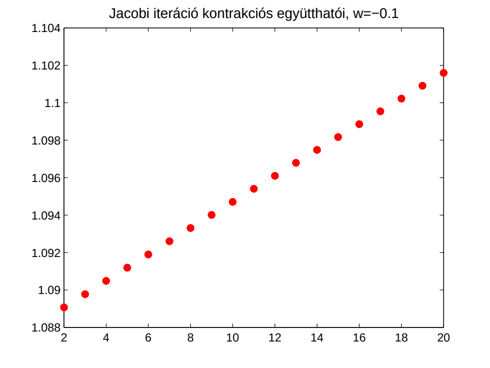
1. Példa:



$q > 1$, divergens sorozat

Tapasztalati kontrakciós együttható vizsgálata

1. Példa:



$q > 1$, divergens sorozat

Numerikus módszerek 1.

9. előadás: Gauss–Seidel iteráció, relaxációs módszer, Richardson típusú iterációk

Krebsz Anna

ELTE IK

Tartalomjegyzék

Az $Ax = b$ LER megoldása érdekében alakítsuk azt át $x = Bx + c$ alakúra, és valamely $x^{(0)}$ kezdőpontból végezzük az

$$x^{(k+1)} = B \cdot x^{(k)} + c \qquad (k \in \mathbb{N}_0)$$

iterációt. A fixponttétel adja meg a sorozat képletét.
A vektorsorozat bizonyos feltételek mellett konvergál a LER megoldásához (lásd fixponttétel, elégséges feltétel, szükséges és elégséges feltétel a konvergenciára a B átmenet mátrixszal).

Volt: Banach-féle fixponttétel, Jacobi-, csillapított Jacobi-iteráció.

Megjegyzés:

- 2–3 változó: felesleges $\Rightarrow$ célja a megértés
- sok változó (100, 1000): használják

Tartalomjegyzék

Átalakítás:

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

Ezek alapján az iteráció a következő.

Definíció: Gauss–Seidel-iteráció

$$x^{(k+1)} = \underbrace{-(L+D)^{-1}U}_{B_S} \cdot x^{(k)} + \underbrace{(L+D)^{-1}b}_{c_S} = B_S \cdot x^{(k)} + c_S$$

Eml.:

$$x^{(k+1)} = -(L+D)^{-1}U \cdot x^{(k)} + (L+D)^{-1}b$$

Írjuk fel koordinátánként! (Kiderül, hogy „helyben” számolható.)

Állítás: a Gauss–Seidel-iteráció komponensenkénti alakja

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

Biz.: Alakítsunk át, majd gondoljunk bele a mátrixszorzásba.

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

Írjuk fel az iteráció reziduum vektoros alakját!

$$(L+D)\cdot x^{(k+1)} = -U\cdot x^{(k)} + b = ((L+D)-A)\cdot x^{(k)} + b =$$
$$= (L+D)\cdot x^{(k)} + \left(-Ax^{(k)} + b\right) = (L+D)\cdot x^{(k)} + r^{(k)}$$
$$\Rightarrow \quad x^{(k+1)} = x^{(k)} + (L+D)^{-1} r^{(k)}$$

Vezessük be az $s^{(k)} := (L+D)^{-1} r^{(k)}$ segédvektort, ezzel egy lépésünk alakja:

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

Az új reziduum vektor:

$$r^{(k+1)} = b - Ax^{(k+1)} = b - A(x^{(k)} + s^{(k)}) = r^{(k)} - As^{(k)}.$$

Algoritmus: Gauss–Seidel-iteráció

$$r^{(0)} := b - Ax^{(0)}$$

$k = 1, \ldots,$ leállásig

$$s^{(k)} := (D+L)^{-1} r^{(k)} \text{ helyett}$$

$$(D+L)\, s^{(k)} = r^{(k)} \text{ LER mo.}$$

$$x^{(k+1)} := x^{(k)} + s^{(k)}$$

$$r^{(k+1)} := r^{(k)} - As^{(k)}$$

Tétel

Ha A szig. diag. dom. a soraira, akkor az $Ax = b$ LER-re felírt Gauss–Seidel-iterációra

$$\|B_S\|_\infty \leq \|B_J\|_\infty < 1$$

teljesül, tehát az konvergens bármely $x^{(0)}$ esetén.

Biz.: Nélkül.

Tétel

Ha A szimmetrikus és pozitív definit, akkor a Gauss–Seidel-iteráció konvergens.

Biz.: Nélkül.

Induljunk a Gauss–Seidel-iteráció következő alakjából:

$$\begin{aligned} (L+D)\cdot x &= -U\cdot x+b & /\cdot\omega \\ D\cdot x &= D\cdot x & /\cdot(1-\omega) \end{aligned}$$

A kettő súlyozott összege:

$$(D+\omega L)\cdot x = [(1-\omega)D-\omega U]\cdot x+\omega b$$

$$x=(D+\omega L)^{-1}[(1-\omega)D-\omega U]\cdot x+(D+\omega L)^{-1}\omega b$$

Ezek alapján az iteráció a következő.

Definíció: relaxált Gauss–Seidel-iteráció ω paraméterrel – $S(\omega)$

$$x^{(k+1)}=\underbrace{(D+\omega L)^{-1}[(1-\omega)D-\omega U]}_{B_{S(\omega)}}\cdot x^{(k)}+\underbrace{\omega(D+\omega L)^{-1}b}_{c_{S(\omega)}}$$

Írjuk fel koordinátánként! (Kiderül, hogy „helyben” számolható.)

Állítás: $S(\omega)$ komponensenkénti alakja

$$x_i^{(k+1)} = (1-\omega)\cdot x_i^{(k)} + \omega \cdot x_{i,S}^{(k+1)},$$

ahol $x_{i,S}^{(k+1)}$ a hagyományos Seidel-módszer ($S = S(1)$) által adott, azaz

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

Minden k. lépés az $i = 1, 2, \ldots, n$ sorrendben számolandó.

Biz.: Alakítsunk át, majd gondoljunk bele a mátrixszorzásba.

$$
\begin{aligned}
(D+\omega L)x^{(k+1)} &= (1-\omega)Dx^{(k)} - \omega Ux^{(k)} + \omega b \\
Dx^{(k+1)} &= (1-\omega)Dx^{(k)} - \omega Lx^{(k+1)} - \omega Ux^{(k)} + \omega b \\
x^{(k+1)} &= (1-\omega)x^{(k)} - \omega \underbrace{D^{-1}\left(Lx^{(k+1)} + Ux^{(k)} - b\right)}_{\text{Lásd } S(1)\text{-nél.}}
\end{aligned}
$$

A koordinátánkénti alakja:

$$
x_i^{(k+1)} = (1-\omega)\cdot x_i^{(k)} - \frac{\omega}{a_{ii}}\left(\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} - b_i\right).
$$

Megj.: Vigyázat! $x^{(k+1)} = (1-\omega)\cdot x^{(k)} + \omega \cdot x_S^{(k+1)}$ nem igaz (tehát az egész vektorra); csak komponensenként.

Relaxált Gauss–Seidel-iteráció

Írjuk fel az iteráció reziduum vektoros alakját!

$$
\begin{aligned}
(D+\omega L)\,x^{(k+1)} &= ((1-\omega)D-\omega U)\cdot x^{(k)}+\omega b = \\
&= D\,x^{(k)}+\omega(\underbrace{(-D-U)}_{L-A})\cdot x^{(k)}+\omega b = \\
&= D\,x^{(k)}+\omega L x^{(k)}+\omega\underbrace{\left(-Ax^{(k)}+b\right)}_{r^{(k)}} = (D+\omega L)\,x^{(k)}+\omega r^{(k)}
\end{aligned}
$$

$$\Rightarrow \quad x^{(k+1)} = x^{(k)}+\omega(D+\omega L)^{-1}r^{(k)}$$

Vezessük be az $s^{(k)} := \omega(D+\omega L)^{-1}r^{(k)}$ segédvektort, ezzel egy lépésünk alakja: $x^{(k+1)} = x^{(k)}+s^{(k)}$.
Az új reziduum vektor:

$$r^{(k+1)} = b-Ax^{(k+1)} = b-A(x^{(k)}+s^{(k)}) = r^{(k)}-As^{(k)}.$$

Algoritmus: relaxált Gauss–Seidel-iteráció $S(\omega)$

$$r^{(0)} := b - Ax^{(0)}$$

$k = 1, \ldots,$ leállásig

$$s^{(k)} := \omega(D+\omega L)^{-1} r^{(k)} \text{ helyett}$$

$$(D+\omega L)\, s^{(k)} = \omega r^{(k)} \text{ LER mo.}$$

$$x^{(k+1)} := x^{(k)} + s^{(k)}$$

$$r^{(k+1)} := r^{(k)} - As^{(k)}$$

Tétel: a relaxált Gauss–Seidel-módszer $S(\omega)$ konvergenciájáról

Ha egy mátrixra az $S(\omega)$ módszer konvergens, akkor $0 < \omega < 2$.

Megjegyzés:

- Ha $\omega \notin (0, 2)$, akkor általában nem konvergál.
- A relaxált Seidel-módszert gyakran alkalmazzák...

Lemma

$$\det(B) = \prod_{i=1}^{n} \lambda_i(B)$$

Biz. lemma: Írjuk fel a B mátrix karakterisztikus polinomját, amelyről tudjuk, hogy gyökei a mátrix sajátértékei; majd rendezzük λ hatványai szerint:

$$p(\lambda) = \det(B - \lambda I) = \prod_{i=1}^{n}(\lambda_i - \lambda) = (-1)^n \cdot \lambda^n + \ldots + \prod_{i=1}^{n} \lambda_i.$$

A $\lambda = 0$ értéket behelyettesítve a konstans tagot kapjuk, amire:

$$p(0) = \det(B) = \prod_{i=1}^{n} \lambda_i.$$

$\square$

Biz. tétel: A konvergencia ekvivalens feltételéből, azaz a

$$\varrho(B_{S(\omega)}) < 1$$

állításból kell ω kívánt becslését előállítanunk. Egyrészt

$$\varrho(B_{S(\omega)}) < 1 \;\Rightarrow\; \left|\lambda_i(B_{S(\omega)})\right| < 1 \;\Rightarrow$$
$$\Rightarrow \left|\prod_{i=1}^{n} \lambda_i(B_{S(\omega)})\right| < 1 \;\Rightarrow\; \left|\det(B_{S(\omega)})\right| < 1.$$

Az iteráció mátrixa

$$B_{S(\omega)} = (D + \omega L)^{-1}[(1-\omega)D - \omega U].$$

Kihasználjuk, hogy háromszögmátrixok determinánsa a főátlóbeli elemek szorzata (tehát nem függ a diagonálison kívüli elemektől).

Biz. tétel folyt.:

$$\left|\det(B_{S(\omega)})\right| = \underbrace{\left|\det\left((D+\omega L)^{-1}\right)\right|}_{1/|\det(D)|} \cdot \underbrace{|\det\left((1-\omega)D - \omega U\right)|}_{|1-\omega|^n \cdot |\det(D)|} =$$

$$= \frac{1}{|\det(D)|} \cdot |1-\omega|^n \cdot |\det(D)| = |1-\omega|^n < 1$$

Ebből pedig $|1-\omega| < 1$ következik, ami ekvivalens a $0 < \omega < 2$ becsléssel. $\square$

Tétel: a relaxált Gauss–Seidel-módszer $S(\omega)$ konvergenciájáról

Ha az egyenletrendszer mátrixa szimmetrikus, pozitív definit és $\omega \in (0, 2)$, akkor az $S(\omega)$ módszer konvergens.

Biz.: nélkül.

Tétel: $S(\omega)$ tridiagonális mátrixokra

Ha a LER mátrixa tridiagonális, akkor a Jacobi- és Gauss–Seidel-iteráció egyszerre konvergens vagy divergens

$$\text{azaz } \varrho(B_S) = \varrho(B_J)^2.$$

Ez azt jelenti, hogy konvergencia esetén a Gauss–Seidel-iteráció kétszer gyorsabb,

Biz.: nélkül.

Tétel: $S(\omega)$ szimmetrikus, pozitív definit és tridiagonális mátrixokra

Ha a LER mátrixa tridiagonális, szimmetrikus és pozitív definit, akkor a Jacobi-, Gauss–Seidel- és relaxált Gauss–Seidel-iteráció is konvergens. Megadható $S(\omega)$-ra optimális paraméter

$$\omega_0 = \frac{2}{1 + \sqrt{1 - \varrho(B_J)^2}}.$$

Továbbá,

- ha $\varrho(B_J) = 0$, akkor $\omega_0 = 1$ és $\varrho(B_S) = \varrho(B_{S(\omega_0)}) = 0$,
- $\varrho(B_J) \neq 0$, akkor $\varrho(B_{S(\omega_0)}) = \omega_0 - 1 < \varrho(B_S) = \varrho(B_J)^2$.

Biz.: nélkül.

Megj.:

- Az utóbbi két tétel blokktridiagonális mátrixokra is igaz, a megfelelő blokkiterációkra.
- Az iterációs módszer konvergencia sebessége a q kontrakciós együtthatótól függ. Minél közelebb van 0-hoz, annál gyorsabb a módszer, míg, ha 1-hez van közel, akkor nagyon lassú. A kontrakciós együtthatót $q = \|B\|$-ként kapjuk.
- Mivel bármely normára $\inf\{\|B\| : B \text{ indukált norma}\} = \varrho(B)$, ezért a spektrálsugár határozza meg a konvergencia sebességét.

Példa

Mit állíthatunk a következő mátrixra felírt Jacobi- és Gauss–Seidel-iterációk konvergenciájáról?

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

A szimmetrikus, pozitív definit és tridiagonális, alkalmazhatók rá a tanult tételek:

- A $J(1)$ iteráció konvergens minden kezdővektorra.
- Ha $\omega \in (0;1)$-re, akkor $J(\omega)$ iteráció konvergens minden kezdővektorra.
- Az $S(1)$ iteráció konvergens minden kezdővektorra.
- Az $S(\omega)$ iteráció konvergens minden kezdővektorra pontosan az $\omega \in (0;2)$ értékekre.

Példa

$$B_J = \frac{1}{2}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad B_S = \frac{1}{8}\begin{bmatrix} 0 & 4 & 0 \\ 0 & 2 & 4 \\ 0 & 1 & 2 \end{bmatrix}$$

B_J sajátértékei: $0, \pm\frac{1}{\sqrt{2}}$, így $\varrho(B_J) = \frac{1}{\sqrt{2}}$.

B_S sajátértékei: $0, 0, \frac{1}{2}$, így $\varrho(B_S) = \frac{1}{2}$.

$S(\omega)$-ra az optimális paraméter:

$$\omega_0 = \frac{2}{1+\sqrt{1-\left(\frac{1}{\sqrt{2}}\right)^2}} = \frac{2}{1+\frac{1}{\sqrt{2}}} = \frac{2\sqrt{2}}{1+\sqrt{2}} = 4-2\sqrt{2} \approx 1,1716...$$

$\varrho(B_{S(\omega_0)}) = \omega_0 - 1 = 3 - 2\sqrt{2} \approx 0,1716...$
Nézzük meg a LER-re a csillapított Jacobi- és a relaxált Gauss–Seidel-iteráció vizsgálatát Maple-ben.

Tekintsük az $Ax = b$ LER-t, ahol A szimmetrikus, pozitív definit mátrix és $p \in \mathbb{R}$.

$$\begin{aligned} Ax &= b \\ p \cdot Ax &= p \cdot b \\ 0 &= -pAx + pb \\ x &= x - pAx + pb = (I - pA)x + pb \end{aligned}$$

Ezek alapján az iteráció a következő.

Definíció: Richardson-iteráció p paraméterrel – $R(p)$

$$x^{(k+1)} = \underbrace{(I - pA)}_{B_{R(p)}} \cdot x^{(k)} + \underbrace{pb}_{c_{R(p)}} = B_{R(p)} \cdot x^{(k)} + c_{R(p)}$$

Írjuk fel az iteráció reziduum vektoros alakját!

$$x^{(k+1)} = x^{(k)} - pAx^{(k)} + pb = x^{(k)} + p \cdot \left(-Ax^{(k)} + b\right) = \\ = x^{(k)} + pr^{(k)}$$

Vezessük be az $s^{(k)} := pr^{(k)}$ segédvektort, ezzel egy lépésünk alakja:

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

Az új reziduum vektor:

$$r^{(k+1)} = b - Ax^{(k+1)} = b - A(x^{(k)} + s^{(k)}) = r^{(k)} - As^{(k)}.$$

Algoritmus: Richardson-iteráció

$$r^{(0)} := b - Ax^{(0)}$$

$k = 1, \ldots,$ leállásig

$$s^{(k)} := pr^{(k)}$$

$$x^{(k+1)} := x^{(k)} + s^{(k)}$$

$$r^{(k+1)} := r^{(k)} - As^{(k)}$$

Megjegyzés: Érdemes meggondolni, hogy ha az $Ax = b$ helyett a $D = \operatorname{diag}(a_{11}, \ldots, a_{nn})$ diagonális mátrix-szal a $D^{-1}Ax = D^{-1}b$ LER-re alkalmazzuk az $R(p)$ iterációt, akkor az eredeti LER-re felírt $J(p)$ csillapított Jacobi-iterációt kapjuk.

Tétel: A Richardson-iteráció konvergenciája

Ha az $A \in \mathbb{R}^{n \times n}$ mátrix szimmetrikus, pozitív definit és sajátértékeire $m = \lambda_1 \leq \cdots \leq \lambda_n = M$ teljesül, akkor $R(p)$ (azaz az $Ax = b$ LER-re felírt $p \in \mathbb{R}$ paraméterű Richardson-iteráció) pontosan a

$$p \in \left(0, \frac{2}{M}\right),$$

paraméter értékekre konvergens minden kezdővektor esetén. Az optimális paraméter $p_0 = \frac{2}{M+m}$, a hozzá kapcsolódó kontrakciós együttható pedig:

$$\varrho(B_{R(p_0)}) := \frac{M-m}{M+m} = \left\|B_{R(p_0)}\right\|_2 = q.$$

Bizonyítás:

1. $B_{R(p)}$ sajátértékei: $\lambda_i(p) = 1 - p \cdot \lambda_i$, hiszen

   $$Av = \lambda_i v \quad \Rightarrow \quad (I - pA)v = v - pAv = v - p\lambda_i v = (1 - p\lambda_i)v.$$

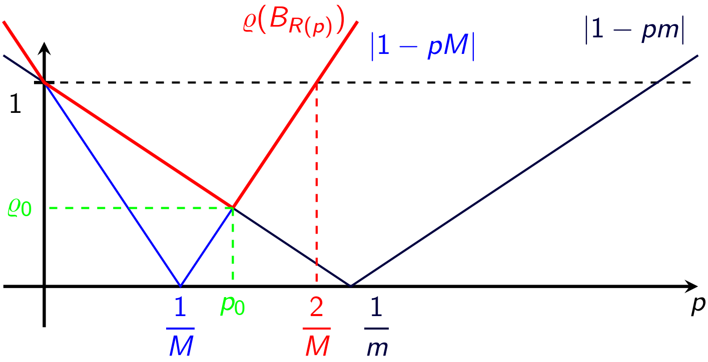
   Vagyis:

   $$\begin{aligned} \lambda_1(p) &= 1 - p \cdot \lambda_1 = 1 - pm, \\ \lambda_2(p) &= 1 - p \cdot \lambda_2, \\ &\vdots \\ \lambda_n(p) &= 1 - p \cdot \lambda_n = 1 - pM. \end{aligned}$$

2. $B_{R(p)}$ spektrálsugara rögzített p-re
   $\varrho(B_{R(p)}) = \max\limits_{i=1}^{n} |1 - p \cdot \lambda_i|$.

❸ Ábrázoljuk az $|1 - p \cdot \lambda_i|$ függvényeket $(i = 1, 2, \ldots, n)$!
(Ezek p-től függenek.)

$$1 - p \cdot \lambda_i = 0 \quad \Longleftrightarrow \quad p = \frac{1}{\lambda_i}$$



4. $R(p)$ konvergens, ha $\varrho(B_{R(p)}) < 1$, azaz ha $p \in \left(0, \dfrac{2}{M}\right)$.
   Ezek az $|1 - pM| = 1$ egyenlet megoldásai.

5. Továbbá az optimális paramétert az

$$|1 - pM| = |1 - pm|$$

   egyenlet megoldása adja. (Nem a 0, hanem a másik.)

$$\begin{aligned} -1 + pM &= 1 - pm \\ pM + pm &= 2 \\ p(M + m) &= 2 \quad \Longrightarrow \quad p_0 = \frac{2}{M + m} \end{aligned}$$

❻

$$\varrho(B_{R(p_0)}) = 1 - p_0 \cdot m = \frac{M+m}{M+m} - \frac{2m}{M+m} = \frac{M-m}{M+m}.$$

❼ Mivel A szimmetrikus, így $B_{R(p)}$ is, ezért a spektrálsugara és kettes normája megegyezik. Az eredményül kapott spektrálsugár egyben kettes normabeli kontrakciós együttható:

$$q = \frac{M-m}{M+m}.$$

□

Példa

Adjuk meg, hogy a Richardson-iteráció mely $p \in \mathbb{R}$ paraméterek mellett konvergens a következő egyenletrendszer esetén – mely ugyanaz, mint az imént. Mi az optimális paraméter és a hozzá tartozó „átmenetmátrix" spektrálsugara?

$$Ax = b, \qquad \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix} \cdot x = \begin{bmatrix} 1 \\ 5 \end{bmatrix}.$$

A mátrix sajátértékei 2 és 4.

$M = 4$, $m = 2$, így a $p \in (0, \frac{2}{M}) = (0; \frac{1}{2})$ értékekre a Richardson-iteráció konvergens bármely kezdővektor esetén. Az optimális paraméter

$$p_0 = \frac{2}{M+m} = \frac{2}{4+2} = \frac{1}{3}$$

és a hozzá tartozó átmenetmátrix spektrálsugara

$$\varrho(B_{R(p_0)}) = \frac{4-2}{4+2} = \frac{1}{3}.$$

Mivel A szimmetriája öröklődik $B_{R(p)}$-re, így az átmenetmátrix is szimmetrikus, így

$$\left\| B_{R(p_0)} \right\|_2 = \varrho(B_{R(p_0)}) = \frac{1}{3} = q$$

a kontrakciós együttható a kettes normában.

Tartalomjegyzék

Példák Matlab-ban

1. A tapasztalati kontrakciós együtthatók szemléltetése a relaxációs módszer esetén.
2. A Richardson-iteráció viselkedésének vizsgálata különböző paraméterek mellett.

1. Példa:

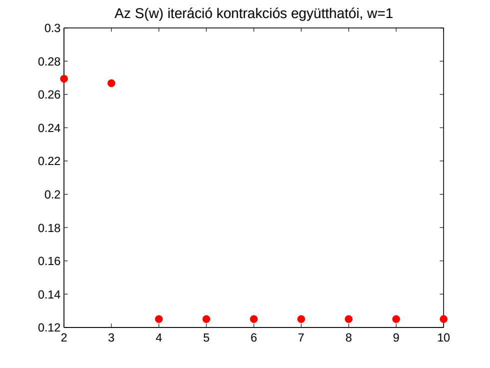
A LER alakja $Ax = b$, ahol

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

Vizsgáljuk a relaxációs módszer tapasztalati kontrakciós együtthatóit $\omega = 1, 0.8, 0.6, 1.033, -0.1, 2, 2.5$ esetén!

Tapasztalati kontrakciós együttható vizsgálata

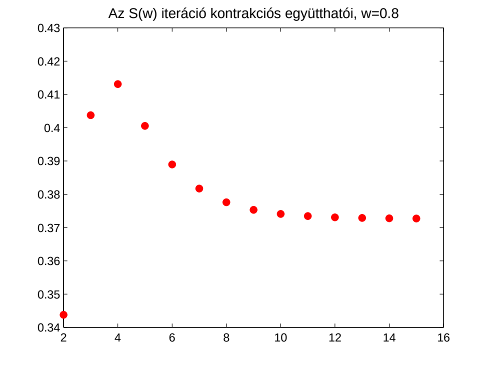
1. Példa:



$q \approx 0.1250$

Tapasztalati kontrakciós együttható vizsgálata

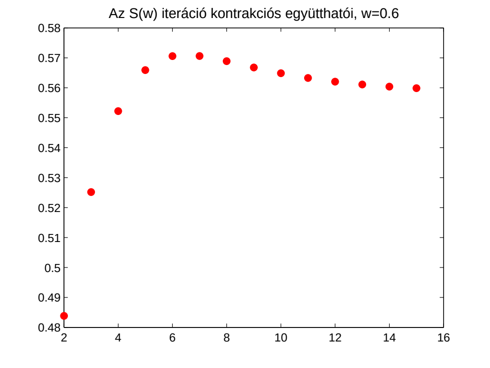
1. Példa:



$q \approx 0.3750$

Tapasztalati kontrakciós együttható vizsgálata

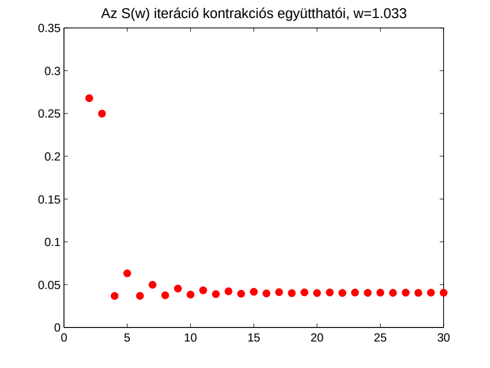
1. Példa:



$q \approx 0.5650$

Tapasztalati kontrakciós együttható vizsgálata

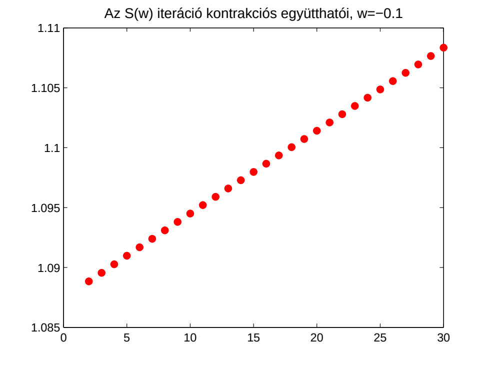
1. Példa:



$q \approx 0.0404$

Tapasztalati kontrakciós együttható vizsgálata

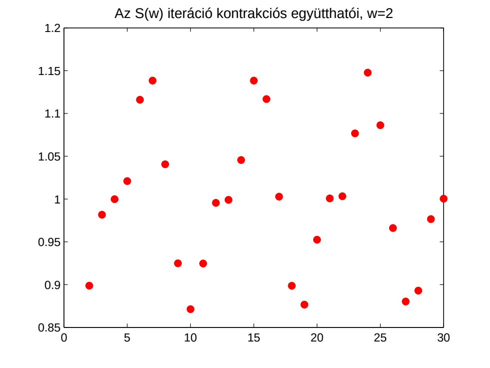
1. Példa:



$q > 1$, divergens iteráció

Tapasztalati kontrakciós együttható vizsgálata

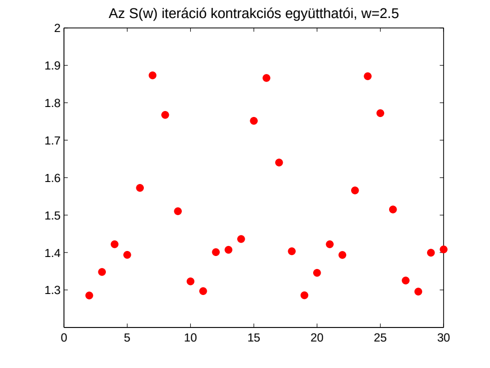
1. Példa:



$q > 1$, divergens iteráció

Tapasztalati kontrakciós együttható vizsgálata

1. Példa:



$q > 1$, divergens iteráció

Numerikus módszerek 1.

10. előadás: Részleges LU-felbontás és algoritmus, kerekítési hibák hatása az iterációkra

Krebsz Anna

ELTE IK

Tartalomjegyzék

Általában:

$$Ax = b, \qquad A = P + Q, \qquad (P + Q)x = b,$$

átrendezve:

$$Px = -Qx + b \quad \Longleftrightarrow \quad x = -P^{-1}Qx + P^{-1}b,$$

iterációs alakban írva:

$$x^{(k+1)} = \underbrace{-P^{-1}Q}_{B} \cdot x^{(k)} + \underbrace{P^{-1}b}_{c}.$$

A továbbiakban olyan $P = LU$ felbontást és $-Q$ mátrixot keresünk, melyre $A = P - Q$. Ekkor a P^{-1}-zel való számolás helyettesíthető két háromszög alakú LER megoldásával, vagyis az iteráció könnyen számolható. Ezzel egy módszercsaládot konstruálunk.

Definíció: ILU-felbontás

- Legyen J a mátrix elemek pozícióinak egy részhalmaza, mely nem tartalmazza a főátlót, azaz $(i,i) \notin J \quad \forall\, i$-re.
  A J halmazt *pozícióhalmaz*nak nevezzük.
- Az A mátrixnak a J pozícióhalmazra illeszkedő *részleges LU-felbontásán* (*ILU-felbontásán*) olyan LU-felbontást értünk, melyre $L \in \mathcal{L}_1$ és $U \in \mathcal{U}$ (tehát a szokásos alakúak), továbbá

$$\forall\,(i,j) \in J: \quad l_{ij} = 0, \;\; u_{ij} = 0 \;\text{ és}$$
$$\forall\,(i,j) \notin J: \quad a_{ij} = (LU)_{ij}.$$

Algoritmus: ILU-felbontás GE-val

$\widetilde{A}_1 := A$
$k = 1, \ldots, n-1$:

(1) Szétbontás: $\widetilde{A}_k = P_k - Q_k$ alakra, ahol

$$\begin{aligned}
(P_k)_{ik} &= 0 \quad (i,k) \in J \\
(P_k)_{kj} &= 0 \quad (k,j) \in J \\
(Q_k)_{ik} &= -\widetilde{a}_{ik}^{(k)} \quad (i,k) \in J \\
(Q_k)_{kj} &= -\widetilde{a}_{kj}^{(k)} \quad (k,j) \in J.
\end{aligned}$$

Ahogy látható, $\widetilde{A}_k$-nak csak k. sorában és k. oszlopában a pozícióhalmazban megadott helyeken változtatunk.

(2) Elimináció P_k-n:

$$\widetilde{A}_{k+1} = L_k P_k$$

Kérdés: az algoritmussal kapott mátrixokból hogyan állítjuk elő az ILU-felbontást?

Tétel: az ILU-felbontásról

Az ILU-felbontás algoritmusával kapott részmátrixokból készítsük el a következőket:

$$\begin{aligned} U &:= \widetilde{A}_n, \\ L &:= L_1^{-1} \cdot \ldots \cdot L_{n-1}^{-1} \quad \text{(összepakolással)}, \\ Q &:= Q_1 + Q_2 + \ldots + Q_{n-1} \quad \text{(összepakolással)}. \end{aligned}$$

Ekkor $A = LU - Q$ és a részleges LU-felbontásra vonatkozó feltételek teljesülnek.

Biz.: A GE $n-1$. lépése után felsőháromszög alakot kapunk, tehát $U := \widetilde{A}_n$ alakja jó és minden $(i,j) \in J, i < j$-re $u_{ij} = 0$. Alkalmazzuk az $n-1$. lépés (2), majd (1) részét:

$$U := \widetilde{A}_n = L_{n-1}P_{n-1} = L_{n-1}\left(\widetilde{A}_{n-1} + Q_{n-1}\right)$$

Az $\widetilde{A}_n$-re kapott rekurziót alkalmazzuk $\widetilde{A}_{n-1}$-re:

$$\widetilde{A}_n = L_{n-1}\left(\widetilde{A}_{n-1} + Q_{n-1}\right) = L_{n-1}\left(L_{n-2}\left[\widetilde{A}_{n-2} + Q_{n-2}\right] + Q_{n-1}\right)$$

Mivel Q_{n-1}-ben az $n-2$. sorban csak nullák vannak, így az $n-2$. GE-s lépés nem változtat rajta, tehát $L_{n-2}Q_{n-1} = Q_{n-1}$. Emiatt Q_{n-1}-et bevihetjük a belső zárójelbe.

$$\widetilde{A}_n = L_{n-1}L_{n-2}\left(\widetilde{A}_{n-2} + Q_{n-2} + Q_{n-1}\right)$$

Biz. folyt.: Folytatva tovább visszafelé a rekurziót

$$U = \widetilde{A}_n = L_{n-1}L_{n-2}\left(\widetilde{A}_{n-2} + Q_{n-2} + Q_{n-1}\right) = \ldots =$$

$$= \underbrace{L_{n-1}L_{n-2}\ldots L_1}_{L^{-1}}\left(A + \underbrace{Q_1 + \ldots + Q_{n-2} + Q_{n-1}}_{Q}\right).$$

$$U = L^{-1}(A+Q) \quad \Leftrightarrow \quad A = LU - Q$$

A kapott mátrixok alakja megfelelő. Az algoritmus (1) lépése garantálja, hogy $\forall\,(i,j) \in J: \;\; l_{ij} = 0, \;\; u_{ij} = 0$, továbbá (2) lépése (GE) miatt $\forall\,(i,j) \notin J: \;\; a_{ij} = (LU)_{ij}$. $\square$

Tétel: szig.diag.dom. mátrix ILU-felbontása

Ha A szigorúan diagonálisan domináns a soraira vagy oszlopaira, akkor a mátrix ILU-felbontása létezik és egyértelmű.

Biz.: az ILU-felbontás (1) lépése a szig. diag. dom. tulajdonságot nem változtatja, mivel átlón kívüli elemet veszünk ki a mátrixból.

A (2) GE-s lépés a szig. diag. dom. tulajdonságot magtartja, lásd GE megmaradási tételek a Schur-komplementerre. $\square$

Megjegyzés:

1. A szig. diag. dom. tulajdonságból következik az összes bal felső részmátrix invertálhatósága, vagyis a főminorok egyike sem nulla.
2. Diff. egyenletek numerikus megoldása során gyakran előforduló M-mátrix osztályra is igaz, hogy egyértelműen létezik az ILU-felbontása.
3. Gyakran csak a főátlót és néhány mellékátlót hagynak ki a J pozícióhalmazból, így a tárigény előre ismert, nem kell a sávon belül feltöltődéssel foglalkozni.

4. Például egy $N^2 \times N^2$-es mátrix esetén, ahol csak a $(-N, -1, 0, 1, N)$ átlókban van nem nulla elem, érdemes J-ből a $(-1, 0, 1)$ átlókat kihagyni.

   Tárolás: L, U csak két-két átlót fog tartalmazni, L átlója egyesekből áll, így 3 db N^2 méretű átlót kell tárolni N^4 elem helyett.

   Műveletigény: az iteráció során a két háromszögmátrixú két átlós LER $2N^2 + \mathcal{O}(1)$ illetve $3N^2 + \mathcal{O}(1)$ művelettel megoldható. (A GE $\frac{2}{3}N^6$-t jelentene.) Gondoljunk arra, hogy $N \approx 10^3$...

1. Példa:

Készítsük el az

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

mátrix $J = \{(1,2),(2,3)\}$ pozícióhalmazhoz illeszkedő ILU-felbontását! A pozícióhalmazt mátrixos alakban is szemléltethetjük, a kinullázandó elemeket $*$-gal jelöljük:

$$\begin{bmatrix} & * & \\ & & * \\ & & \end{bmatrix}.$$

1. lépés: (1) **szétbontás:** olyan pozíciókat keresünk J-ben, melyek az 1. sorhoz illetve az 1. oszlophoz tartoznak: $(1,2)$.
Ezt a pozíciót kinullázzuk P_1-ben és a (-1)-szeresét Q_1-be tesszük.

$$A = \widetilde{A}_1 = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix} - \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = P_1 - Q_1$$

(2) **Elimináció:** P_1-en elvégezzük az 1. GE-s lépést:

$$\widetilde{A}_2 = L_1 P_1 = \begin{bmatrix} 4 & 0 & 2 \\ 0 & 4 & \frac{1}{2} \\ 0 & 1 & 3 \end{bmatrix}, \quad L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix}.$$

2. lépés: (1) **szétbontás:** olyan pozíciókat keresünk J-ben, melyek a 2. sorhoz illetve az 2. oszlophoz tartoznak: $(2,3)$.
Ezt a pozíciót kinullázzuk P_2-ben és a (-1)-szeresét Q_2-be tesszük.

$$\widetilde{A}_2 = \begin{bmatrix} 4 & 0 & 2 \\ 0 & 4 & \frac{1}{2} \\ 0 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 2 \\ 0 & 4 & 0 \\ 0 & 1 & 3 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix} = P_2 - Q_2$$

(2) **Elimináció:** P_2-en elvégezzük a 2. GE-s lépést:

$$\widetilde{A}_3 = L_2 P_2 = \begin{bmatrix} 4 & 0 & 2 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{4} & 1 \end{bmatrix}.$$

A tétel alapján összerakjuk az ILU-felbontást:

$$U = \widetilde{A}_3 = \begin{bmatrix} 4 & 0 & 2 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad Q = Q_1 + Q_2 = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix}$$

Mivel az egyes lépésekben a Q_k mátrixok különböző sorait és oszlopait töltjük, így elegedő a gyakorlatban egy Q mátrixot tárolni. Az iterációnál látni fogjuk, hogy Q-ra a végrehajtáshoz nincs szükség.

Összepakolással:

$$L = L_1^{-1} L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{4} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix}.$$

Ellenőrizhetjük, hogy $A = LU - Q$

$$\begin{bmatrix} 4 & 1 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 & 0 & 2 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}}_{\begin{bmatrix} 4 & 0 & 2 \\ 1 & 4 & \frac{1}{2} \\ 2 & 1 & 4 \end{bmatrix}} - \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix}.$$

Teljesíti a ILU-felbontásra tett összes követelményt. $\square$

1. Példa tömören

Tömör írásmódban: Csak egy Q mátrixot tárolunk, ebbe pakoljuk a Q_k mátrixok nem nulla elemeit. Az GE eredményét illetve a GE-s hányadosokat, vagyis az $\widetilde{A}_k, L_k$ mátrixokat is egyben tároljuk. Vonalakkal jelezzük, hogy itt már tárolásról is szó van.

1. lépés: (1) **szétbontás:**

$$A = \widetilde{A}_1 = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix}, \quad P_1 = \begin{bmatrix} 4 & 0 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

(2) **Elimináció** P_1**-en:**

$$\begin{bmatrix} 4 & 0 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix} \to \left[\begin{array}{c|cc} 4 & 0 & 2 \\ \hline \frac{1}{4} & 4 & \frac{1}{2} \\ \frac{2}{4} & 1 & 3 \end{array}\right]$$

2. lépés: Ugyanúgy dolgozunk tovább, de most már csak a jobb alsó 2×2-es mátrix részen, a többit változatlanul leírjuk.

(1) **szétbontás:**

$$\left[\begin{array}{c|cc} 4 & 0 & 2 \\ \hline \frac{1}{4} & 4 & \frac{1}{2} \\ \frac{1}{2} & 1 & 3 \end{array}\right] \rightarrow \left[\begin{array}{c|cc} 4 & 0 & 2 \\ \hline \frac{1}{4} & 4 & \color{red}{0} \\ \frac{1}{2} & 1 & 3 \end{array}\right] \qquad Q = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & \color{red}{-\frac{1}{2}} \\ 0 & 0 & 0 \end{bmatrix}$$

(2) **Elimináció:**

$$\left[\begin{array}{c|cc} 4 & 0 & 2 \\ \hline \frac{1}{4} & 4 & 0 \\ \frac{1}{2} & 1 & 3 \end{array}\right] \rightarrow \left[\begin{array}{c|cc} 4 & 0 & 2 \\ \hline \frac{1}{4} & 4 & 0 \\ \cline{2-3} \frac{1}{2} & \multicolumn{1}{c|}{\color{red}{\frac{1}{4}}} & \color{red}{3} \end{array}\right] = L \text{ és } U \text{ együtt}$$

$\square$

2. Példa:

Készítsük el az

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

mátrix $J = \{(1,2),(1,3),(2,1),(2,3),(3,1),(3,2)\}$ pozícióhalmazhoz illeszkedő ILU-felbontását! A pozícióhalmazt mátrixos alakban is szemléltethetjük, a kinullázandó elemeket $*$-gal jelöljük:

$$\begin{bmatrix} & * & * \\ * & & * \\ * & * & \end{bmatrix}.$$

A lehető legbővebb pozícióhalmazt adtuk meg.

1. lépés: (1) **szétbontás:** olyan pozíciókat keresünk J-ben, melyek az 1. sorhoz illetve az 1. oszlophoz tartoznak: $(1,2),(1,3),(2,1),(3,1)$.
Ezeket a pozíciókat kinullázzuk P_1-ben és a (-1)-szeresüket Q_1-be tesszük.

$$A = \widetilde{A}_1 = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix} - \begin{bmatrix} 0 & -1 & -2 \\ -1 & 0 & 0 \\ -2 & 0 & 0 \end{bmatrix} = P_1 - Q_1$$

(2) **Elimináció:** P_1-en elvégezzük az 1. GE-s lépést (valójában nem kell eliminálnunk a kinullázások miatt):

$$\widetilde{A}_2 = L_1 P_1 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix}, \quad L_1^{-1} = I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

2. lépés: (1) **szétbontás:** olyan pozíciókat keresünk J-ben, melyek a 2. sorhoz illetve az 2. oszlophoz tartoznak: $(2,3),(3,2)$. Ezeket a pozíciókat kinullázzuk P_2-ben és a (-1)-szeresüket Q_2-be tesszük.

$$\widetilde{A}_2 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix} = P_2 - Q_2$$

(2) **Elimináció:** P_2-en elvégezzük a 2. GE-s lépést (valójában nem kell eliminálnunk a kinullázások miatt):

$$\widetilde{A}_3 = L_2 P_2 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad L_2^{-1} = I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

A tétel alapján összerakjuk az ILU-felbontást:

$$U = \widetilde{A}_3 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad Q = Q_1 + Q_2 = \begin{bmatrix} 0 & -1 & -2 \\ -1 & 0 & -1 \\ -2 & -1 & 0 \end{bmatrix}$$

Mivel az egyes lépésekben a Q_k mátrixok különböző sorait és oszlopait töltjük, így elegedő a gyakorlatban egy Q mátrixot tárolni. Az iterációnál látni fogjuk, hogy Q-ra a végrehajtáshoz nincs szükség.

Összepakolással:

$$L = L_1^{-1} L_2^{-1} = I.$$

Ellenőrizhetjük, hogy $A = LU - Q$

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

Teljesíti a *ILU*-felbontásra tett összes követelményt. $\square$

2. Példa tömören

Tömör írásmódban: Csak egy Q mátrixot tárolunk, ebbe pakoljuk a Q_k mátrixok nem nulla elemeit. Az GE eredményét illetve a GE-s hányadosokat, vagyis az $\widetilde{A}_k, L_k$ mátrixokat is egyben tároljuk. Vonalakkal jelezzük, hogy itt már tárolásról is szó van.

1. lépés: (1) **szétbontás:**

$$A = \widetilde{A}_1 = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix}, \quad P_1 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & -1 & -2 \\ -1 & 0 & 0 \\ -2 & 0 & 0 \end{bmatrix}$$

(2) **Elimináció** P_1**-en:** valójában nem kell eliminálni.

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix} \to \left[\begin{array}{c|cc} 4 & 0 & 0 \\ \hline 0 & 4 & 1 \\ 0 & 1 & 4 \end{array}\right]$$

2. lépés: Ugyanúgy dolgozunk tovább, de most már csak a jobb alsó 2×2-es mátrix részen, a többit változatlanul leírjuk.

(1) **szétbontás:**

$$\left[\begin{array}{c|cc} 4 & 0 & 0 \\ \hline 0 & 4 & 1 \\ 0 & 1 & 4 \end{array}\right] \to \left[\begin{array}{c|cc} 4 & 0 & 0 \\ \hline 0 & 4 & \color{red}{0} \\ 0 & \color{red}{0} & 4 \end{array}\right] \qquad Q = \begin{bmatrix} 0 & -1 & -2 \\ -1 & 0 & \color{red}{-1} \\ -2 & \color{red}{-1} & 0 \end{bmatrix}$$

(2) **Elimináció:** valójában nem kell eliminálni.

$$\left[\begin{array}{c|cc} 4 & 0 & 0 \\ \hline 0 & 4 & 0 \\ 0 & 0 & 4 \end{array}\right] \to \left[\begin{array}{c|c|c} 4 & 0 & 0 \\ \hline 0 & 4 & 0 \\ \cline{2-3} 0 & \color{red}{0} & \color{red}{4} \end{array}\right] = L \text{ és } U \text{ együtt}$$

$\square$

3. Példa:

Készítsük el az

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

mátrix $J = \{(1,2),(1,3),(2,3)\}$ pozícióhalmazhoz illeszkedő ILU-felbontását! A pozícióhalmazt mátrixos alakban is szemléltethetjük, a kinullázandó elemeket $*$-gal jelöljük:

$$\begin{bmatrix} & * & * \\ & & * \\ & & \end{bmatrix}.$$

A felsőháromszögrész minden átlón kívüli elemét megjelöltük.

1. lépés: (1) **szétbontás:** olyan pozíciókat keresünk J-ben, melyek az 1. sorhoz illetve az 1. oszlophoz tartoznak: $(1,2),(1,3)$. Ezeket a pozíciókat kinullázzuk P_1-ben és a (-1)-szeresüket Q_1-be tesszük.

$$A = \widetilde{A}_1 = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix} - \begin{bmatrix} 0 & -1 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = P_1 - Q_1$$

(2) **Elimináció:** P_1-en elvégezzük az 1. GE-s lépést:

$$\widetilde{A}_2 = L_1 P_1 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix}, \quad L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix}.$$

2. lépés: (1) **szétbontás:** olyan pozíciókat keresünk J-ben, melyek a 2. sorhoz illetve az 2. oszlophoz tartoznak: $(2,3)$.
Ezt a pozíciót kinullázzuk P_2-ben és a (-1)-szeresét Q_2-be tesszük.

$$\widetilde{A}_2 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 1 & 4 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} = P_2 - Q_2$$

(2) **Elimináció:** P_2-en elvégezzük a 2. GE-s lépést:

$$\widetilde{A}_3 = L_2 P_2 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{4} & 1 \end{bmatrix}.$$

A tétel alapján összerakjuk az ILU-felbontást:

$$U = \widetilde{A}_3 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad Q = Q_1 + Q_2 = \begin{bmatrix} 0 & -1 & -2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

Mivel az egyes lépésekben a Q_k mátrixok különböző sorait és oszlopait töltjük, így elegedő a gyakorlatban egy Q mátrixot tárolni. Az iterációnál látni fogjuk, hogy Q-ra a végrehajtáshoz nincs szükség.

Összepakolással:

$$L = L_1^{-1} L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{4} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix}.$$

Ellenőrizhetjük, hogy $A = LU - Q$

$$\begin{bmatrix} 4 & 1 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}}_{\begin{bmatrix} 4 & 0 & 0 \\ 1 & 4 & 0 \\ 2 & 1 & 4 \end{bmatrix}} - \begin{bmatrix} 0 & -1 & -2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Teljesíti a *ILU*-felbontásra tett összes követelményt. $\square$

Tömör írásmódban: Csak egy Q mátrixot tárolunk, ebbe pakoljuk a Q_k mátrixok nem nulla elemeit. Az GE eredményét illetve a GE-s hányadosokat, vagyis az $\widetilde{A}_k, L_k$ mátrixokat is egyben tároljuk. Vonalakkal jelezzük, hogy itt már tárolásról is szó van.

1. lépés: (1) **szétbontás:**

$$A = \widetilde{A}_1 = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix}, \quad P_1 = \begin{bmatrix} 4 & \color{red}{0} & \color{red}{0} \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & \color{red}{-1} & \color{red}{-2} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

(2) **Elimináció P_1-en:**

$$\begin{bmatrix} 4 & 0 & 0 \\ 1 & 4 & 1 \\ 2 & 1 & 4 \end{bmatrix} \to \left[\begin{array}{c|cc} 4 & 0 & 0 \\ \hline \color{red}{\frac{1}{4}} & \color{red}{4} & \color{red}{1} \\ \color{red}{\frac{1}{2}} & \color{red}{1} & \color{red}{4} \end{array}\right]$$

2. lépés: Ugyanúgy dolgozunk tovább, de most már csak a jobb alsó 2×2-es mátrix részen, a többit változatlanul leírjuk.

(1) **szétbontás:**

$$\left[\begin{array}{c|cc} 4 & 0 & 0 \\ \hline \frac{1}{4} & 4 & 1 \\ \frac{1}{2} & 1 & 4 \end{array}\right] \to \left[\begin{array}{c|cc} 4 & 0 & 0 \\ \hline \frac{1}{4} & 4 & \color{red}{0} \\ \frac{1}{2} & 1 & 4 \end{array}\right] \qquad Q = \begin{bmatrix} 0 & -1 & -2 \\ 0 & 0 & \color{red}{-1} \\ 0 & 0 & 0 \end{bmatrix}$$

(2) **Elimináció:**

$$\left[\begin{array}{c|cc} 4 & 0 & 0 \\ \hline \frac{1}{4} & 4 & 0 \\ \frac{1}{2} & 1 & 4 \end{array}\right] \to \left[\begin{array}{c|cc} 4 & 0 & 0 \\ \hline \frac{1}{4} & 4 & 0 \\ \cline{2-3} \frac{1}{2} & \multicolumn{1}{c|}{\color{red}{\frac{1}{4}}} & \color{red}{4} \end{array}\right] = L \text{ és } U \text{ együtt}$$

$\square$

Átalakítás:

$$
\begin{aligned}
Ax = b, \quad A = P - Q, \quad P = LU \\
(P - Q)x &= b \\
Px &= Qx + b \\
x &= P^{-1}Qx + P^{-1}b
\end{aligned}
$$

Ezek alapján az iteráció a következő.

Definíció: ILU-algoritmus

$$x^{(k+1)} = \underbrace{P^{-1}Q}_{B_{ILU}} \cdot x^{(k)} + \underbrace{P^{-1}b}_{c_{ILU}} = B_{ILU} \cdot x^{(k)} + c_{ILU}$$

Írjuk fel az iteráció reziduum vektoros alakját!

$$A = P - Q \quad \Leftrightarrow \quad Q = P - A$$

$$\begin{aligned} P \cdot x^{(k+1)} &= Q \cdot x^{(k)} + b = (P - A) \cdot x^{(k)} + b = \\ &= P \cdot x^{(k)} + \left(-Ax^{(k)} + b\right) = P \cdot x^{(k)} + r^{(k)} \\ \Rightarrow \quad x^{(k+1)} &= x^{(k)} + P^{-1} r^{(k)} \end{aligned}$$

Vezessük be az $s^{(k)} := P^{-1} r^{(k)}$ segédvektort, ezzel egy lépésünk alakja:

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

Az új reziduum vektor:

$$r^{(k+1)} = b - Ax^{(k+1)} = b - A(x^{(k)} + s^{(k)}) = r^{(k)} - As^{(k)}.$$

Algoritmus: *ILU*-algoritmus

$$r^{(0)} := b - Ax^{(0)}$$

$k = 1, \ldots,$ leállásig

$$\begin{aligned} s^{(k)} &:= P^{-1} r^{(k)} \ \text{ helyett} \\ LU\, s^{(k)} &= r^{(k)} \ \text{ (2 db háromszögű LER mo.)} \\ x^{(k+1)} &:= x^{(k)} + s^{(k)} \\ r^{(k+1)} &:= r^{(k)} - As^{(k)} \end{aligned}$$

Megjegyzés:

1. Az átmenetmátrix

$$B_{ILU} = P^{-1}Q = P^{-1}(P - A) = I - P^{-1}A.$$

   Legyen P az A-hoz közeli, mert ekkor $\|B_{ILU}\|$ kicsi és így az iteráció gyors.

2. Ha L, U-ban csak kevés nem nulla átló van, akkor az iteráción belüli LER megoldás műveletigénye kicsi.
3. Láttuk, hogy az iteráció végrehajtásakor Q-ra nincs szükségünk.

Általánosítás az *ILU*-algoritmusból:

$$P\,(x^{(k+1)} - x^{(k)}) = r^{(k)} \quad \Leftrightarrow \quad P\,(x^{(k+1)} - x^{(k)}) + A\,x^{(k)} = b.$$

Definíció: általános kétrétegű iterációs eljárás

A

$$P\,(x^{(k+1)} - x^{(k)}) + A\,x^{(k)} = b$$

iterációt *általános kétrétegű iterációs eljárás*nak nevezzük.
P: a *prekondicionáló mátrix*.

Megjegyzés: A korábbi összes iterációs módszerünk ilyen alakú:

$$P\,(x^{(k+1)} - x^{(k)}) + A\,x^{(k)} = b.$$

1. Ha $P = D$, akkor a $J(1)$ iterációt kapjuk.
2. Ha $P = \frac{1}{\omega}\,D$, akkor a $J(\omega)$ iterációt kapjuk.
3. Ha $P = D + L$, akkor az $S(1)$ iterációt kapjuk.
4. Ha $P = D + \omega\,L$, akkor az $S(\omega)$ iterációt kapjuk.
5. Ha $P = \frac{1}{p}\,I$, akkor az $R(p)$ iterációt kapjuk.
6. Ha $P = LU$ az *ILU*-felbontásból, akkor az *ILU* iterációt kapjuk.

Példa:

A korábbi *ILU*-felbontás példákhoz készítsük el a megfelelő *ILU*-algoritmusok átmenetmátrixát és hasonlítsuk össze az egyes iterációk gyorsaságát!

1. Példa:

$$P = \begin{bmatrix} 4 & 0 & 2 \\ 1 & 4 & \frac{1}{2} \\ 2 & 1 & 4 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix}, \quad B_{ILU} = P^{-1}Q,$$

$$\|B_{ILU}\|_2 \approx 0.3601, \quad \|B_{ILU}\|_\infty \approx 0.3438$$

2. Példa: Jacobi-iteráció

$$P = 4\,I, \;\; Q = \begin{bmatrix} 0 & -1 & -2 \\ -1 & 0 & -1 \\ -2 & -1 & 0 \end{bmatrix}, \;\; B_{ILU} = P^{-1}Q = \frac{1}{4}\begin{bmatrix} 0 & -1 & -2 \\ -1 & 0 & -1 \\ -2 & -1 & 0 \end{bmatrix},$$

$$\|B_{ILU}\|_2 \approx 0.6830, \quad \|B_{ILU}\|_\infty \approx 0.75$$

3. Példa: Gauss–Seidel-iteráció

$$P = \begin{bmatrix} 4 & 0 & 0 \\ 1 & 4 & 0 \\ 2 & 1 & 4 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & -1 & -2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_{ILU} = P^{-1}Q,$$

$$\|B_{ILU}\|_2 \approx 0.6408, \quad \|B_{ILU}\|_\infty \approx 0.75$$

Látjuk, hogy az 1. példabeli *ILU*-felbontást alkalmazó *ILU*-algoritmus a leggyorsabb a három közül.

Kerekítési hibák hatása az iterációkra

Tekintsük az iteráció szokásos alakját!

$$x^{(k+1)} = Bx^{(k)} + c$$

Vizsgáljuk meg, hogyan változik az iteráció, ha a $k+1$. lépésben *kicsit* $\varepsilon^{(k)}$*-val* megváltoztatjuk! (Számolási pontatlanság, kerekítési hiba, ...)

1. **Eredeti:**

$$x^{(k+1)} = Bx^{(k)} + c$$

2. **Módosult:**

$$y^{(k+1)} = By^{(k)} + c + \varepsilon^{(k)}$$

Nyilván a lépésenkénti $\varepsilon^{(k)}$ hiba miatt *kicsit* más lesz az iteráció ...

Tétel: a kerekítési hibák hatása az iterációkra

Tegyük fel, hogy

- iterációnk bármely kezdőértékre konvergens,
- a lépésenkénti hiba felülről korlátos, vagyis létezik $\varepsilon > 0$, melyre $\left\|\varepsilon^{(k)}\right\| \leq \varepsilon$ minden k-ra.

Ekkor a $z^{(k)}$ hibasorozatra

$$\lim_{k\to\infty} \left\|z^{(k)}\right\| \leq \frac{\varepsilon}{1-\|B\|}.$$

Biz.: A $z^{(k)} := x^{(k)} - y^{(k)}$ hibavektorra írjuk fel a rekurziót:

$$z^{(k+1)} = x^{(k+1)} - y^{(k+1)} = \left(Bx^{(k)} + c\right) - \left(By^{(k)} + c + \varepsilon^{(k)}\right) =$$

$$= B\left(x^{(k)} - y^{(k)}\right) - \varepsilon^{(k)} = Bz^{(k)} - \varepsilon^{(k)}.$$

Biz. folyt.: A konvergenciából következik, hogy létezik olyan indukált mátrixnorma, melyben $\|B\| < 1$. A hozzá illeszkedő vektornormában becsüljünk:

$$\begin{aligned}
\left\|z^{(k+1)}\right\| &\leq \|B\| \cdot \left\|z^{(k)}\right\| + \left\|\varepsilon^{(k)}\right\| \leq \|B\| \cdot \left\|z^{(k)}\right\| + \varepsilon \leq \\
&\leq \|B\| \left(\|B\| \cdot \left\|z^{(k-1)}\right\| + \varepsilon\right) + \varepsilon \leq \ldots \leq \\
&\leq \|B\|^{k+1} \cdot \left\|z^{(0)}\right\| + \varepsilon \cdot \left(\|B\|^{k} + \ldots + \|B\| + 1\right) < \\
&< \varepsilon \|B\|^{k+1} + \varepsilon \cdot \frac{1}{1 - \|B\|}.
\end{aligned}$$

Innen $k \to \infty$ határátmenettel adódik a bizonyítandó állítás. $\square$

Numerikus módszerek 1.

11. előadás: Nemlineáris egyenletek numerikus megoldása

Krebsz Anna

ELTE IK

Tartalomjegyzék

Feladat

Keressük meg egy $f \in \mathbb{R} \to \mathbb{R}$ nemlineáris függvény gyökét, avagy zérushelyét. ($\exists$?, 1, több?)

$$f(x^*) = 0, \qquad x^* = ?$$

Ekvivalens módon átfogalmazható (általában): keressük meg egy $\varphi \in \mathbb{R} \to \mathbb{R}$ nemlineáris függvény fixpontját.

$$x^* = \varphi(x^*), \qquad x^* = ?$$

Tartalomjegyzék

Lásd Analízis. . .

Tétel: Bolzano-tétel

Ha $f \in C[a; b]$ és $f(a) \cdot f(b) < 0$, akkor $\exists\, x^* \in (a; b) : f(x^*) = 0$.

Megjegyzés:

- $a, b \in \mathbb{R}$, $a < b$, $[a; b]$ zárt intervallum,
- $C[a; b]$: az $[a; b]$ (zárt) intervallumon folytonos függvények halmaza,
- $f(a) \cdot f(b) < 0$: $f(a)$ és $f(b)$ különböző előjelűek
- van gyök az $(a; b)$ (nyílt) intervallumban

Biz. (Bolzano-tétel): az intervallumfelezés módszerével

1. Legyen $x_0 := a,\ y_0 := b$.
2. Ismételjük:
   - Legyen $s_k := \frac{1}{2}(x_k + y_k)$, az intervallum fele.
   - Ha $f(x_k) \cdot f(s_k) < 0$, akkor $x_{k+1} := x_k,\ y_{k+1} := s_k$.
   - Ha $f(x_k) \cdot f(s_k) > 0$, akkor $x_{k+1} := s_k,\ y_{k+1} := y_k$.
3. Álljunk meg, ha
   - egyenlőség teljesül, ekkor $x^* = s_k$, vagy
   - elértük a kívánt pontosságot, ekkor $x^* \in (x_k, y_k)$, és

$$y_k - x_k = \frac{y_{k-1} - x_{k-1}}{2}$$

teljesül. $\square$

Megjegyzés:

- Általában nem tapasztalunk egyenlőséget.
- Az (x_k) és (y_k) sorozatok konvergenciájának részletes tárgyalása: Analízis...
- **Hibabecslések:**

$$|x_k - x^*| < \frac{b-a}{2^k}, \quad |y_k - x^*| < \frac{b-a}{2^k},$$

$$|s_k - x^*| < \frac{b-a}{2^{k+1}}.$$

Példa

Közelítsük a $P(x) = x^3 + 3x - 2$ polinom egyik gyökét 0.1 pontossággal. Hány lépés szükséges?

Próbálkozhatunk a $[0; 1]$ intervallummal...

A $P(x) = x^3 + 3x - 2$ polinom gyökét keressük intervallumfelezéssel a $[0; 1]$ intervallumon:

$$P(0) = -2 < 0, \quad P(1) = 1 + 3 - 2 = 2 > 0$$
$$\Rightarrow \quad \exists\, x^* \in (0; 1) : P(x^*) = 0.$$

Hibabecslés:

$$\frac{1}{2^k} < \frac{1}{10} \quad \Leftrightarrow \quad k > 3,$$

tehát legalább 4 lépésre van szükségünk. Lassú ... $\square$

Tétel: gyök egyértelmű létezéséről

1. Ha $f \in C[a;b]$, $f(a) \cdot f(b) < 0$,
2. valamint $f \in D(a;b)$ és $f' > 0$ (vagy < 0),

akkor $\exists!\ x^* \in (a;b) : f(x^*) = 0$.

Biz.: A Bolzano-tételből következik, hogy van gyök.
f szigorúan monoton, ezért egyértelmű is. $\square$

Tartalomjegyzék

Emlékeztető: Iterációs módszerek LER-ek esetén.

$$Ax = b \qquad \Longleftrightarrow \qquad x = Bx + c$$

$$x^{(k+1)} = \varphi(x^{(k)}) = B \cdot x^{(k)} + c$$

Ötlet: Most, nemlineáris függvények zérushelyéhez:

$$f(x) = 0 \qquad \Longleftrightarrow \qquad x = \varphi(x)$$

$$x_{k+1} = \varphi(x_k) = \dots$$

Emlékeztető: fixpont, kontrakció

Emlékeztető: fixpont

Az $x^* \in \mathbb{R}^n$ pontot a $\varphi\colon \mathbb{R}^n \to \mathbb{R}^n$ leképezés *fixpontjának* nevezzük, ha $x^* = \varphi(x^*)$.

Emlékeztető: kontrakció

A $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ leképezés *kontrakció*, ha $\exists\, q \in [0, 1)$, hogy

$$\|\varphi(x) - \varphi(y)\| \le q \cdot \|x - y\|\,, \qquad \forall x, y \in \mathbb{R}^n.$$

Megj.:

- kontrakció $\approx$ összehúzás, q: kontrakciós együttható
- most $n = 1$, $\|.\| = |.|$; $\mathbb{R}$ helyett $[a; b] \subset \mathbb{R}$, így jobban használható

Definíció: kontrakció

A $\varphi : [a;b] \to \mathbb{R}$ leképezés *kontrakció* $[a;b]$-n, ha $\exists\, q \in [0,1)$, hogy

$$|\varphi(x) - \varphi(y)| \le q \cdot |x - y|\,, \qquad \forall x, y \in [a;b].$$

Állítás

1. $\varphi\colon [a;b] \to \mathbb{R}$ függvény, $\varphi \in C^1[a;b]$ és
2. $|\varphi'(x)| < 1$ $(\forall\, x \in [a;b])$,

akkor φ kontrakció $[a;b]$-n.

Megj.:

- C^1: egyszer folyotnosan differenciálható, vagyis a deriváltja folytonos.
- A kontrakciós tulajdonság függ az intervallumtól.

Biz.: A Lagrange-féle középértéktétel segítségével.

$$q := \max_{x \in [a;b]} \left|\varphi'(x)\right| < 1$$

$\forall x, y \in [a; b]\ (x < y): \quad \exists\, \xi \in (x; y):$

$$|\varphi(x) - \varphi(y)| = \left|\varphi'(\xi)\right| \cdot |x - y| \leq q \cdot |x - y|\,.$$

$\square$

Tétel: Brouwer-féle fixponttétel

1. Ha $\varphi\colon [a;b] \to [a;b]$
2. és $\varphi \in C[a;b]$,

akkor $\exists\, x^* \in [a;b] : \varphi(x^*) = x^*$.

Biz.: Definiáljuk a $g(x) = x - \varphi(x)$ függvényt, majd alkalmazzuk a Bolzano-tételt.

Biz. folyt.:

1. Mivel $\varphi(a), \varphi(b) \in [a; b] \quad \Rightarrow$

$$g(a) = a - \varphi(a) \leq 0, \quad g(b) = b - \varphi(b) \geq 0$$
$$\Rightarrow \quad g(a) \cdot g(b) \leq 0.$$

2. Ha $g(a) \cdot g(b) = 0$, akkor $g(a) = 0$ vagy $g(b) = 0$.
Ez azt jelenti, hogy első esetben a, második esetben b fixpont.

3. Ha $g(a) \cdot g(b) < 0$, akkor a Bolzano-tétel miatt van g-nek gyöke $(a; b)$-ben, azaz

$$\exists\, x^* \in (a; b) : g(x^*) = x^* - \varphi(x^*) = 0 \quad \Leftrightarrow \quad \varphi(x^*) = x^*$$

$\square$

Tétel: Banach-féle fixponttétel $[a; b]$-re

Ha a $\varphi\colon [a; b] \to [a; b]$ függvény kontrakció $[a; b]$-n q kontrakciós együtthatóval, akkor

1. $\exists!\, x^* \in [a; b] : x^* = \varphi(x^*)$, azaz létezik fixpont,
2. $\forall\, x_0 \in [a; b]$ esetén az $x_{k+1} = \varphi(x_k),\ k \in \mathbb{N}_0$ sorozat konvergens és $\lim\limits_{k\to\infty} x_k = x^*$,
3. továbbá a következő hibabecslések teljesülnek:
   - $|x_k - x^*| \le q^k \cdot |x_0 - x^*| \le q^k(b-a)$,
   - $|x_k - x^*| \le \dfrac{q^k}{1-q} \cdot |x_1 - x_0|$.

Biz.: Már volt, csak most $\mathbb{R}^n$ helyett $\mathbb{R}$ $(n = 1)$, sőt $[a; b]$. $\square$

Következmény: iteráció konvergenciájának elégséges feltétele

1. Ha $\varphi\colon [a;b] \to [a;b]$,
2. $\varphi \in C^1[a;b]$ és
3. $|\varphi'(x)| < 1 \quad \forall\, x \in [a;b]$,

akkor az $x_{k+1} = \varphi(x_k)$ iteráció konvergens $\forall x_0 \in [a;b]$ esetén.

Megj.: Attól még lehet konvergens a sorozat, ha valahol $|\varphi'| \geq 1$. (Nem szükséges feltétel.)

Tétel Lokális fixponttétel

Legyen $\varphi\colon [a;b] \to \mathbb{R}$ függvény.

1. Ha $\varphi \in C^1[a;b]$ és
2. $\exists\, \xi \in [a;b]$ és $\delta > 0$, melyre

$$\left|\varphi'(x)\right| \leq q < 1 \quad \forall\, x \in [\xi - \delta; \xi + \delta] \subset [a;b].$$

3. Ha $\exists\, r:\ 0 < r \leq \delta$, melyre

$$|\varphi(\xi) - \xi| \leq (1-q)r,$$

(azaz ξ a fixpont egy elég jó közelítése,)

akkor φ kontrakció $[\xi - r; \xi + r]$-n és

$$\forall\, x \in [\xi - r; \xi + r]:\ \ \varphi(x) \in [\xi - r; \xi + r].$$

Biz.: A tétel feltételeiből következik, hogy φ kontrakció $[\xi-\delta;\xi+\delta]$-n.

Gondoljuk meg, hogy a kontrakciós tulajdonság a $[\xi-r;\xi+r]\subset[\xi-\delta;\xi+\delta]$ részintervallumra is teljesül a q kontrakciós együtthatóval.

Tetszőleges $x\in[\xi-r;\xi+r]$ esetén

$$\begin{aligned}|\varphi(x)-\xi| &= |\varphi(x)-\varphi(\xi)+\varphi(\xi)-\xi|\leq\\ &\leq |\varphi(x)-\varphi(\xi)|+|\varphi(\xi)-\xi|\leq\\ &\leq q\cdot\underbrace{|x-\xi|}_{\leq r}+(1-q)\cdot r=r\end{aligned}$$

Tehát φ az $x\in[\xi-r;\xi+r]$ intervallumba beleképez. $\square$

Következmény:

Ha a lokális fixponttétel feltételei teljesülnek, akkor valójában a Banach-féle fixponttétel feltételei teljesülnek az $[\xi - r; \xi + r]$ intervallumra, így

1. $\exists!\, x^* \in [\xi - r; \xi + r] : x^* = \varphi(x^*)$, azaz létezik fixpont,
2. $\forall\, x_0 \in [\xi - r; \xi + r]$ esetén az $x_{k+1} = \varphi(x_k),\ k \in \mathbb{N}_0$ sorozat konvergens és $\lim\limits_{k\to\infty} x_k = x^*$,
3. továbbá a következő hibabecslések teljesülnek:
   - $|x_k - x^*| \le q^k \cdot |x_0 - x^*| \le q^k(b-a)$,
   - $|x_k - x^*| \le \dfrac{q^k}{1-q} \cdot |x_1 - x_0|$.

1. Példa

A zsebszámológépünkbe írjunk be egy 0 és 1 közötti számot, majd nyomjuk meg az x^2 billentyűt. A sokadik gombnyomás után mit tapasztalunk?

Valójában az $x = x^2$ egyenlet fixpontját keressük az

$$x_{k+1} = x_k^2$$

iterációval. Két fixpontja van 0 és 1, de

- $0 \leq x_0 < 1$ esetén $\lim(x_k) = 0$.
- $x_0 = 1$ esetén $\lim(x_k) = 1$.

2. Példa

A zsebszámológépünkbe írjunk be egy 0 és 1 közötti számot, majd nyomjuk meg a $\sqrt{x}$ billentyűt. A sokadik gombnyomás után mit tapasztalunk?

Valójában az $x = \sqrt{x}$ egyenlet fixpontját keressük az

$$x_{k+1} = \sqrt{x_k}$$

iterációval. Két fixpontja van 0 és 1, de

- $x_0 = 0$ esetén $\lim(x_k) = 0$.
- $0 < x_0 \leq 1$ esetén $\lim(x_k) = 1$.

3. Példa

A zsebszámológépünkbe írjunk be egy 0 és 1 közötti számot, majd nyomjuk meg a $\cos(x)$ billentyűt. A sokadik gombnyomás után mit tapasztalunk?

Valójában az $x = \cos(x)$ fixpontegyenlet megoldását keressük a $[0, 1]$ intervallumon az

$$x_{k+1} := \cos(x_k), \; x_0 \in [0, 1]$$

iterációval. Egyértelmű-e a megoldás? Konvergens ez a sorozat? Adjunk hibabecslést! Hány lépés után kapjuk a megoldást 0.1-es pontossággal?

1. Belátjuk, hogy a $\varphi(x) := \cos(x)$ függvény a $[0;1]$ intervallumot a $[0;1]$-be képezi:
   - Mivel $\varphi'(x) = -\sin(x) < 0$, $\forall\, x \in [0;1]$, ezért φ szigorúan monoton fogyó $[0;1]$-en.
   - $\varphi([0;1]) = [\varphi(1);\varphi(0)] = [\cos(1), 1] \subset [0;1]$, tehát $\varphi:\ [0;1] \to [0;1]$.

2. Belátjuk, hogy a $\varphi(x) = \cos(x)$ függvény kontrakció $[0;1]$-en. Tetszőleges $x, y \in [0;1]$-re a Lagrange-középértéktételt alkalmazva $\exists\, \xi \in (0;1)$, melyre

$$|\varphi(x) - \varphi(y)| = |\varphi'(\xi)| \cdot |x - y| \leq q \cdot |x - y|\,,$$

ahol a kontrakciós együttható $q := \max_{\xi \in [0;1]} |-\sin(\xi)| = \sin(1) \approx 0.8415 < 1.$

3. A Banach-féle fixponttétel feltételei teljesülnek, így annak állításai felhasználhatóak, ezzel a fixpont létezését, egyértelműségét és a konvergenciát beláttuk.

4. Hibabecslése:

$$|x_k - x^*| \leq 0.8415^k \cdot \underbrace{|x_0 - x^*|}_{<1} \leq 0.8415^k.$$

5. A megadott pontosság eléréséhez szükséges lépésszám:

$$0.8415^k < \frac{1}{10} \quad \Leftrightarrow \quad k > \frac{-1}{\lg(0.8415)} \approx 13.34.$$

Nagyon lassú ...

Definíció: konvergencia rend

Az (x_k) konvergens sorozat – határértékét jelölje x^* – *p-edrendben konvergens*, ha $\exists c \in (0; +\infty) \subset \mathbb{R}$, hogy

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = c.$$

Megjegyzés:

- p egyértelmű, $p \geq 1$,
- p nem feltétlenül egész (A szelőmódszernél $p = \frac{1+\sqrt{5}}{2}$.)
- $p = 1$: elsőrendű vagy lineáris konvergencia (ekkor $c \leq 1$)
  $p = 2$: másodrendű vagy kvadratikus konvergencia
- $p > 1$: szuperlineáris konvergencia

Megjegyzés folyt.:

- Gyakorlatban a legalább p-edrendű konvergencia megfogalmazása:

$$\exists K \in \mathbb{R}^+ : \forall k \in \mathbb{N}_0 : |x_{k+1} - x^*| \leq K \cdot |x_k - x^*|^p$$

- A fixponttételek nem mondanak konvergencia rendet. (Csak annyit, hogy legalább elsőrendű.)
- Ha $c = 0$, akkor a keresett konvergencia rend nagyobb a megadottnál.
- Ha $c = \infty$, akkor a keresett konvergencia rend kisebb a megadottnál.

Példa

Mennyi a konvergenciarendje a következő nullsorozatoknak?

$$\left(\frac{1}{n^2}\right); \qquad \left(\frac{1}{2^n}\right); \qquad (q^n)\ \ (|q| < 1); \qquad \left(\frac{1}{2^{2^n}}\right);$$

Vizsgáljuk az egyik sorozatot, a többit gyakorlaton..

Példa

Tekintsük az $(x_k) = \left(\frac{1}{2^k}\right)$, $(k \in \mathbb{N})$ sorozatot.

1. Tippeljük $p = 2$-re a konvergencia rendet:

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = \lim_{k\to\infty} \frac{\left|\frac{1}{2^{k+1}} - 0\right|}{\left|\frac{1}{2^k} - 0\right|^2} = \lim_{k\to\infty} \frac{2^{2k}}{2^{k+1}} = \lim_{k\to\infty} 2^{k-1} = \infty.$$

   Látjuk, hogy a határérték ∞, vagyis kisebb p-vel kell próbálkoznunk.

2. Tippeljük $p = 1$-re a konvergencia rendet.

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = \lim_{k\to\infty} \frac{\left|\frac{1}{2^{k+1}} - 0\right|}{\left|\frac{1}{2^k} - 0\right|} = \lim_{k\to\infty} \frac{2^k}{2^{k+1}} = \lim_{k\to\infty} \frac{1}{2} = \frac{1}{2}.$$

   Látjuk, hogy a határértek rendben van, a konvergencia elsőrendű.

Mit jelent az első- és másodrendű konvergencia számokban? ($\sqrt{2}$)

1. $p = 1$, $|x_{k+1} - x^*| \leq K \cdot |x_k - x^*|^1$

   $\underline{1.414}184570312500$

   $\underline{1.4142}45605468750$

   $\underline{1.41421}5087890625$

   Minden lépésben kb. egy újabb tizedesjegy pontos.

2. $p = 2$, $|x_{k+1} - x^*| \leq K \cdot |x_k - x^*|^2$

   $\underline{1.41}6666666666667$

   $\underline{1.41421}5686274510$

   $\underline{1.41421356237}4690$

   Minden lépésben kb. kétszer annyi tizedesjegy pontos.

Tétel: *p*-edrendben konvergens iterációk

1. Legyen $\varphi\colon \mathbb{R} \to \mathbb{R}$, $\varphi \in C^p[a;b]$ és
2. az $x_{k+1} = \varphi(x_k)$ sorozat konvergens, határértéke x^*.
3. Ha $\varphi'(x^*) = \cdots = \varphi^{(p-1)}(x^*) = 0$, de $\varphi^{(p)}(x^*) \neq 0$,

akkor a konvergencia p-edrendű és hibabecslése:

$$\left|x_{k+1} - x^*\right| \leq \frac{M_p}{p!} \left|x_k - x^*\right|^p,$$

ahol $M_p = \max\limits_{\xi \in [a;b]} \left|\varphi^{(p)}(\xi)\right|$.

Biz.: Írjuk fel a φ függvény x^* körüli Taylor-polinomját a maradéktaggal.

$$\exists\, \xi \in (x, x^*) \ \ (\text{vagy} \ \ (x^*, x)):$$

$$\varphi(x) = \varphi(x^*) + \varphi'(x^*)(x - x^*) + \dots + \frac{\varphi^{(p-1)}(x^*)}{(p-1)!}(x - x^*)^{p-1} +$$

$$+ \frac{\varphi^{(p)}(\xi)}{p!}(x - x^*)^p$$

Vizsgáljuk ezt az $x = x_k$ helyen, kihasználva a deriváltak zérus voltát is. ($\exists \xi_k$):

$$x_{k+1} = \varphi(x_k) = \underbrace{\varphi(x^*)}_{x^*} + \frac{\varphi^{(p)}(\xi_k)}{p!}(x_k - x^*)^p$$

Biz. folyt.: átrendezve

$$x_{k+1} - x^* = \frac{\varphi^{(p)}(\xi_k)}{p!}(x_k - x^*)^p.$$

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = \lim_{k\to\infty} \frac{\left|\varphi^{(p)}(\xi_k)\right|}{p!} = \frac{\left|\varphi^{(p)}(x^*)\right|}{p!} \neq 0.$$

Tehát (x_k) egy p-adrendben konvergens sorozat.

Vegyük szemügyre a $k+1$-edik és a k-adik tag hibáját.

$$|x_{k+1} - x^*| = \frac{\left|\varphi^{(p)}(\xi_k)\right|}{p!} \cdot |x_k - x^*|^p \leq \frac{M_p}{p!} |x_k - x^*|^p ,$$

ahol $M_p = \max_{\xi\in[a,b]} \left|\varphi^{(p)}(\xi)\right|$. $\square$

Következmény

1. Ha $\varphi\colon [a;b] \to [a;b]$ kontrakció,
2. x^* a φ fixpontja és
3. $\varphi'(x^*) = \cdots = \varphi^{(p-1)}(x^*) = 0$, de $\varphi^{(p)}(x^*) \neq 0$,

akkor

1. a fixpont egyértelmű,
2. $\forall x_0 \in [a;b]$ esetén az $x_{k+1} = \varphi(x_k),\ k \in \mathbb{N}_0$ sorozat konvergens és $\lim\limits_{k\to\infty} x_k = x^*$,
3. és a következő hibabecslés teljesül: $\left|x_{k+1} - x^*\right| \leq \frac{M_p}{p!} \left|x_k - x^*\right|^p$.

Biz.: Ez a Banach-féle fixponttétel és a p-edrendben konvergens iterációk tételének összeházasításaként adódik. $\square$

Még egy példa egyszerű iterációra

Példa

Írjunk fel fixpont-iteráció(ka)t az $x^3 - x - 1 = 0$ egyenlet megoldására, bizonyítsuk a konvergenciát.

a) $x = x^3 - 1$,

b) $x = \sqrt[3]{x + 1}$.

Lásd gyakorlat...
A két sorozat közül az egyik konvergens, a másik divergens.
Melyik-melyik? Milyen intervallumon konvergens? Indokoljuk.

Tartalomjegyzék

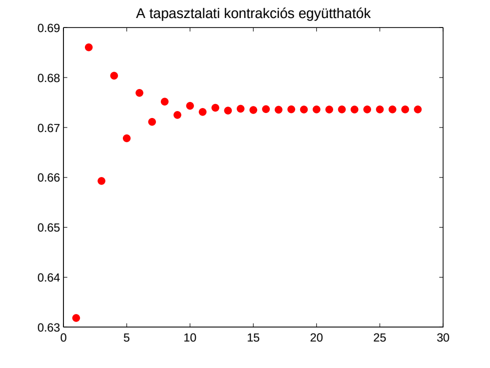
Példák Matlab-ban

1. Intervallumfelezés számolása és szemléltetése.
2. Egyszerű iterációk és fixpontok elemzése az $x = \cos(x)$ egyenlet példáján keresztül.
3. Tapasztalati kontrakciós együtthatók szemléltetése.
4. $\sqrt{2}$ közelítése különböző iterációkkal ($p = 1, 2, 3$ rendűek).
5. A logisztikus leképezés viselkedésének bemutatása érdekességképpen.

Tapasztalati kontrakciós együttható vizsgálata

1. Példa:

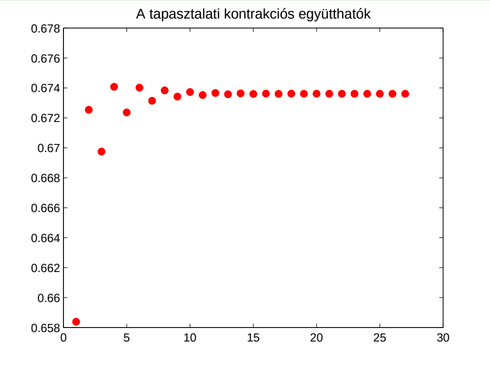
$$x_{k+1} := \cos(x_k),\ x_0 \in [0,1]$$



$q \approx 0.6736$

Tapasztalati kontrakciós együttható vizsgálata

1. Példa:



Az egymást követő tapasztalati kontrakciós együtthatók mértani középét rajzoltuk ki. $q \approx 0.6736$

2. Példa:

Matlab segítségével vizsgáljuk a következő sorozatokat:

1. A $\sqrt{2}$ lánctörtkifejtéséből: $(p = 1)$

$$x_{k+1} = 1 + \frac{1}{1 + x_k}.$$

2. Az $f(x) = x^2 - 2$ függvényre alkalmaztuk a Newton-módszert, analízisből ismerős lehet... $(p = 2)$

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

3. Másodfokú Taylor-polinom közelítéssel: $(p = 3)$

$$x_{k+1} = x_k \cdot \frac{x_k^2 + 6}{3x_k^2 + 2}.$$

Logisztikus leképezés

Az ökológusok gyakran vizsgálnak olyan - időszakosan szaporodó - populációkat (pl. gyümölcsöskerti kártevők), amelyekben nincs átfedés az egyes generációk között. A kutatások célja ilyenkor annak megértése, hogy az $n+1$-edik generáció számossága (N_{n+1}) hogyan függ az előző, n-edik generáció számosságától (N_n). Az ismert tendenciát figyelembe véve, nevezetesen, hogy az utódok száma (N_{n+1}) általában nő, ha a populáció számossága kicsi, és csökken, ha N_n értéke nagy, egy egyszerű nemlineáris differenciaegyenletet írhatunk fel:

$$N_{n+1} = kN_n - bN_n^2 = N_n(k - bN_n),$$

amelyet logisztikus differenciaegyenletnek neveznek, és amelyben k és b a populációk növekedésének, illetve csökkenésének mértékét megszabó paraméterek.

$$N_{n+1} = kN_n\left(1 - \frac{bN_n}{k}\right) \;\Leftrightarrow\; \frac{bN_{n+1}}{k} = k\frac{bN_n}{k}\left(1 - \frac{bN_n}{k}\right)$$

Az $x_n = bN_n/k$ jelölést bevezetve az egyenlet a következő egyszerű alakra hozható:

$$x_{n+1} = kx_n(1 - x_n),$$

amit logisztikus leképezésnek nevezünk.

A logisztikus leképezés egyik nagy előnye az, hogy $1 < k < 4$ esetén a megoldás mindig a $0 < x < 1$ intervallumban marad. A $k < 1$ esetben az összes megoldás az $x = 0$ ponthoz tart, azaz a populáció kihal.

k értéke és a megfigyelt dinamikai viselkedés:

- 3.0000 : a fixpont instabilissá válik, megjelenik az oszcilláció
- 3.4500 : a periódusketttőződés kezdete
- 3.5700 : a $2n$ periódusú oszcillációk torlódási pontja, a kaotikus tartomány kezdete
- 3.6786 : az első páratlan periódusú oszcilláció megjelenése
- 3.8284 : a háromperiódusú oszcilláció megjelenése
- 4.0000 : a kaotikus tartomány vége.

Irodalom: Gáspár Vilmos: Játsszunk káoszt! (Természet Világa cikk)

Példa

Vizsgáljuk meg az $x_0 \in [0,1]$, $x_{k+1} = \alpha \cdot x_k(1-x_k)$ iterációk (logisztikus leképezés) viselkedését különböző $\alpha \in [0,4]$ paraméterek esetén.

Megj.: Általában nem kontrakció. Könnyen eljuthatunk differenciaegyenletek bifurkációinak és a káoszelmélet alapjainak vizsgálatához. . .

Numerikus módszerek 1.

12. előadás: A Newton-módszer és társai

Krebsz Anna

ELTE IK

Tartalomjegyzék

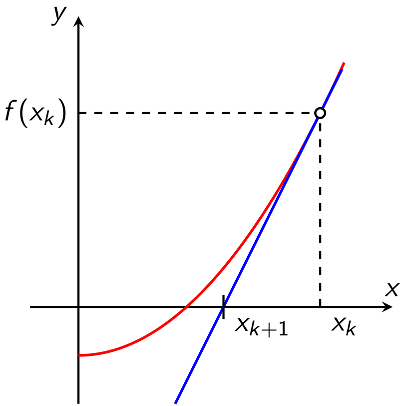
Feladat

Keressük meg egy $f : \mathbb{R} \to \mathbb{R}$ nemlineáris függvény gyökét, avagy zérushelyét. ($\exists$?, 1, több?)

$$f(x^*) = 0, \qquad x^* = ?$$

Geometriai megközelítés:

$$f, x_k \to \text{érintő} \to \text{zérushely (y=0)} \to x_{k+1}$$



Az érintő egyenlete:

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

Analitikus megközelítés:

$$f \text{ gyöke} \approx x_k \text{ körüli Taylor-polinomának gyöke}$$

$$0 = f(x) = f(x_k) + f'(x_k) \cdot (x - x_k) + \ldots$$

Definíció: Newton-módszer

Adott $f : \mathbb{R} \to \mathbb{R}$ differenciálható függvény és $x_0 \in \mathbb{R}$ kezdőpont esetén a *Newton-módszer* alakja:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \qquad (k = 0, 1, 2, \ldots).$$

Példa

Írjuk fel a Newton-módszert a $\sqrt{2}$ értékének közelítésére, és számoljuk ki a közelítő sorozat első néhány elemét valamely kezdőpontból!

Megj.: babiloni módszer ($\sqrt{n}$ számítása).

Általában másodrendben konvergens!

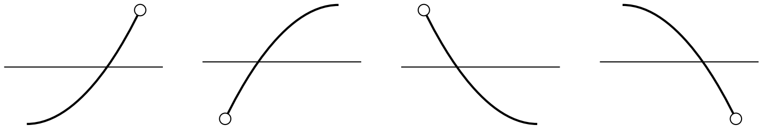
Tétel: monoton konvergencia tétele

Ha $f \in C^2[a;b]$ és

1. $\exists\, x^* \in [a;b] : f(x^*) = 0$, azaz van gyök,
2. f' és f'' állandó előjelű,
3. $x_0 \in [a;b] : f(x_0) \cdot f''(x_0) > 0$,

akkor az x_0 pontból indított Newton-módszer (által adott (x_k) sorozat) monoton konvergál x^*-hoz.

Megj.: 4 eset van:



Biz.: Csak az $f' > 0,\ f'' > 0$ esetre (a többi hasonló)
$\Rightarrow \quad f(x_0) > 0.$

1. Taylor-formula másodfokú maradéktaggal, x_k középponttal:
   $\exists \xi_k \in (x, x_k)$ vagy (x_k, x):

$$f(x) = f(x_k) + f'(x_k) \cdot (x - x_k) + \frac{f''(\xi_k)}{2} \cdot (x - x_k)^2.$$

Az x_{k+1} helyen: $\exists \xi_k \in (x_{k+1}, x_k)$ vagy (x_k, x_{k+1})

$$f(x_{k+1}) = \underbrace{f(x_k) + f'(x_k) \cdot (x_{k+1} - x_k)}_{=0 \text{ (def. alapján)}} + \underbrace{\frac{f''(\xi_k)}{2}}_{>0} \cdot \underbrace{(x_{k+1} - x_k)^2}_{>0}.$$

Tehát $f(x_k) > 0 \ (\forall k \in \mathbb{N})$.

Newton-módszer – monoton konvergencia

2. Az (x_k) sorozat monoton fogyó,

$$x_{k+1} = x_k - \underbrace{\frac{f(x_k)}{f'(x_k)}}_{>0} < x_k;$$

valamint az (x_k) sorozat alulról korlátos,

$$0 = f(x^*) < f(x_k),\ f \text{ szig. mon. nő} \quad \Longrightarrow \quad x^* < x_k$$

így az (x_k) sorozat konvergens, $\hat{x} := \lim\limits_{k\to\infty} x_k$.

3. Kell: $\hat{x} = x^*$. Elég: $f(\hat{x}) = 0$. ($f \in C[a;b]$, f szig. mon.)

$$f(\hat{x}) = \lim_{k\to\infty} f(x_{k+1}) = \lim_{k\to\infty} \underbrace{\frac{f''(\xi_k)}{2}}_{\text{korlátos}} \cdot \underbrace{(x_{k+1} - x_k)^2}_{\to 0 \text{ (Cauchy)}} = 0. \quad \square$$

Tétel: lokális konvergencia tétele

Ha $f \in C^2[a;b]$ és

1. $\exists\, x^* \in [a;b] : f(x^*) = 0$, azaz van gyök,
2. f' állandó előjelű,
3. $m_1 = \min\limits_{x \in [a;b]} |f'(x)| > 0$,
4. $M_2 = \max\limits_{x \in [a;b]} |f''(x)| < +\infty$, innen $M = \frac{M_2}{2 \cdot m_1}$.
5. $x_0 \in [a;b] : |x_0 - x^*| < r := \min\left\{ \frac{1}{M}, |x^* - a|, |x^* - b| \right\}$,

akkor az x_0 pontból indított Newton-módszer másodrendben konvergál a gyökhöz, és az

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

hibabecslés érvényes.

Röviden: Ha elég közelről indulunk, akkor gyorsan odatalálunk.

Megjegyzés:

- $|x_0 - x^*| < r := \min\left\{ \frac{1}{M}, |x^* - a|, |x^* - b| \right\}$, azaz legyünk „elég közel”, de azért mindenesetre legyünk $[a; b]$-n belül is.
- A monoton konvergencia feltételeinek esetén is másodrendű lesz a konvergencia, hiszen előbb-utóbb „elég közel” kerülünk a gyökhöz.

Biz.:

1. Alkalmazzuk az f függvényre a Taylor-formulát, x_k középppponttal az x^* helyen, másodfokú maradéktaggal. $\exists \xi_k \in (x_k, x^*)$ (vagy (x^*, x_k)):

$$0 = f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{f''(\xi_k)}{2}(x^* - x_k)^2.$$

2. Mindkét oldalt $f'(x_k)$-val osztva, majd átrendezve és a Newton-módszer képletét felismerve kapjuk, hogy

$$0 = \frac{f(x_k)}{f'(x_k)} + x^* - x_k + \frac{f''(\xi_k)}{2 \cdot f'(x_k)}(x^* - x_k)^2,$$

$$\left(x_k - \frac{f(x_k)}{f'(x_k)}\right) - x^* = x_{k+1} - x^* = \frac{f''(\xi_k)}{2 \cdot f'(x_k)}(x^* - x_k)^2,$$

$$|x_{k+1} - x^*| \leq \frac{M_2}{2 \cdot m_1} \cdot |x_k - x^*|^2 = M \cdot |x_k - x^*|^2,$$

ahol M, m_1, M_2 a tételben definiált mennyiségek.

Newton-módszer – lokális konvergencia

3. Bevezetve az $\varepsilon_k := x_k - x^*$ jelölést, így is írhatjuk:

$$|\varepsilon_{k+1}| \leq M \cdot |\varepsilon_k|^2 .$$

Ezzel beláttuk, hogy ha (x_k) konvergál és határértéke x^*.

4. A Taylor-formából

$$\frac{|x_{k+1} - x^*|}{|x_k - x^*|^2} = \frac{|f''(\xi_k)|}{2\,|f'(x_k)|}.$$

Határértéket véve

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^2} = \lim_{k\to\infty} \frac{|f''(\xi_k)|}{2\,|f'(x_k)|} = \frac{|f''(x^*)|}{2\,|f'(x^*)|} \neq 0,$$

tehát legalább másodrendben konvergens a sorozat.

Newton-módszer – lokális konvergencia

5. Teljes indukcióval belátjuk, hogy a sorozat minden tagja a $K_r(x^*)$ környezetben marad. $|x_0 - x^*| < r$ feltétel volt. Tegyük fel, hogy $|x_k - x^*| = |\varepsilon_k| < r \leq \frac{1}{M}$, ekkor

$$|\varepsilon_{k+1}| = |x_{k+1} - x^*| \leq M \cdot |\varepsilon_k|^2 = \underbrace{(M|\varepsilon_k|)}_{<1} \cdot |\varepsilon_k| < |\varepsilon_k| < r.$$

6. A konvergencia bizonyításához belátjuk, hogy az $|\varepsilon_k|$ hibakorlátok sorozata 0-hoz tart. Bevezetjük a $d_k := M \cdot |\varepsilon_k|$ jelölést.

$$|\varepsilon_{k+1}| \leq M \cdot |\varepsilon_k|^2 \quad \Longrightarrow \quad M \cdot |\varepsilon_{k+1}| \leq (M \cdot |\varepsilon_k|)^2 \quad \Longrightarrow$$

$$d_{k+1} \leq d_k^2 \quad \Longrightarrow \quad d_k \leq d_{k-1}^2 \leq d_{k-2}^{2\cdot 2} \leq \ldots \leq d_0^{2^k},$$

$$M \cdot |\varepsilon_k| \leq (M \cdot |\varepsilon_0|)^{2^k} \quad \Longrightarrow \quad |\varepsilon_k| \leq \frac{1}{M} \cdot (M \cdot |\varepsilon_0|)^{2^k}.$$

7. Mivel $|\varepsilon_0| = |x_0 - x^*| < \frac{1}{M}$, így $M \cdot |\varepsilon_0| < 1$, ezért $|\varepsilon_k| \to 0$, ami az (x_k) sorozat konvergenciáját jelenti. □

Megjegyzés:

- Ha $f'(x_k) = 0$, akkor x_{k+1} nincs értelmezve.
- Néha a konvergencia csak elsőrendű (vagy instabillá válik). Például ha $f'(x^*) = 0$, azaz x^* többszörös gyök. A Newton-módszerrel x^* közelében $\frac{0}{0}$ alakú osztást végzünk.
- Többszörös gyök esetén például alkalmazzuk a $g(x) := \frac{f(x)}{f'(x)}$ függvényre a Newton-módszert.
- Másik lehetőség: ha x^* r-szeres gyök, akkor az

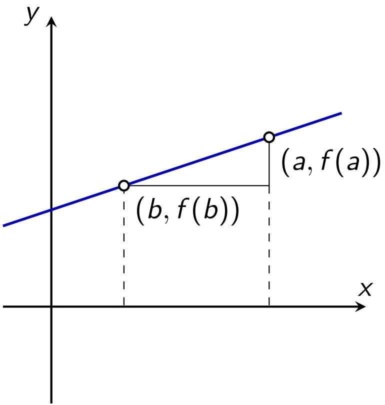
  $$x_{k+1} := x_k - r \cdot \frac{f(x_k)}{f'(x_k)}$$

  módosítást használjuk, amivel másodrendű iterációt kapunk.
- Néha akár harmadrendű is lehet (v.ö. magasabbrendű konvergencia tétel).

Megjegyzés folyt.:

- Használhattuk volna a magasabbrendű konvergencia tételt is a Newton-módszer lokális konvergencia tételének bizonyítására a $\varphi(x) := x - \frac{f(x)}{f'(x)}$ megfeleltetéssel, de akkor $f \in C^3[a; b]$-t kellett volna feltennünk.
- Hívják Newton–Raphson-, ill. Newton–Fourier-módszernek is.
- A módszer nem biztos, hogy konvergál.
- Ciklusba is kerülhet (pontos számolás esetén...).
- A gyökök „vonzásterületein” kívül kaotikus jelenségek. . .

Ismétlés: Két adott ponton átmenő egyenes egyenlete.



Az egyenes meredeksége:

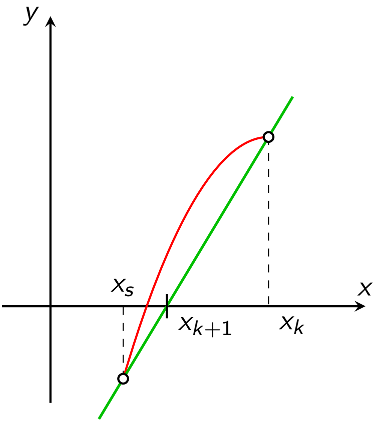
$$\frac{f(a) - f(b)}{a - b}.$$

Az egyenes egyenlete:

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

Ennek zérushelye ($y = 0$):

$$x = a - \frac{f(a) \cdot (a - b)}{f(a) - f(b)}.$$



Definíció: húrmódszer

Az $f \in C[a;b]$ függvény esetén, ha $f(a) \cdot f(b) < 0$, akkor a *húrmódszer* alakja:

$$x_0 := a, \quad x_1 := b,$$

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

$$(k = 0, 1, 2, \dots),$$

ahol s a legnagyobb olyan index, amelyre $f(x_k) \cdot f(x_s) < 0$.

Tétel: a húrmódszer konvergenciája

Ha $f \in C^2[a;b]$ és

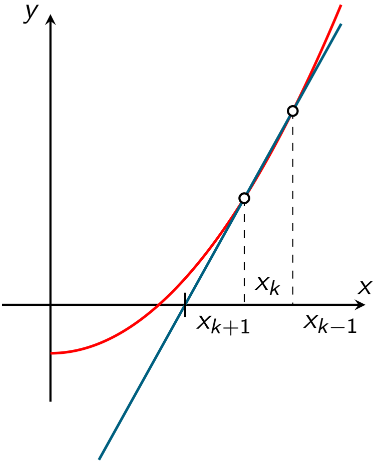
1. $f(a) \cdot f(b) < 0$,
2. $M \cdot (b-a) < 1$,

akkor a húrmódszer elsőrendben konvergál az x^* gyökhöz és

$$|x_k - x^*| \leq \frac{1}{M} \cdot (M \cdot |x_0 - x^*|)^k$$

teljesül, ahol $M = \frac{M_2}{2 \cdot m_1}$ ugyanúgy, mint korábban.

Biz.: nélkül.



Definíció: szelőmódszer

Az $f \in C[a;b]$ függvény esetén a *szelőmódszer* alakja:

$$x_0, x_1 \in [a;b],$$

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

$$(k = 0, 1, 2, \dots).$$

Tétel: a szelőmódszer konvergenciája

Ha $f \in C^2[a;b]$ és

1. $\exists x^* \in [a;b] : f(x^*) = 0$, azaz van gyök,
2. f' állandó előjelű,
3. $x_0, x_1 \in [a;b]$:

$$\left.\begin{array}{l} |x_0 - x^*| \\ |x_1 - x^*| \end{array}\right\} < r := \min\left\{ \frac{1}{M}, |x^* - a|, |x^* - b| \right\},$$

akkor a szelőmódszer $p = \dfrac{1+\sqrt{5}}{2}$ rendben konvergál az x^* gyökhöz. (M a szokásos.)

Biz.: nélkül.

Többváltozós nemlineáris egyenletrendszerek

Feladat

$$F\colon \mathbb{R}^n \to \mathbb{R}^n, \quad F(x) = 0, \qquad x = ?, \quad (x \in \mathbb{R}^n)$$

Legtöbb módszerünk általánosítható többváltozós esetre.

Egyszerű iteráció

$$F(x) = 0 \quad \Longleftrightarrow \quad x = \Phi(x)$$

Banach-féle fixponttétel szerint. . .

Többváltozós Newton-módszer

Közelítsük F-et az elsőfokú Taylor-polinomjával.

$$F(x) \approx F(x^{(k)}) + F'(x^{(k)}) \cdot (x - x^{(k)}),$$
$$F'(x^{(k)}) = \left(\frac{\partial f_i(x^{(k)})}{\partial x_j} \right)_{i,j=1}^{n} \in \mathbb{R}^{n \times n}$$

Ezen közelítés zérushelye lesz $x^{(k+1)}$:

1. $F'(x^{(k)}) \cdot \underbrace{(x^{(k+1)} - x^{(k)})}_{s^{(k)}} = -F(x^{(k)})$ LER megoldás ($\rightsquigarrow s^{(k)}$),
2. $x^{(k+1)} = x^{(k)} + s^{(k)}$, $s^{(k)}$ a továbblépés iránya.

Definíció: a többváltozós Newton-módszer képlete

$$x^{(k+1)} = x^{(k)} - \left(F'(x^{(k)})\right)^{-1} \cdot F(x^{(k)})$$

Megj.: A módszer javítható pl. úgy, hogy ne kelljen minden lépésben invertálni és deriváltat számolni $\rightsquigarrow$ Broyden-módszer (lassabb).

Példák Matlab-ban

1. A $\sqrt{2}$ értékének másodrendben konvergens közelítése.
2. Példák a Newton-módszer működésére: konvergencia, divergencia, ciklizálás, fraktálszerű jelenségek.

Példa:

Alkalmazzuk a következő kétváltozós függvényre a Newton-módszert!

$$F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad F : \mathbb{R}^2 \to \mathbb{R}^2,$$

ahol $f_1(x) = x_1^2 + x_2^2 - 1$, $f_2(x) = -x_1^2 - x_2$.
Geometriailag egy fordított parabola és az origó körüli egy sugarú kör metszéspontját keressük.

Megj.:

- Bizonyos pontokban a Newton-módszer nem értelmezett, mert $\det(f'(x^{(k)})) = 0$.

  $$\det(F'(x)) = \begin{vmatrix} 2x_1 & 2x_2 \\ -2x_1 & -1 \end{vmatrix} = -2x_1 + 4x_1x_2 = 2x_1(2x_2 - 1) = 0$$

  $x_1 = 0$ és $x_2 = 0.5$ esetén a módszer nem értelmezett.
- Divergens például $x_0 = \begin{bmatrix} \pm 1 & 1 \end{bmatrix}^T$-ből úgy, hogy az első koordináta sorozat konvergens (de a határérték rossz).

Numerikus módszerek 1.

13. előadás: Polinomokról: gyökök becslése, Horner-algoritmus

Krebsz Anna

ELTE IK

Tartalomjegyzék

Tartalomjegyzék

Vizsgáljunk n-edfokú polinomokat, melyek alakja:

$$P(x) = a_n \cdot x^n + a_{n-1} \cdot x^{n-1} + \cdots + a_1 \cdot x + a_0$$
$$a_i \in \mathbb{R}, \quad a_0 \neq 0, \quad a_n \neq 0.$$

Megjegyzés:

- Akár $a_i \in \mathbb{C}$ is lehet...
- Ha $a_0 = 0$, akkor az $x = 0$ gyök, leoszthatunk x-szel $\rightsquigarrow$ egyszerűbb polinomot vizsgálhatunk.
- Ha $a_n = 0$, akkor nem is n-edfokú...

Példa

Vizsgáljuk meg néhány polinom gyökeinek elhelyezkedését.
Komplex gyökök is szóba jöhetnek.

Tétel: Becslés polinom gyökeinek elhelyezkedésére

A $P(x) = a_n \cdot x^n + a_{n-1} \cdot x^{n-1} + \cdots + a_1 \cdot x + a_0$ polinom esetén, ha $a_0 \neq 0$ és $a_n \neq 0$, akkor P bármely x_k gyökére:

$$r < |x_k| < R,$$

ahol

$$R = 1 + \frac{\max_{i=0}^{n-1} |a_i|}{|a_n|}, \quad r = \frac{1}{1 + \frac{\max_{i=1}^{n} |a_i|}{|a_0|}}.$$

Megjegyzés: Ezzel a gyökök elhelyezkedésére egy origó középpontú nyílt körgyűrűt adtunk meg a komplex számsíkon.

Biz.:

1. Megmutatjuk, hogy ha $|x| \geq R$ (x a külső körön kívül van), akkor $|P(x)| > 0$ (x nem gyöke P-nek). A becsléshez a kétféle háromszög-egyenlőtlenséget használjuk:

$$|P(x)| \geq |a_n x^n| - \left|a_{n-1}x^{n-1} + \ldots a_1 x + a_0\right|$$

   A továbbiakban lefelé akarunk becsülni, így a kivonandó összeget növelnünk kell:

$$\left|a_{n-1}x^{n-1} + \ldots + a_0\right| \leq |a_{n-1}| \cdot |x|^{n-1} + \ldots + |a_0| \leq$$

$$\leq \left(\max_{i=0}^{n-1} |a_i|\right) \cdot \left(|x|^{n-1} + \ldots + 1\right) = \left(\max_{i=0}^{n-1} |a_i|\right) \cdot \frac{|x|^n - 1}{|x| - 1} <$$

$$< \left(\max_{i=0}^{n-1} |a_i|\right) \cdot \frac{|x|^n}{|x| - 1}.$$

Biz. folyt: Folytassuk $|P(x)|$ becslését és vizsgáljuk meg, mikor pozitív.

$$|P(x)| > |a_n| \cdot |x|^n - \left(\max_{i=0}^{n-1} |a_i| \right) \cdot \frac{|x|^n}{|x| - 1} \geq 0$$

Rendezzük át az egyenlőtlenséget, szorozzunk be $|x| - 1 > 0$-val és osszunk le $|a_n| \cdot |x|^n$-vel

$$|P(x)| > 0 \quad \Leftrightarrow \quad |a_n| \cdot |x|^n \geq \left(\max_{i=0}^{n-1} |a_i| \right) \cdot \frac{|x|^n}{|x| - 1} \quad \Leftrightarrow$$

$$|x| - 1 \geq \left(\max_{i=0}^{n-1} |a_i| \right) \cdot \frac{|x|^n}{|a_n| \cdot |x|^n} \quad \Leftrightarrow$$

$$|x| \geq 1 + \frac{\max_{i=0}^{n-1} |a_i|}{|a_n|} =: R.$$

Biz. folyt: Azt kaptuk, hogy ha $|x| \geq R$, akkor $|P(x)| > 0$, vagyis x nem gyök. Ezzel beláttuk a tétel első felét.

2. Az alsó becslést úgy nyerjük, hogy az imént belátott becslést alkalmazzuk $P(x)$ reciprok-polinomjára.
   Vezessük be az $y := \frac{1}{x}$ új változót ($x \neq 0$):

$$P(x) = P\left(\frac{1}{y}\right) = a_n\left(\frac{1}{y}\right)^n + a_{n-1}\left(\frac{1}{y}\right)^{n-1} + \ldots + a_1\left(\frac{1}{y}\right) + a_0 =$$
$$= \left(\frac{1}{y}\right)^n \cdot \underbrace{\left(a_n + a_{n-1}y + \ldots + a_1 y^{n-1} + a_0 y^n\right)}_{Q(y)} = x^n \cdot Q\left(\frac{1}{x}\right).$$

   A Q polinomot a P *reciprok-polinomjának* nevezzük. Ekkor

$$P(x_k) = 0 \quad \Leftrightarrow \quad Q\left(\frac{1}{x_k}\right) = 0,$$

   vagyis Q gyökei P gyökeinek reciprokai.

Biz. folyt: Alkalmazzuk a már belátott becslésünket Q-ra:

$$\frac{1}{|x_k|} < 1 + \frac{\max\limits_{i=1}^{n} |a_i|}{|a_0|} = \frac{1}{r} \quad \Rightarrow \quad |x_k| > r.$$

$\square$

Megjegyzés: Akár komplex együtthatós polinomokat is megengedhetünk a tételben, a bizonyítás menetén nem változtat.

Polinomok és deriváltjaik helyettesítési értékeinek kiszámítására.

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 = \ldots$$

Átzárójelezzük:

$$\begin{aligned} P(x) &= \underbrace{\left(a_n x^{n-1} + a_{n-1}x^{n-2} + \cdots + a_2 x + a_1\right)}_{a_1^{(1)}} \cdot x + a_0 = \\ &= \Big(\underbrace{\left(a_n x^{n-2} + a_{n-1}x^{n-3} + \cdots + a_2\right)}_{a_2^{(1)}} \cdot x + a_1\Big) \cdot x + a_0 = \\ &= \cdots = \Big(\ldots \underbrace{\left(a_n x + a_{n-1}\right)}_{a_{n-1}^{(1)}} \cdot x + \ldots\Big) \cdot x + a_0. \end{aligned}$$

Megj.: Más elnevezés: Horner-módszer, Horner-elrendezés.

Definíció: Horner-algoritmus

A $P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ polinom adott ξ helyen vett helyettesítési értéke számolható a következő módon:

1. $a_n^{(1)} := a_n$,
2. $a_k^{(1)} := a_k + \xi \cdot a_{k+1}^{(1)} \quad (k = n-1, \ldots, 1, 0)$,

ekkor $P(\xi) = a_0^{(1)}$.

Állítás: A Horner-algoritmus műveletigénye

Egy n-edfokú polinom adott helyen felvett értéke kiszámítható n szorzás és n összeadás által, azaz $\mathcal{O}(n)$ művelettel.

Biz.: ✓ $\square$

Táblázat $P(\xi)$ kézi számolásához:

a_n	a_{n-1}	a_{n-2}	$\ldots$	a_k	$\ldots$	a_1	a_0
$\boxed{\xi}$	$\xi \cdot a_n^{(1)}$	$\xi \cdot a_{n-1}^{(1)}$	$\ldots$	$\xi \cdot a_{k+1}^{(1)}$	$\ldots$	$\xi \cdot a_2^{(1)}$	$\xi \cdot a_1^{(1)}$
$a_n^{(1)}$	$a_{n-1}^{(1)}$	$a_{n-2}^{(1)}$	$\ldots$	$a_k^{(1)}$	$\ldots$	$a_1^{(1)}$	$\boxed{a_0^{(1)}}$

Példa

Számítsuk ki a $P(x) = x^5 + 6x^4 - x^3 + 3x^2 - 15x - 7$ polinom helyettesítési értékét a $\xi = 2$ helyen.

1	6	-1	3	-15	-7
$\boxed{2}$	$2 \cdot 1$	$2 \cdot 8$	$2 \cdot 15$	$2 \cdot 33$	$2 \cdot 51$
1	8	15	33	51	$\boxed{95}$

Tehát $P(2) = 95$, amihez összesen 10 műveletet végeztünk.

Állítás: Horner-algoritmus és a derivált

A P polinom felírható a következő alakban:

$$P(x) = a_0^{(1)} + (x - \xi) \cdot \overbrace{(a_1^{(1)} + \cdots + a_n^{(1)} x^{n-1})}^{P_1(x)},$$

ahol az $a_i^{(1)}$ $(i = 0, \ldots, n)$ értékeket a Horner-algoritmus adja.
Továbbá

$$P'(\xi) = P_1(\xi) = a_1^{(2)}.$$

Megj.: $\sim$Taylor-polinom ξ körül.

$$P(x) = a_0^{(1)} + (x-\xi)\cdot\underbrace{(a_1^{(1)} + \dots + a_k^{(1)}x^{k-1} + a_{k+1}^{(1)}x^k \dots + a_n^{(1)}x^{n-1})}_{P_1(x)}$$

Biz.:

1. P-ben x^k $(k = 0, \dots, n-1)$ együtthatója
   - külön: x^n együtthatói a két oldalon: $a_n = a_n^{(1)}$, ✓
   - bal oldalon definícó szerint: a_k,
   - a fenti alak szerint a jobb oldalon: $a_k^{(1)} - \xi \cdot a_{k+1}^{(1)}$.
   - A Horner-algoritmus szerint: $a_k^{(1)} = a_k + \xi \cdot a_{k+1}^{(1)}$. ✓
2. P deriváltja a fenti alakból (összeg, szorzat):

$$P'(x) = 1 \cdot P_1(x) + (x-\xi)\cdot P_1'(x) \quad \Rightarrow \quad P'(\xi) = P_1(\xi).$$

Biz. folyt: $P_1(\xi)$ kiszámítása ugyanúgy, Horner-algoritmussal, P_1 együtthatói: $a_n^{(1)}, \dots, a_1^{(1)}$.

1. $a_n^{(2)} := a_n^{(1)}$,
2. $a_k^{(2)} := a_k^{(1)} + \xi \cdot a_{k+1}^{(2)} \quad (k = n-1, \dots, 1)$,

ekkor $P_1(\xi) = P'(\xi) = a_1^{(2)}$. $\square$

Folytatjuk a táblázatot:

a_n	a_{n-1}	a_{n-2}	$\dots$	a_1	a_0
$\boxed{\xi}$	$\xi \cdot a_n^{(1)}$	$\xi \cdot a_{n-1}^{(1)}$	$\dots$	$\xi \cdot a_2^{(1)}$	$\xi \cdot a_1^{(1)}$
$a_n^{(1)}$	$a_{n-1}^{(1)}$	$a_{n-2}^{(1)}$	$\dots$	$a_1^{(1)}$	$\boxed{a_0^{(1)} = P(\xi)}$
$\boxed{\xi}$	$\xi \cdot a_n^{(1)}$	$\xi \cdot a_{n-1}^{(1)}$	$\dots$	$\xi \cdot a_2^{(1)}$	
$a_n^{(2)}$	$a_{n-1}^{(2)}$	$a_{n-2}^{(2)}$	$\dots$	$\boxed{a_1^{(2)} = P_1(\xi)}$	

Tovább is folytathatjuk... $P(x) = a_0^{(1)} + (x-\xi) \cdot P_1(x)$

Állítás: Horner-algoritmus és a magasabbrendű deriváltak

A P polinom felírható a következő alakban:

$$P(x) = a_0^{(1)} + a_1^{(2)}(x-\xi) + a_2^{(3)}(x-\xi)^2 + \cdots + a_n^{(n+1)}(x-\xi)^n,$$

ahol az $a_i^{(j+1)}$ ($j = 0, \ldots, n$; $i = j, \ldots, n$) értékeket a Horner-módszer adja. Továbbá:

$$\frac{P^{(j)}(\xi)}{j!} = P_j(\xi) = a_j^{(j+1)},$$

ahol $P_j(x) = a_j^{(j)} + \cdots + a_n^{(j)} x^{n-j}$.

Biz.: indukcióval, nem kell.

Megjegyzés: Ha a táblázatot addig folytatjuk, míg csak 1 elemet kapunk, akkor az átlóban találjuk a P polinom ξ körüli Taylor-polinomjának együtthatóit.

Példa

Határozzuk meg a $P(x) = x^3 - x^2 + x - 1$ polinom $\xi = 1$ körüli Taylor-polinomját a Horner-módszer segítségével!

Példa

$P(x) = x^4 - 2x^3 + 3x^2 - x + 1 =$

$= 1 \cdot (x-1)^4 + 2 \cdot (x-1)^3 + 3 \cdot (x-1)^2 + 3 \cdot (x-1) + 2$

az 1 körüli Taylor-polinomot kaptuk.

1	-2	3	-1	1
$\boxed{1}$	$1 \cdot 1$	$1 \cdot (-1)$	$1 \cdot 2$	$1 \cdot 1$
1	-1	2	1	$\boxed{2 = P(1)}$
$\boxed{1}$	$1 \cdot 1$	$1 \cdot 0$	$1 \cdot 2$	
1	0	2	$\boxed{3 = P'(1)}$	
$\boxed{1}$	$1 \cdot 1$	$1 \cdot 1$		
1	1	$\boxed{3 = \frac{P''(1)}{2}}$		
$\boxed{1}$	$1 \cdot 1$			
1	$\boxed{2 = \frac{P'''(1)}{3!}}$			

Példák Matlab-ban

1. Véletlen (valós és komplex) együtthatós magasabbfokú ($n = 5, 10, 50, 100$) polinomok gyökeinek és a rájuk adott korlátoknak szemléltetése.