

# Numerikus módszerek 1 jegyzet

Toffalini Leonardo, Havasi Ágnes

2024. április 16.

ELTE



# Tartalomjegyzék

---

<b>1. Bevezetés</b>	<b>1</b>
<b>2. Numerikus modellezés</b>	<b>3</b>
2.1. Numerikus modellezés lépései	3
2.2. Hibaforrások	4
2.3. Hibafogalmak	4
2.4. Az alpműveletek hibája	5
2.5. Korrekt kitűzésű feladatok	5
<b>3. Normált terek</b>	<b>7</b>
3.1. Normált tér	7
3.2. Fontos fogalmak normált terekben	8
3.3. Mátrixnormák	9
3.4. Kondíciós szám	10
<b>4. Lineáris algebrai egyenletrendszerek megoldása</b>	<b>13</b>
4.1. Gauss-elimináció	13
4.2. Főelem kiválasztás (pivoting)	16
4.3. Klasszikus iterációs módszerek	17
4.4. Richardson-iteráció	18
4.5. Jacobi-iteráció	18
4.6. Gauss-Seidel-iteráció	18
4.7. Stacionárius-iteráció	19
4.8. Stacionárius iteráció konvergenciája	20
4.9. SOR-módszer konvergenciája	21
<b>5. Gradiens alapú módszerek</b>	<b>23</b>
5.1. Gradiens módszer	25
5.2. Konjugált gradiens-módszer	25
<b>6. Általános algebrai egyenletek megoldása</b>	<b>27</b>
6.1. Gyökök stabilitása	27
6.2. Konvergencia sebesség	28
6.3. Intervallum felezés	29
6.4. Egyszerű iteráció (fixpont-iteráció)	31
6.5. Newton módszer (érintő módszer)	32
6.6. Egyenletrendszerek megoldása	34
<b>7. Interpolációs feladatok</b>	<b>35</b>
7.1. Interpolációs alapfeladat	35
7.2. Függvény approximáció interpolációval	37
7.3. Hermite-interpoláció	38
7.4. Spline-interpoláció	39
7.5. Lineáris spline-interpoláció	39
7.6. Kvadratikus spline-interpoláció	40
7.7. Legkisebb négyzetek módszere	40
<b>8. Közelítő integrálás</b>	<b>43</b>



# 1. fejezet

## Bevezetés

Az alábbi egy jegyzet Havasi Ágnesnek a 2023/2024-es tavaszi félévében tartott Numerikus Módszerek 1 előadásáról. A jegyzet nem teljeskörű dokumentációja az előadáson elhangzottaknak és nem vállal felelősséget az esetleges hibákért.



## 2. fejezet

# Numerikus modellezés

Ebben a fejezetben tárgyalni fogjuk az alapvető lépéseit és fogalmait a numerikus modellezésnek és a numerikus módszereknek.

### 2.1. Numerikus modellezés lépései

#### 1. Valódi probléma

Halpopuláció időbeli fejlődése.

#### 2. Tudományos modell

Vannak zsákmányhalak és ragadozó halak. A zsákmányhalak és a ragadozóhalak populációját befolyásolni, többek között:

- természetes szaporulat
- ragadozók esznek zsákmány halakat
- természetes pusztulás

#### 3. Matematikai modell

- jelölje  $x(t)$  a zsákmányhalak  $t$  időbeli össztömegét
- jelölje  $y(t)$  a ragadozóhalak  $t$  időbeli össztömegét

Ezekkel a jelölésekkel felírhatjuk a változók közti összefüggést egy differenciálegyenlettel:

$$\begin{aligned}x' &= ax - bxy \\y' &= -cy + dxy \\x(0) &= x_0 \\y(0) &= y_0\end{aligned}$$

#### 4. Numerikus modell

Közelítő módszert alkalmazunk az előző, úgy nevezett *Lotka Volterra* egyenletre.

#### 5. Számítógépes modell

Lekódoljuk és futtatjuk a numerikus modellnek a programját.

## 2.2. Hibaforrások

### 1. Modellhiba

A tudományos és a matematikai modellben éltünk egyszerűsítésekkel, melyek nem pontosan ábrázolták a valóságot.

### 2. Képlethiba

A matematikai és a numerikus modellben egy egyszerűbb kifejezéssel helyettesítettünk egy bonyolultabb kifejezést. Tipikusan egy Taylor-sorral helyettesítettünk egy nehezen leírható függvényt.

Például:

$$\exp(2) = \sum_{k=0}^{\infty} \frac{2^k}{k!} \approx \sum_{k=0}^N \frac{2^k}{k!}$$

A képlethibának az egyik fajtája a *diszkretizációs hiba*, melynek tipikus esetei:

- folytonos függvényt helyettesítünk rácspont függvénnyel
- deriváltat helyettesítünk differenciáhányadossal
- integrált helyettesítünk egy véges összeggel
- végtelent helyettesítünk egy tetszőlegesen nagy termiszetes számmal

### 3. A bemenő adatok hibája

Gyakran nem pontosan kapjuk meg az adatokat és így számolnunk kell ezzel a hibaforrással. Ez gyakran mérési hibából következik.

### 4. Számábrázolási hiba

A való életben nem szimbólikusan számolunk valós számokkal, hanem egy számítógépre hagyjuk a számításokat. A számítógépünk viszont csak egy véges részhalmozát képes ábrázolni a valósszámoknak, így ha egy valós számot adunk meg egy számítógépnek, akkor az a hozzá legközelebb álló ábrázolható számot fogja helyette használni.

## 2.3. Hibafogalmak

Szeretnénk számszerűen megfogalmazni, hogy mennyire pontosan számoltunk és, hogy mennyire tér el a számított érték a valódi értéktől. A továbbiakban jelölje  $a \in \mathbb{R}$  a pontos értéket és  $\tilde{a} \in \mathbb{R}$  a számított értéket.

**Definíció 2.3.1** Az  $\tilde{a}$  abszolút hibájának a  $\Delta a := a - \tilde{a}$  számot értjük.

**Definíció 2.3.2** A  $\Delta a \in \mathbb{R}_0^+$  számot az  $\tilde{a}$  egy abszolút hibakorlátjának nevezzük, ha  $|\Delta a| \leq \Delta a$

Jelölésben  $a = \tilde{a} \pm \Delta a$

**Definíció 2.3.3**  $\tilde{a}$  relatív hibájának nevezzük a következőt:  $\delta a = \frac{\Delta a}{|\tilde{a}|}$

**Definíció 2.3.4**  $\tilde{a}$  relatív hibakorlátjának nevezzük a következőt:  $\delta a \in \mathbb{R}_0^+$  szám melyre  $|\delta a| \leq \delta a$



## 2.4. Az alpműveletek hibája

A következőkben keressük, hogy mennyire hibázunk, amikor számábrázolási hibából következően nem a pontos értékekkel végezzük el az alpműveleteket.

Tegyük fel, hogy  $x, y \in \mathbb{R}$  helyett a hibás  $\tilde{x}, \tilde{y} \in \mathbb{R}$  számokkal végezzük el az alpműveleteket.

### 1. Összeadás

$$\begin{aligned} |(x + y) - (\tilde{x} + \tilde{y})| &= |x - \tilde{x} + y - \tilde{y}| \\ &\leq |x - \tilde{x}| + |y - \tilde{y}| \\ &\leq \Delta_x + \Delta_y \end{aligned}$$

### 2. Kivonás

$$\begin{aligned} |(x - y) - (\tilde{x} - \tilde{y})| &= |x - \tilde{x} + \tilde{y} - y| \\ &\leq |x - \tilde{x}| + |\tilde{y} - y| = |x - \tilde{x}| + |y - \tilde{y}| \\ &\leq \Delta_x + \Delta_y \end{aligned}$$

### 3. Szorzás

$$\begin{aligned} |xy - \tilde{x}\tilde{y}| &= |xy + x\tilde{y} - x\tilde{y} - \tilde{x}\tilde{y}| \\ &= |x(y - \tilde{y}) + \tilde{y}(x - \tilde{x})| \\ &\approx |\tilde{x}(y - \tilde{y}) + \tilde{y}(x - \tilde{x})| \\ &\leq |\tilde{x}|\Delta_y + |\tilde{y}|\Delta_x := \Delta_{xy} \end{aligned}$$

### 4. Hányados

$$\left| \frac{x}{y} - \frac{\tilde{x}}{\tilde{y}} \right| \leq \frac{\Delta_{xy}}{\tilde{y}^2}$$

## 2.5. Korrekt kitűzésű feladatok

Mielőtt nekiállnánk egy feladatot megoldani érdemes elgondolkoznunk azon, hogy egyáltalán van-e értelme megoldani, vagy korrekten van-e kitűzve a feladat.

Ha kapunk egy feladatot, akkor a következők korrekt elvárások:

- Létezzen megoldás (*egzisztencia*)
- Csak egy megoldás létezzen (*unicitás*)
- A feladat pontos megoldása folytonosan függjön a bemenő adatoktól.

Például az  $Ax = b$  nem ilyen, mert ha egy kicsit megváltoztatjuk az  $A$  együttható mátrix elemét, akkor a megoldás nagy mértékben változhat.



## 3. fejezet

# Normált terek

Eddig csak valós számokra alkalmaztuk az abszolútérték függvényt, amikor hibafogalmakról beszéltünk. Megeshet, hogy a keresett érték nem egy valós szám, hanem például egy mátrix vagy egy függvény vagy egy tetszőleges operátor. Ilyenkor nem tudjuk alkalmazni a szokásos abszolút érték függvényt, mert nem tudjuk, hogy mit jelent egy mátrix abszolútértéke.

Ennek érdekében bevezetünk egy olyan teret, melynek elemeire lehet a kiterjesztett abszolútérték függvényt használni.

### 3.1. Normált tér

Ahhoz, hogy kiterjesszük az abszolútérték függvényt tekintsük a tulajdonságait, hogy mit kéne örökölnie egy tágabb hossz fogalomnak:

1.  $|x| \geq 0 \quad \forall x \in \mathbb{R}$  és  $|x| = 0 \iff x = 0$
2.  $|\lambda x| = |\lambda| \cdot |x|$  (abszolút homogenitás)
3.  $|x + y| \leq |x| + |y| \quad \forall x, y \in \mathbb{R}$  (háromszög egyenlőtlenség)

**Definíció 3.1.1** Legyen  $X$  tetszőleges vektortér, és  $\|\cdot\| : X \rightarrow \mathbb{R}$  egy függvény a következő tulajdonságokkal:

1.  $\|x\| \geq 0 \quad \forall x \in X$  és  $\|x\| = 0 \iff x = 0_X$  ( $X$  nullvektora)
2.  $\|\lambda x\| = |\lambda| \cdot \|x\| \quad \forall x \in X, \forall \lambda \in \mathbb{R}$
3.  $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in X$

Ekkor ezen  $\|\cdot\|$  függvényt normának nevezzük és a normált tér (N.T.) a következő rendezett pár:  $(X, \|\cdot\|)$ .

**Definíció 3.1.2** Ha  $(X, \|\cdot\|)$  Normált tér, akkor  $x, y \in X$  elemek távolságán az  $\|x - y\|$  számot értjük.

**Megjegyzés 1** Ezt a  $\|x - y\|$  távolságot szokás a norma által indukált metrikának nevezni.

**Példa 1** Példák normákra és normált terekre:

1.  $X = \mathbb{R}$  és  $\|\cdot\| = |\cdot|$

2.  $X = \mathbb{R}^n$  a következő normákkal:

(i)  $\|x\|_1 := \sum |x_j|$

(ii)  $\|x\|_2 := \sqrt{\sum |x_j|^2}$

(iii)  $\|x\|_\infty := \max\{|x_j|\}$

(iv)  $\|x\|_p := (\sum |x_j|^p)^{1/p}$

Ha  $p \rightarrow \infty$  akkor  $\|x\|_p \rightarrow \|x\|_\infty \quad \forall x \in X$

3.  $X = C[a, b]$ , azaz az  $[a, b]$  intervallumon értelmezett folytonos függvények, a következő normákkal:

(i)  $\|f\|_\infty := \max_{x \in [a, b]} |f(x)|$

(ii)  $\|f\|_f := \int_a^b |f(x)| dx$

## 3.2. Fontos fogalmak normált terekben

Most hogy már kiterjesztettük a hossz fogalmát normált terekre, így képesek vagyunk az előző fejezetekben bevezetett fogalmakat analóg módon megfogalmazni a tér normájával.

### 1. Hibafogalmak

Legyen  $(X, \|\cdot\|)$  egy tetszőleges Normált tér és  $a, \tilde{a} \in X$ . Ekkor

- $\tilde{a}$  abszolút hibája:  $a - \tilde{a} \in X$
- $\tilde{a}$  abszolút hibakorlátja:  $\Delta_a \in \mathbb{R}$  szám, melyre  $\|a - \tilde{a}\| \leq \Delta_a$
- $\tilde{a}$  relatív hibája:  $\frac{a - \tilde{a}}{\|\tilde{a}\|} \in X$
- $\tilde{a}$  relatív hibakorlátja:  $\frac{\|a - \tilde{a}\|}{\|\tilde{a}\|} \leq \delta_a \in \mathbb{R}$

### 2. Konvergencia

**Definíció 3.2.1** Azt mondjuk, hogy az  $(x_n) \subset X$  sorozat konvergens, ha  $\exists x \in X$ , melyre  $\|x_n - x\| \rightarrow 0$  ha  $n \rightarrow \infty$ .

### 3.3. Mátrixnormák

Tudjuk, hogy az  $\mathbb{R}^{n \times n}$ -beli mátrixok a rajta értelmezett  $+$  (összeadás) és  $\lambda$ -val való szorzás műveletekkel vektorteret alkotnak.

**Kérdés 1** *Hogyan definiálható ezen a vektortéren norma?*

**Definíció 3.3.1** *Legyen  $\|\cdot\|_{\mathbb{R}^n}$  egy  $\mathbb{R}^n$ -beli vektornorma. Ekkor az  $A \in \mathbb{R}^{n \times n}$  mátrix ezen vektornorma által indukált mátrixnormáján a következő számot értjük:*

$$\|A\| := \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}}$$

Magyarázó jelentések a definícióhoz:

- $\|Ax\|_{\mathbb{R}^n}$  - az  $Ax$  vektor "hossza"
- $\frac{\|Ax\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}}$  - hányszorosára nyújtotta az  $A$  mátrix az  $x$  vektort
- $\sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}}$  - lehetséges legnagyobb megnyújtásnak az értéke

**Példa 2** *Tekintsük pár mátrixnak pár mátrixnormáját.*

1.

$$\|I\| = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ix\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}} = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|x\|}{\|x\|} = \sup 1 = 1$$

Tehát bármelyik  $\mathbb{R}^n$ -beli norma által indukált mátrixnormában az identitás mátrix normája 1, azaz  $\|I\| = 1$ .

2. A sup-norma kiszámítása a tanult vektornormák esetén: Ha  $\|\cdot\|_{\mathbb{R}^n} = \|\cdot\|_1$ , akkor:

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

max oszlopösszeg!

Például:

$$\begin{bmatrix} -2 & 1 \\ 0 & 3 \end{bmatrix} \Rightarrow \|A\|_1 = \max\{|-2| + |0|, |1| + |3|\} = 3$$

3. Ha  $\|\cdot\| = \|\cdot\|_2$ , akkor:

$$\|A\| = \|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$$

ahol  $\lambda_{\max}$  a legnagyobb sajátértéket jelöli. Ezt a normát szokás *spektrálnormának* nevezni, mert a sajátértékek halmazát *spektrál*-nak nevezik.

4. Ha  $\|\cdot\| = \|\cdot\|_\infty$ , akkor:

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

max sorösszeg!

Például:

$$\begin{bmatrix} -2 & 1 \\ 0 & 3 \end{bmatrix} \Rightarrow \|A\|_\infty = \max\{|-2| + |1|, |0| + |3|\} = 3$$

**Állítás 3.3.1** Az indukált mátrix normákra igazak a következők:

1.  $\|Ax\| \leq \|A\| \cdot \|x\| \quad \forall A \in \mathbb{R}^{n \times n}, \forall x \in \mathbb{R}^n.$
2.  $\|I\| = 1$  (láttuk).
3.  $\|A \cdot B\| \leq \|A\| \cdot \|B\| \quad \forall A, B \in \mathbb{R}^{n \times n}$  (szub multiplikatívitas).

**Megjegyzés 2** Vannak egyéb, nem indukált, mátrix normák. például:

1.  $\|A\|' = \max_{i,j} |a_{ij}|$  (maximális elem)
2.  $\|A\|'' = \sum_{i,j=1}^n |a_{ij}|$  (elemösszeg)
3.  $\|A\|_F = \sqrt{\sum_{i,j=1}^n a_{ij}^2}$  (Frobenius norma)

Ezekre a nem indukált mátrix normákra nem feltétlenül teljesülnek a 3.3.1-beli tulajdonságok.

### 3.4. Kondíciósám

Az előbb meggondoltuk, hogy egy lineáris egyenletrendszernek,  $Ax = b$ -nek, az  $A$  együtthatómátrixának egy elemét kicsit perturbálva a megoldás drasztikusan változhat. Célunk, hogy megfogalmazzuk, hogy mennyire változhat a megoldás kis perturbációra.

A továbbiakban a következő egyenletrendszerrel fogunk foglalkozni.

$$Ax = b \tag{3.1}$$

Ahol  $A \in \mathbb{R}^{n \times n}$ ,  $\det A \neq 0$ ,  $b \in \mathbb{R}^n$

Tegyük fel, hogy  $b$  helyett a perturbált  $\tilde{b}$  van a jobb oldalon:

$$A\tilde{x} = \tilde{b}$$

Jelölje:

$$\begin{aligned}\Delta x &= x - \tilde{x} \implies \tilde{x} = x - \Delta x \\ \Delta b &= b - \tilde{b} \implies \tilde{b} = b - \Delta b\end{aligned}$$

Ekkor:

$$\begin{aligned}A\tilde{x} &= \tilde{b} \\ A(x - \Delta x) &= b - \Delta b \\ Ax - A\Delta x &= b - \Delta b \\ A\Delta x &= \Delta b \\ \Delta x &= A^{-1}\Delta b\end{aligned}$$

Nézzük  $\|\Delta x\|$ -át valamelyik  $\mathbb{R}^n$ -beli normában:

$$\|\Delta x\| = \|A^{-1}\Delta b\| \leq \|A^{-1}\| \cdot \|\Delta b\|$$

Most alkalmazzuk a 3.1-es egyenletrendszerre a normát.

$$b = Ax$$

$$\begin{aligned}\|b\| &= \|Ax\| \leq \|A\| \cdot \|x\| \\ \frac{1}{\|x\|} &\leq \|A\| \cdot \frac{1}{\|b\|} \\ \implies \frac{\|\Delta x\|}{\|x\|} &\leq \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\Delta b\|}{\|b\|}\end{aligned}$$

Tehát azt kaptuk, hogy minél nagyobb  $\|A^{-1}\| \cdot \|A\|$  annál pontatlanabb a becslés.

**Definíció 3.4.1** Az  $\|A^{-1}\| \cdot \|A\|$  számot az  $A$  mátrix kondíció számának nevezzük és  $\text{cond}(A)$ -val jelöljük.

**Definíció 3.4.2** Azt mondjuk, hogy a 3.1-es egyenletrendszer rosszul kondicionált, ha  $\text{cond}(A) \gg 1$ .

**Példa 3** Nezzük meg a már említett példának a kondíció számát.

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1.01 \end{bmatrix}$$

Alkalmazzuk az  $\|\cdot\|_1$  által indukált mátrix normát.

$$\|A\|_1 = \max\{1 + 1, 1 + 1.01\} = 2.01$$

$$A^{-1} = \begin{bmatrix} 101 & -100 \\ -100 & 100 \end{bmatrix} \implies \|A^{-1}\|_1 = \max\{101 + 100, 100 + 100\} = 201$$

$$\text{cond}(A) = 201 \cdot 2.01 = 404.01 \gg 1$$

Tehát valóban rosszul kondicionált volt az egyenlet rendszer.





## 4. fejezet

# Lineáris algebrai egyenletrendszerek megoldása

Lineáris algebrai egyenletrendszerek megoldásaira két féle megoldási módszert fogunk tanulni. Direkt megoldókat és iterációs módszereket. Az előzőhöz tartozik például a Cramer-szabály vagy a Gauss-elimináció. Az iterációs módszereknek viszont a lényege az, hogy egy vektorsorozatot generálnak, melyek tartanak a pontos megoldáshoz.

### 4.1. Gauss-elimináció

Megoldandó egyenletrendszer:  $Ax = b$ ,  $A \in \mathbb{R}^{m \times m}$ ,  $\det A \neq 0$ ,  $b \in \mathbb{R}^m$

Lineáris algebrából tudjuk, hogy ezek a feltételek mellett egyértelműen létezik megoldás, tehát korrekt kitűzésű a feladat és van értelme nekiállni megoldani.

A lineáris egyenletrendszer teljes anyakönyvezet néven a következő:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m &= b_1 & (1) \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mm}x_m &= b_m & (m) \end{aligned}$$

I. alakban: Átalakítjuk az egyenletrendszert normált felső háromszög mátrixúvá. Tehát a főátlóban legyenek egyesek és a főátló alatt csupa nulla.

1. lépés: Tegyük fel, hogy  $a_{11} \neq 0$  ekkor

$$x_1 + \frac{a_{12}}{a_{11}}x_2 + \frac{a_{13}}{a_{11}}x_3 + \frac{a_{1m}}{a_{11}}x_m = \frac{b_1}{a_{11}} = y_1 \quad (1) \quad (4.1)$$

2. lépés: 4.1 segítségével a másodiktól az  $m$ -edik egyenletekből elimináljuk  $x_1$ -et, kivonva belőlük a 4.1-nek az  $a_{i1}$ -szeresét.

$$\begin{aligned} x_1 + \frac{a_{12}}{a_{11}}x_2 + \frac{a_{13}}{a_{11}}x_3 + \frac{a_{1m}}{a_{11}}x_m &= \frac{b_1}{a_{11}} = y_1 \\ a_{22}^{(1)}x_2 + \dots &= y_2 \\ &\vdots \\ a_{m2}^{(1)}x_2 + \dots + a_{mm}^{(1)}x_m &= b_m \end{aligned}$$

3. lépés: Nem írom tovább mert mindenki tud Gauss-eliminálni...

**Kérdés 2** Mikor hajtható végre a Gauss-elimináció?

I. szakaszban  $Ax = b \implies Ux = y$

**Kérdés 3** Mi a kapcsolat  $y$  és  $b$  között?

$$\begin{aligned} b_1 &= a_{11}y_1 \\ b_2 &= a_{21}y_1 + a_{22}^{(1)}y_2 \\ &\vdots \\ b_m &= l_{j1}y_1 + l_{j2}y_2 + \dots + l_{mm}y_m \end{aligned}$$

Ahol  $l_{jj} = a_{jj}^{(j-1)}$

Kompaktabb mátrix formába átírva:

$$\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22}^{(1)} & \dots & 0 \\ & & \dots & a_{mm}^{(m-1)} \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

Ha a Gauss-elimináció elvégezhető akkor a fenti mátrix invertálható, azaz a főátlóban nincs 0, tehát  $\exists L^{-1}$ , ahol  $L$  a fenti alsó háromszög mátrix.

$$\text{Tehát } Ly = b \implies y = L^{-1}b \implies Ux = L^{-1}b \implies L Ux = b$$

Ebből adódik egy új módszer ( $LU$  felbontás):

1. Felírjuk az  $A$ -t  $A = LU$  alakban, ahol  $L$  invertálható alsó háromszög mátrix és  $U$  olyan felső háromszög mátrix melynek a főátlójában csak egyesek vannak.
2. Megoldjuk az  $Ly = b$  egyenletrendszert, ebből kapunk egy értéket  $y$ -ra.
3. Megoldjuk az  $Ux = y$  egyenletrendszert, amiből megkapjuk  $x$ -et.

Belátható, hogy az LU felbontás első és második lépse ekvivalens a Gauss-elimináció első szakaszával és harmadik lépés ekvivalens a Gauss-elimináció második szakaszával. Tehát ez a módszer a Gauss-elimináció módosított algoritmusa.

Ahhoz, hogy megválaszoljuk, hogy mikor végezhető el a Gauss-elimináció elég megválaszolnunk, hogy mikor létezik LU felbontás.

A következőképpen jelöljük a balfelső sarokdeterminánsokat (*főminorokat*):

$$\Delta_1 := a_{11}, \quad \Delta_2 := \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \dots, \Delta_m := \det A$$

**Állítás 4.1.1** *Ha  $\Delta_j \neq 0, \forall j \in \{1, \dots, m\}$ , akkor létezik LU felbontása  $A$ -nak, és az egyértelmű.*

*Bizonyítás:* Csak az  $A \in \mathbb{R}^{2 \times 2}$  esetre mutatjuk meg, magasabb dimenzióra teljes indukcióval lehet belátni az állítást.

Először bizonyítsuk a létezést.

$$A = L \cdot U = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \cdot \begin{bmatrix} 1 & u_{12} \\ 0 & 1 \end{bmatrix}$$

felbontás létezik  $\iff$  létezik  $l_{11}, l_{21}, l_{22}, u_{12}$  ismeretlenekre nézve megoldása a következő egyenlet rendszernek.

$$\begin{aligned} l_{11} &= a_{11} \\ l_{11}u_{12} &= a_{12} \\ l_{21} &= a_{21} \\ l_{21}u_{12} + l_{22} &= a_{22} \end{aligned}$$

és a következő  $L$  mátrixnak létezen inverze

$$L = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix}$$

azaz  $l_{11} \neq 0, l_{22} \neq 0$ .

Ha  $a_{11} \neq 0$ , akkor látható, hogy ennek az egyenletrendszernek egyértelműen létezik megoldása és az a következő:

$$l_{11} = a_{11}, \quad u_{12} = \frac{a_{12}}{a_{11}}, \quad l_{21} = a_{21}, \quad l_{22} = a_{22} - a_{21} \frac{a_{12}}{a_{11}}$$

Továbbá,  $l_{11} \neq 0$ , mert  $a_{11} \neq 0$  és  $l_{22} \neq 0$ , mert  $l_{22} = \frac{\det A}{a_{11}} \implies \exists L^{-1}$

Most lássuk be, hogy egyértelműen létezik.

Tegyük fel, hogy  $A = L_1 U_1 = L_2 U_2$

$$\begin{aligned} L_2^{-1} L_1 U_1 &= U_2 \\ L_2^{-1} L_1 &= U_2 U_1^{-1} \end{aligned}$$

Mivel az alsóháromszög mátrixok és a felső háromszög mátrixok is egy-egy csoportot alkotnak, ezért a fenti csak akkor igaz, ha  $L_2^{-1}L_1$  és  $U_2U_1^{-1}$  is diagonális. Továbbá  $U_{1,2}$ -nek a főátlójában egyesek vannak, tehát  $U_2U_1^{-1}$ -nek is a főátlójában egyesek vannak. Tehát mindkét oldalon az egység mátrix van.

$$\implies L_2^{-1}L_1 = I = U_2U_1^{-1} \implies L_1 = L_2, \quad U_1 = U_2$$

Megmutatható, hogy ha  $\Delta_j \neq 0$  valamely  $j$ -re, akkor  $\exists$  LU-felbontása  $A$ -nak.  $2 \times 2$  esetben jól látszik:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 0 & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$\begin{bmatrix} 0 & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \cdot \begin{bmatrix} 1 & u_{12} \\ 0 & 1 \end{bmatrix} \implies l_{11} = 0 \implies \text{ekkor } L \text{ nem invertálható}$$

**Következmény 1** A Gauss-elimináció pontosan akkor hajtható végre, ha  $A$  összes bal felső sarokdeterminánsa nem 0.

**Megjegyzés 3** Pár észrevétel a Gauss-elimináció és az LU felbontással kapcsolatban:

1.  $A \Delta_j \neq 0, \quad \forall j = 1, \dots, m$  teljesül, ha  $A$  szimmetrikus pozitív definit mátrix (szpd).
2.  $A \Delta_j \neq 0, \quad \forall j = 1, \dots, m$  teljesül, ha  $A$  szigorúan domináns főátlójú, tehát  $\forall i = 1, \dots, n$ -re  $2|a_{ii}| > \sum_{j=1}^m |a_{ij}|$ .
3. Ha  $\det A \neq 0$ , akkor mindig  $\exists P \in \mathbb{R}^{n \times m}$  permutáló mátrix, hogy  $PA$ -nak  $\exists$  LU felbontása.
4. Ha  $A$  szimmetrikus pozitív definit mátrix, akkor létezik egy másik felbontása is:  $A = G \cdot G^T$ , ahol  $G$  alsó háromszög mátrix, pozitív főátlóval. (Cholesky-felbontás)

## 4.2. Főelem kiválasztás (pivoting)

A Gauss-elimináció során a  $j$ -edik lépésben a  $j$ -edik sort elosztjuk  $a_{jj}$ -vel. Tehát minél kisebb  $a_{jj}$ , annál pontatlanabb az osztás. Ennek orvosolására valahogyan meg kéne oldanunk, hogy egy nagyobb elemmel osszunk, de a Gauss-elimináció lényegét tartjuk meg.

**Részleges főelem kiválasztás:** Sorcserével a főátlóba hozzuk az  $a_{jj}$  alatti legnagyobb abszolútértékű elemet.

**Teljes főelem kiválasztás:** Sorcserével és oszlopcserével az  $A[j : n, j : n]$  jobb alsó részmátrix legnagyobb abszolútértékű elemet visszük a főátlóba. Itt figyelni kell arra, hogy oszlop cserénél az  $x$  elemeket is cseréljük. Tehát ha egy  $P$  mátrixszal permutáljuk az oszlopait  $A$ -nak, akkor mikor visszaolvassuk  $x$  megoldást, akkor  $P^{-1}$ -el meg kell szorozni előtte.

### 4.3. Klasszikus iterációs módszerek

**Definíció 4.3.1** Azt mondjuk, hogy az  $x^* \in \mathbb{R}^n$  az  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  függvény fixpontja, ha  $f(x^*) = x^*$

**Definíció 4.3.2** az  $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$  függvény kontrakció az  $\|\cdot\|$   $\mathbb{R}^n$ -beli normában, ha  $\exists q \in [0, 1]$  melyre:

$$\|f(x) - f(y)\| \leq q \cdot \|x - y\| \quad \forall x, y \in D(f)$$

**Tétel 4.3.1** (Banach fixpont tétel)

Ha  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  az egész  $\mathbb{R}^n$ -en értelmezett kontrakció ( $q$ -val), akkor:

1.  $f$ -nek egyértelműen létezik  $x^*$  fixpontja.
2. Tetszőleges  $x^0 \in \mathbb{R}^n$  vektorból indítva  $x^{n+1} = f(x^n)$  rekurzióval felépített  $(x)_n$  sorozat konvergens, és  $x_n \rightarrow x^*$ .
3.  $\|x^n - x^*\| \leq \frac{q^n}{1-q} \|x^1 - x^0\|$

**Kérdés 4** Hogyan alkalmazhatjuk ezt a tételt lineáris algebrai egyenletrendszerek megoldására?

$$Ax = b, \quad A \in \mathbb{R}^{m \times m}, \quad \det A \neq 0, \quad b \in \mathbb{R}^m \quad (4.2)$$

Tegyük fel, hogy 4.2 átírható a következő alakra:

$$x = Qx + r, \quad Q \in \mathbb{R}^{m \times m}, \quad r \in \mathbb{R}^m \quad (4.3)$$

Ekkor az  $f(x) := Qx + r$  ejöléssel a feladat megoldása az  $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$  függvény fixpontja. Ezt a fixpontot keressük iterációval.

**Kérdés 5** Mikor lesz  $f$  kontrakció?

$$x, y \in \mathbb{R}^m, \quad f(x) - f(y) = Qx + r - Qy - r = Q(x - y)$$

$$\|f(x) - f(y)\|_{\mathbb{R}^m} = \|Q(x - y)\|_{\mathbb{R}^m} \leq \|Q\| \cdot \|x - y\|_{\mathbb{R}^m}$$

Tehát be kell látni, hogy  $\|Q\| < 1$ , akkor  $f$  kontrakció és  $q = \|Q\|$ .

Banach fixpont tételből következik, hogy a  $x^{n+1} = Qx^n + r$  rekurzióval definiált vektorsorozat konvergens (bármely  $\mathbb{R}^m$ -beli vektornormában), és  $x_n \rightarrow x^*$ , ahol  $x^*$  4.2 megoldása.

**Kérdés 6** Hogyan írhatjuk át 4.2-et olyan alakra amilyen 4.3?

**Kérdés 7** Mikor fog teljesülni, hogy  $\|Q\| < 1$  valamelyik indukált mátrixnorma szerint?

#### 4.4. Richardson-iteráció

A *Richardson iteráció* vagy másnéven egyszerű iteráció, ahogyan a név is sugallja a legegyszerűbb módon alakítja át az  $Ax = b$  egyenletet  $f(x) = x$  alakúra. Pusztán annyi átalakítás történik, hogy nullára rendezzük az egyenletet és mindkét oldalhoz hozzáadunk  $x$ -et.

$$\begin{aligned} Ax &= b \\ 0 &= b - Ax \\ x &= x - Ax + b \\ x &= (I - A)x + b \end{aligned}$$

Tehát  $f(x) = (I - A)x$  függvénynek fixpontjaként kapjuk az  $Ax = b$  egyenlet megoldását a Banach-fixpont tétel alapján.

#### 4.5. Jacobi-iteráció

A célunk még mindig, hogy egy függvénynek a fixpontjaként írjuk fel a lineáris egyenletrendszer megoldását. Ezt megtehetjük, ha a következőképpen felbontjuk az együttható mátrixot és egy kis algebrai manipulációt végzünk.

$$\begin{aligned} Ax &= b \\ A &= L + D + U \\ (L + D + U)x &= b \\ Dx &= -(L + U)x + b \\ x &= D^{-1}(b - (L + U)x) \\ &= -D^{-1}(L + U)x + D^{-1}b \\ Q_J &= -D^{-1}(L + U), \quad r_J = D^{-1}b \end{aligned}$$

Ekkor kapjuk, hogy a Jacobi fixpont iterációra rögzítsük  $x^0 \in \mathbb{R}^m$  kezdőpontot és legyen az általános lépés:

$$x^{n+1} = -D^{-1}(L + U)x^n + D^{-1}b$$

##### Állítás 4.5.1

$$\|Q_J\|_\infty < 1 \iff A \text{ szigorúan domináns főátlójú}$$

**Következmény 2** Ha  $A$  szigorúan domináns főátlójú, akkor a Jacobi-iteráció konvergens.

#### 4.6. Gauss-Seidel-iteráció

A célunk még mindig, hogy egy függvénynek a fixpontjaként írjuk fel a lineáris egyenletrendszer megoldását. Ezt megtehetjük, ha a következőképpen felbontjuk az együttható mátrixot és egy kis algebrai manipulációt végzünk.

$$\begin{aligned}
Ax &= b \\
A &= L + D + U \\
(L + D + U)x &= b \\
(L + D)x &= -Ux + b \\
x &= -(L + D)^{-1}Ux + (L + D)^{-1}b \\
Q_{GS} &= -(L + D)^{-1}U, \quad r_{GS} = (L + D)^{-1}b
\end{aligned}$$

## 4.7. Stacionárius-iteráció

*Észrevétel:* A Jacobi-iteráció átírható a következő módon:

$$\begin{aligned}
x^{n+1} &= -D^{-1}(L + U)x^n + D^{-1}b \\
Dx^{n+1} &= -(L + U)x^n + b \\
Dx^{n+1} &= -(A - D)x^n + b \\
D(x^{n+1} - x^n) + Ax^n &= b
\end{aligned}$$

A fentit a Jacobi-iteráció kanonikus alakjának szokás nevezni.

Hasonló módon át tudjuk írni a Gauss-Seidel iterációt is:

$$(D + L)(x^{n+1} - x^n) + Ax^n = b \quad (\text{SI})$$

A fentit a Gauss-Seidel-iteráció kanonikus alakjának szokás nevezni.

**Definíció 4.7.1** Legyen  $B \in \mathbb{R}^{m \times m}$ , és  $\tau > 0$  szám. Ekkor a következő iterációt stacionárius-iterációnak nevezzük.

$$B \cdot \frac{x^{n+1} - x^n}{\tau} + Ax^n = b$$

**Megjegyzés 4** Az előbb említett iterációs módszerek összegezve:

- *Jacobi:*  $B = D, \quad \tau = 1$
- *Gauss-Seidel:*  $B = D + L, \quad \tau = 1$
- *Még általánosabb:*  $B \leftrightarrow B_n, \quad \tau \leftrightarrow \tau_n$

Említés szintjén még egy stacionárius iteráció a *Túlrelaxációs módszer* vagy angolul *Successive overrelaxation method (SOR)*:

$$B = D + \omega L, \quad \tau = \omega, \quad \text{ahol } \omega > 0 \text{ adott paraméter}$$

$$(D + \omega L) \cdot \frac{x^{n+1} - x^n}{\omega} + Ax^n = b$$

**Megjegyzés 5** A SOR módszert  $\omega = 1$ -el írva visszakapjuk a Gauss-Seidel-iterációt.

## 4.8. Stacionárius iteráció konvergenciája

Emlék:

$$B \frac{x^{n+1} - x^n}{\tau} + A \cdot x^n = b \quad (\text{SI})$$

$$Ax = b$$

Tegyük fel, hogy  $A$  szimmetrikus pozitív definit (szpd). Tehát  $A = A^T$ ,  $x^T A x > 0$ , ha  $x \neq 0$ . Másképpen,  $\exists \delta > 0 : (Ax, x) \geq \delta \cdot \|x\|^2$ .

Jelölje  $x^*$  a 3.1 egyenlet megoldását, azaz  $Ax^* = b$  és  $e_n := x^n - x^*$  (az  $n$ -edik iteráció hibáját).

**Definíció 4.8.1** Azt mondjuk hogy a stacionárius iteráció (SI) konvergens, ha  $\exists \lim x_n$  és  $x_n \rightarrow x^*$ , azaz  $\lim_{n \rightarrow \infty} e_n = 0$ .

**Állítás 4.8.1** Tegyük föl, hogy  $A$  szpd. Ha  $\exists B^{-1}$ , és  $\tau > 0$  paraméter olyan, hogy  $B - 0.5\tau A$  szpd, akkor a stacionárius iteráció konvergens.

Bizonyítás:

$$x^n = e_n + x^*, \quad x^{n+1} = e_{n+1} + x^* \rightsquigarrow (\text{SI})\text{-be beírva}$$

$$B \frac{e_{n+1} + x^* - e_n - x^*}{\tau} + A e_n + A x^* = b$$

$$B \frac{e_{n+1} - e_n}{\tau} + A e_n = 0 \quad (3) \text{ hibaegyenlet}$$

Fejezzük ki  $e_{n+1}$ -el

$$B e_{n+1} = (B - \tau A) e_n$$

$$e_{n+1} = (I - \tau B^{-1} A) e_n$$

$$A e_{n+1} = (A - \tau A B^{-1} A) e_n$$

$$\implies (A e_{n+1}, e_{n+1}) = (A e_n - \tau A B^{-1} A e_n, e_n - \tau B^{-1} A e_n)$$

$$= (A e_n, e_n) - \tau (A B^{-1} A e_n, e_n) - \tau (A e_n, B^{-1} A e_n) + \tau^2 (A B^{-1} A e_n, B^{-1} A e_n)$$

Tudjuk, hogy

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

Tehát

$$(A e_{n+1}, e_{n+1}) = (A e_n, e_n) - 2\tau (A B^{-1} A e_n, e_n) + \tau^2 (A B^{-1} A e_n, B^{-1} A e_n)$$

Jelölje  $J_n = (A e_n, e_n)$ . Ezzel

$$J_{n+1} = J_n - 2\tau (A e_n, B^{-1} A e_n) + \tau^2 (A B^{-1} A e_n, \overbrace{B^{-1} A e_n}^{y_n})$$

Ezzel  $B y_n = A e_n$

$$= J_n - 2\tau (B y_n, y_n) + \tau^2 (A y_n, y_n) = J_n - 2\tau \left( (B y_n, y_n) - \frac{\tau}{2} (A y_n, y_n) \right)$$



$$\begin{aligned} \leadsto J_{n+1} &= J_n - 2\tau((B - 0.5\tau A)y_n, y_n) \\ \implies J_{n+1} &\leq J_n \end{aligned} \quad (4.4)$$

Mert feltétel szerint  $\tau > 0$  és  $(B - 0.5\tau A)$  szpd, tehát pozitív szor pozitív tagot vonunk ki, tehát egy pozitív számot vonunk ki. Ezért a jobb oldal kisebb mint  $J_n$ . Így a  $(J_n)$  sorozat monoton csökkenő, és  $J_n \geq 0$ , (mert  $J_n = (Ae_n, e_n)$ ), tehát ez a sorozat alulról korlátos. Tehát  $(J_n)$  konvergens, jelölés  $J^* := \lim_{n \rightarrow \infty} J_n$

4.4-ban vegyünk limeszt  $\leadsto$

$$\begin{aligned} J^* &= J^* - 2\tau \lim_{n \rightarrow \infty} ((B - 0.5\tau A)y_n, y_n) \\ \implies \lim_{n \rightarrow \infty} ((B - 0.5\tau A)y_n, y_n) &= 0 \end{aligned}$$

Mivel  $B - 0.5\tau A$  szpd, ezért  $\exists \delta > 0 : ((B - 0.5\tau A)y_n, y_n) \geq \delta \cdot \|y_n\|^2$ . Rendőrelv:

$$\begin{aligned} ((B - 0.5\tau A)y_n, y_n) &\geq \delta \cdot \|y_n\|^2 \geq 0 \\ \lim_{n \rightarrow \infty} ((B - 0.5\tau A)y_n, y_n) &\geq \lim_{n \rightarrow \infty} \delta \cdot \|y_n\|^2 \geq 0 \\ \implies 0 &\geq \lim_{n \rightarrow \infty} \delta \cdot \|y_n\|^2 \geq 0 \implies \lim_{n \rightarrow \infty} \delta \cdot \|y_n\|^2 = 0 \end{aligned}$$

Mivel  $y_n = B^{-1}Ae_n \leadsto e_n = A^{-1}By_n$ , ezért

$$0 \leq \|e_n\| = \|A^{-1}By_n\| \leq \|A^{-1}B\| \cdot \|y_n\| \rightarrow 0 \implies \|e_n\| \rightarrow 0 \implies e_n \rightarrow 0$$

Ezzel beláttuk, hogy konvergens, mert a hiba 0-hoz tart.

## 4.9. SOR-módszer konvergenciája

**Kérdés 8** *Hogyan válasszuk meg  $\omega$  paramétert, hogy konvergáljon?*

Észrevétel:  $\omega$  választása erősen függ  $A$ -tól.

**Állítás 4.9.1** *Tetszőleges  $A \in \mathbb{R}^{n \times m}$  esetén a SOR-módszer konvergenciájához szükséges, hogy  $\omega \in (0, 2)$ .*

**Állítás 4.9.2** *Ha  $A$  szpd, akkor  $\omega \in (0, 2)$  elégséges is a konvergenciához.*

**Következmény 3** *Ha  $A$  szimmetrikus pozitív definit (szpd), akkor a Gauss-Seidel iteráció konvergens, mert a Gauss-Seidel iteráció pont a SOR-módszer  $\omega = 1$ -el.*



## 5. fejezet

# Gradiens alapú módszerek

Tekintsük megint a következő egyenletet.

$$Ax = b \quad (1)$$

Tegyük fel, hogy  $A$  szimmetrikus pozitív definit (szpd).

**Definíció 5.0.1** *Definiáljuk a következő  $\phi : \mathbb{R}^m \rightarrow \mathbb{R}$  függvényt:*

$$\phi(x) := \frac{1}{2}(x, Ax) - (x, b)$$

ez differenciálható  $\mathbb{R}^m$ -en.

Célunk, hogy a  $\phi(x)$  függvényt minimalizáljuk, tehát nézzük meg, hogy hol lesz 0 a gradiense.

$$\phi'(x) = \nabla \phi(x) = Ax - b \quad (\text{számolással ellenőrizhető})$$

Ekkor pont az  $r := b - Ax$  maradékvektor  $-1$  szeresét kapjuk.

Hol 0 a gradiens?

$$\phi'(x) = Ax - b = 0 \leadsto x = A^{-1}b$$

ez éppen a 3.1 megoldása, tehát a  $\phi(x)$  függvényt minimalizálni ekvivalens azzal, hogy megoldjuk a 3.1 egyenletet.

$$\phi''(x) = A$$

Mivel feltettük, hogy  $A$  szpd, ezért  $\phi''(x)$  pozitív definit, tehát ahol a gradiens nulla ott lokális minimum hely van.

$\implies x^*$  az egyetlen lokális minimum hely / globális minimum helye  $\phi$ -nek.

**Kérdés 9** *Hogy néz ki a  $\phi$  függvény?*

**Példa 4** *Tekintsünk egy két dimenziós példát, ahol már a következő egyenletrendszerrel tartunk:*

$$2x_1 = 4$$

$$8x_2 = 8$$

*Megoldás:*

Ránézésre látszik, hogy a megoldás  $x_1^* = 2$ ,  $x_2^* = 1$

Írjuk ki  $A$  és  $b$  teljes alakját.

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix}, \quad b = \begin{bmatrix} 4 \\ 8 \end{bmatrix}$$

Helyettesítsük be  $A$ -t és  $b$  a  $\phi(x)$  függvénybe.

$$\begin{aligned} \phi(x) &= \frac{1}{2}(x, Ax) - (x, b) \\ \phi(x) &= \frac{1}{2} \left( \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} 2x_1 \\ 8x_2 \end{bmatrix} \right) - \left( \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} 4 \\ 8 \end{bmatrix} \right) \\ &= \frac{1}{2}x_1 2x_1 + \frac{1}{2}x_2 8x_2 - 4x_1 - 8x_2 = (x_1 - 2)^2 + 4(x_2 - 1)^2 - 8 \end{aligned}$$

Vizsgáljuk a szintvonalait ennek a függvénynek.

$Ac = 0$ -hoz tartozó szintvonal:

$$(x_1 - 2)^2 + 4(x_2 - 1)^2 - 8 = 0$$

$$\frac{(x_1 - 2)^2}{8} + \frac{(x_2 - 1)^2}{2} = 1$$

Tehát azt kaptuk, hogy ez egy  $(2, 1)$  középpontú ellipszis  $\sqrt{8}, \sqrt{2}$  hosszú főtengelyekkel. Azaz valóban  $(2, 1)$  a megoldás.

Tehát a függvény szintvonalai koncentrikus hiperellipszoidok!

Először gondoljuk meg, hogy egy  $x \in \mathbb{R}^m$  pontot és egy  $p \neq 0$  vektort rögzítve  $p$  irány mentén hol veszi fel a  $\phi$  a legkisebb értéket?

Jelölés:  $g(\alpha) := \phi(x + \alpha p)$

**Kérdés 10** Mely  $\alpha$ -ra lesz  $g(\alpha)$  függvény értéke minimális?

**Állítás 5.0.1** A  $g(\alpha) = \phi(x + \alpha p)$  függvény egyértelmű minimumát az

$$\alpha = \frac{(p, r)}{(p, Ap)}$$

megvalósztás esetén veszi föl!

*Bizonyítás:* Faragó I. Numerikus módszerek jegyzet 83. oldalán található.[1]

**Kérdés 11** Hogyan válasszuk meg  $p_1, p_2, \dots$  keresési irányokat?

## 5.1. Gradiens módszer

Tudjuk: A  $\nabla\phi$ -vel ellentétes irányban a legmeredekebb a lejtés.

$x_i$  pontban  $p_{i+1}$ -el jelölve a keresési irányt:

$$\begin{aligned} p_{i+1} &:= -\nabla\phi(x_i) \\ \nabla\phi(x) &= Ax - b = -r \\ \implies p_{i+1} &:= -\nabla\phi(x_i) = b - Ax_i = r_i \end{aligned}$$

ami éppen az  $x_i$  pontbeli maradékvektor.

$$x_i \rightsquigarrow x_{i+1} = x_i + \alpha \cdot p_{i+1} = x_i + \frac{(p_{i+1}, r_i)}{(p_{i+1}, Ap_{i+1})} \cdot p_{i+1} = x_i + \frac{(r_i, r_i)}{(r_i, Ar_i)} \cdot r_i$$

**Kérdés 12** *Mi lesz  $x_{i+1}$ -ben a maradékvektor?*

$$r_{i+1} = b - Ax_{i+1} = b - A \cdot \left( x_i + \frac{(r_i, r_i)}{(r_i, Ar_i)} \cdot r_i \right)$$

Vegyük észre:  $r_i \perp r_{i+1}$ , mert addig megyünk  $p_i$  irányban ameddig nem érintjük a következő szintvonalat, amire a következő gradiens merőleges.

Ez előző vizuálisan magyarázza az egymást követő irányok merőlegességét, de bizonyítsuk be formálisabban. Írjuk fel a skaláris szorzatát az egymást követő irányoknak!

$$\begin{aligned} (r_i, r_{i+1}) &\stackrel{?}{=} 0 \\ (r_i, r_{i+1}) &= \left( r_i, b - A \left( x_i + \frac{(r_i, r_i)}{(r_i, Ar_i)} \cdot r_i \right) \right) \\ &= (r_i, r_i) - (r_i, A \frac{(r_i, r_i)}{(r_i, Ar_i)} \cdot r_i) \\ &= (r_i, r_i) - \frac{(r_i, r_i)}{(r_i, Ar_i)} (r_i, Ar_i) \\ &= (r_i, r_i) - (r_i, r_i) = 0 \end{aligned}$$

**Megjegyzés 6** *Ha  $\text{cond}_2(A)$  nagy, akkor lassú a konvergencia.*

## 5.2. Konjugált gradiens-módszer

Az előbb láttuk be, hogy a gradiens módszernél a  $p_1$  keresési irány  $(r_0) \perp r_1$ . Azaz:

$$0 = (p_1, r_1) = (p_1, b - Ax_1) = (p_1, Ax^* - Ax_1) = (p_1, A(x^* - x_1))$$

**Definíció 5.2.1** Legyen  $A \in \mathbb{R}^{m \times m}$  szimmetrikus pozitív definit (szpd). Azt mondjuk, hogy  $x$  és  $y \in \mathbb{R}^m$  vektorok  $A$ -konjugáltak/ortogonálisak, ha  $(x, Ay) = 0$ .

Tehát olyan keresési irányt lenne érdemes választani, amely  $p_1$ -re  $A$ -ortogonális!  
Keressük  $p_2$ -t a következő alakban:

$$\begin{aligned} p_2 &= r_1 - \beta_1 \cdot p_1 \\ (p_1, A(r_1 - \beta_1 \cdot p_1)) &= 0 \\ \beta_1 &= ? \end{aligned}$$

$$\begin{aligned} (p_1, Ar_1) - \beta_1(p_1, Ap_1) &= 0 \\ \implies \beta_1 &= \frac{(p_1, Ar_1)}{(p_1, Ap_1)} \end{aligned}$$

Ezen  $\beta_1$ -et választva, a  $p_2 = r_1 - \beta_1 \cdot p_1$  irányba lépve az  $x^*$  minimum helybe lépünk! Tehát  $m = 2$  esetén 2 lépésben meg tudjuk határozni a lineáris egyenletrendszer megoldását.

**Megjegyzés 7**  $A \in \mathbb{R}^{m \times m}$  esetén is általánosítható az eljárás. Ekkor legfeljebb  $m$  lépésben megkapjuk a megoldást.

## 6. fejezet

# Általános algebrai egyenletek megoldása

Ebben a fejezetben egyismeretlenes valós egyenletekkel foglalkozunk. Egy ilyen egyenlet mindig felírható a következő alakban:

$$f(x) = 0 \tag{6.1}$$

ahol  $f : \mathbb{R} \rightarrow \mathbb{R}$  függvény.

Ezzel 6.1-nek a megoldása ugyanaz mint  $f$  zérushelye. Ezt keressük a továbbiakban!

### 6.1. Gyökök stabilitása

**Kérdés 13** *Mennyire érzékeny a megoldás  $f$  kis megváltoztatására?*

Tegyük fel, hogy 6.1 helyett az

$$\tilde{f}(x) = 0 \tag{6.2}$$

Egyenletet oldjuk meg, és tegyük fel, hogy 6.1-nek és 6.2-nek is  $\exists!$  megoldása, melyek  $x^*$  illetve  $\tilde{x}^*$  rendre.

A következő legyen a mérőszámunk az eltérésre:

$$|x^* - \tilde{x}^*| \leq ?$$

Ha  $f$  és  $\tilde{f}$  csak *kicsit* tér el egymástól, akkor legfeljebb mennyire tér el  $x^*$  és  $\tilde{x}^*$ ? Mérje  $\max_{[a,b]} |f - \tilde{f}|$  az  $f$  és  $\tilde{f}$  eltérését.

Tegyük fel, hogy  $f \in C[a, b] \cap D(a, b)$

*Ismétlés:* (Lagrange-közéérték tétel) Tegyük fel, hogy  $f \in C[a, b] \cap D(a, b)$ . Ekkor  $\exists c \in (a, b)$  úgy, hogy

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

Továbbá tegyük fel, hogy  $x^*$  és  $\tilde{x}^* \in [a, b]$ , és  $\max_{[a, b]} |f - \tilde{f}| < \varepsilon$ . Alkalmazzuk a Lagrange-közéérték tételt az  $[x^*, \tilde{x}^*]$  intervallumon (feltéve, hogy  $x^* < \tilde{x}^*$ ):

$$\exists c \in (x^*, \tilde{x}^*) : f(\tilde{x}^*) - f(x^*) = f'(c)(\tilde{x}^* - x^*)$$

Tegyük fel, hogy  $f'(x) \neq 0 \quad \forall x \in (x^*, \tilde{x}^*)$ .

$$\iff |\tilde{x}^* - x^*| = \left| \frac{f(\tilde{x}^*)}{f'(c)} \right| = \frac{|f(\tilde{x}^*) - \tilde{f}(\tilde{x}^*)|}{|f'(c)|} < \frac{\varepsilon}{\min_{[a, b]} |f'|}$$

**Definíció 6.1.1** Az  $M := \frac{1}{\min_{[a, b]} |f'|}$  számot a 6.1 egyenlet kondicionáltsági számának nevezzük.

Tehát ha  $\max_{[a, b]} |f - \tilde{f}| < \varepsilon$ , akkor  $|\tilde{x}^* - x^*| < M \cdot \varepsilon$ .

## 6.2. Konvergencia sebesség

Tegyük fel, hogy  $\lim_{k \rightarrow \infty} x_k = x^*$ , és legyen  $e_k := x_k - x^*$ . ( $\lim_{k \rightarrow \infty} e_k = 0$  vagy  $\lim_{k \rightarrow \infty} |e_k| = 0$ )

**Definíció 6.2.1** Azt mondjuk, hogy az  $(x_k)$  sorozat konvergencia rendje  $p \geq 1$ , ha

$$\lim_{k \rightarrow \infty} \frac{\log |e_k|}{\log |e_{k-1}|} = p$$

- Ha  $p = 1$ , akkor lineáris vagy elsőrendű konvergenciáról beszélünk.
- Ha  $p = 2$ , akkor másodrendű vagy kvadratis konvergenciáról beszélünk.

**Példa 5** Elsőrendű és másodrendű konvergens sorozatok hibatagjainak lecsengésére példák.

**Elsőrendű:**

	$ e_k $	$\frac{\log  e_k }{\log  e_{k-1} }$
$k = 1$	$10^{-3}$	$N/A$
$k = 2$	$10^{-4}$	1.33
$k = 3$	$10^{-5}$	1.25



**Másodrendű:**

	$ e_k $	$\frac{\log e_k }{\log e_{k-1} }$
$k = 1$	$10^{-3}$	$N/A$
$k = 2$	$10^{-6}$	2
$k = 3$	$10^{-12}$	2

**Állítás 6.2.1** Tegyük fel, hogy  $|e_k| = c_k \cdot |e_{k-1}|$ ,  $k = 1, 2, \dots$  ahol  $0 < \underline{c} \leq c_k \leq \bar{c} < 1$ . Valamilyen  $\underline{c}$  és  $\bar{c}$  konstansokra.

Ekkor  $x_k \rightarrow x^*$  monoton módon és elsőrendben.

*Bizonyítás:* Monotonan, mivel  $0 < c_k < 1 \implies |e_k| < |e_{k-1}| \quad \forall k = 1, 2, \dots$   
 $\implies (|e_k|)$  sorozat monoton csökkenő.

*Konvergál,* mivel  $|e_k| = c_k \cdot |e_{k-1}| \leq \bar{c} \cdot |e_{k-1}| \leq \bar{c} \cdot \bar{c} \cdot |e_{k-2}| \leq \dots \leq \bar{c}^k \cdot |e_0|$ . Mivel  $\bar{c} < 1$  ezért tényleg  $\lim_{k \rightarrow \infty} |e_k| = 0$ .

A feltételben lévő egyenletnek mindkét oldalán logaritmust véve:

$$\begin{aligned} \log|e_k| &= \log c_k + \log|e_{k-1}| \\ \implies \frac{\log|e_k|}{\log|e_{k-1}|} &= \frac{\log c_k}{\log|e_{k-1}|} + 1 \end{aligned}$$

Ltszik, hogy  $\log|e_{k-1}| \rightarrow -\infty$ . Mostmár elegendő lenne belátni, hogy  $\log c_k$  korlátos.

$$\log \underline{c} < \log c_k \leq 0$$

Tehát  $\frac{\log c_k}{\log|e_{k-1}|} \rightarrow 0 \implies$  a jobb oldal  $\rightarrow 1 \implies p = 1$  a konvergencia rendje, azaz elsőrendű a konvergencia.

**Állítás 6.2.2** Tegyük fel, hogy  $|e_k| = c_k \cdot |e_{k-1}|^p$   $k = 1, 2, \dots$  ahol  $p > 1$  és  $0 < \underline{c} \leq c_k \leq \bar{c} < +\infty$ . Valamilyen  $\underline{c}$  és  $\bar{c}$  konstansokra. Továbbá  $\bar{c}^{1/p-1} \cdot |e_0| < 1$ . Ekkor  $(x_k)$  konvergens és a konvergencia rendje  $p$ .

**Megjegyzés 8** Az utóbbi feltétel azt jeletnti, hogy a konvergencia csak akkor következik, ha  $x_0$  elég közel van  $x^*$ -hoz. Ugyanakkor  $\bar{c} < +\infty$ , és nem kell teljesülnie, hogy  $\bar{c} < 1$ .

## 6.3. Intervallum felezés

Megoldandó feladat:  $f(x) = 0$

Feltevés:

- $f \in C[a, b]$
- $f(a)f(b) < 0$

Ekkor a Bolzano tétel szerint  $\exists x^* \in (a, b)$ , ahol  $f(x^*) = 0$ . Miután a Bolzano tétel biztosítja nekünk a gyök létezését, keressük meg, hogy hol van ez a gyök.

Felépítünk egy intervallumsorozatot:  $I_0 := [a, b]$  Felezzük meg ezt az intervallumot, legyen  $c = \frac{a+b}{2}$ . Ezután vizsgáljuk  $f(c)$  előjelét:

- $f(c) = 0$  ekkor készen is vagyunk mert találtunk egy gyököt.
- Ha  $f(c) \neq 0$ , akkor  $I_1 := [a, c]$  vagy  $I_1 := [c, b]$ , azt az intervallumot választuk melyben az intervallum szélein az  $f$  értéke ellentétes előjelű.

Megfelezzük  $I_1$ -et és folytatjuk az eljárást. Tehát megint megnézzük az intervallum felét és választjuk azt a felet, melyben az intervallum szélein az  $f$  értéke ellentétes előjelű.

Az iteráció során mindig marad gyök az aktuálisan vizsgált intervallumban és mindig feleződik az intervallum hossza.

Látszik, hogy nem mindig fogunk olyan esetre találni, ahol  $f(c) = 0$  ls pontosan megtaláltuk a függvény gyökét, például  $f(x) = x - \sqrt{2}$  függvénynek irracionális a gyöke de az iteráció során csak racionális pontokat vizsgálunk.

Tehát érdemes megbeszélni, hogy milyen pontossággal szeretnénk közelíteni a gyököt és mikor álljuk le.

Folytassuk addig az iterációt ameddig az aktuálisan vizsgált intervallum hossza nem éri el az előírt  $\varepsilon > 0$  pontosságot. Ekkor leállunk és válasszuk az aktuálisan vizsgált intervallum bármelyik pontját közelítő megoldásnak, mert az intervallumban minden pont legfeljebb  $\varepsilon$  távolságra lesz a valós gyöktől.

Meg lehet mondani előre, hogy hány iteráció után kell majd leállnunk?

Jelölés:  $\text{diam}(I_k) := I_k$  hossza

$\text{diam}(I_k) = \frac{b-a}{2^k} < \varepsilon$  ebből következik, hogy  $k > \frac{\log(\frac{b-a}{\varepsilon})}{\log(2)}$ . Észrevétel: A lépésszám teljesen független az  $f$  függvénytől, de hát miért is függne, mert mindig csak intervallumokkal dolgozunk és az  $f$  függvényt csak a következő intervallum kiválasztására használjuk, ami lehetne akár egy pénzérme dobás is.

Érdemes lenne beszélni még a konvergencia sebességéről.

$$|x_k - x^*| \leq \text{diam}(I_k)$$

ezen felsőkorlátok sorozata lineárisan konvergens, mert  $\text{diam}(I_k)$  mindig feleződik és az előző fejezetben megbeszéltük, hogy ha a hibatag valahányadrészt csökken akkor a konvergencia lineáris. ( $c_k := \frac{1}{2} \quad \forall k$ , láasd első állítás múlt óráról)

## 6.4. Egyszerű iteráció (fixpont-iteráció)

Megoldandó feladat:  $f(x) = 0$

Írjuk át a következő alakra:

$$\varphi(x) = x$$

ahol  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  valamilyen függvény. Ekkor  $f$  gyöke pontosan a  $\varphi$  fixpontja.

Érvényes a fixponttétel a következő változata:

**Tétel 6.4.1** Legyen  $H \subset \mathbb{R}$  zárt halmaz, és  $\varphi : H \rightarrow H$  kontrakció, tehát  $\exists q \in (0, 1)$ , melyre  $|\varphi(x) - \varphi(y)| \leq q \cdot |x - y| \quad \forall x, y \in H$ . Ekkor

- egyértelműen létezik  $\varphi$ -nek fixpontja, azaz  $\exists! x^*$  melyre  $\varphi(x^*) = x^*$
- tetszőleges  $x_0 \in H$  kezdőpontot választva a következő módon definiált sorozat konvergens és tart  $x^*$ -hoz

$$x_{k+1} = \varphi(x_k)$$

- a következő módon tudjuk becsülni a konvergencia sebességét  $|x_k - x^*| \leq \frac{q^k}{1-q} \cdot |x_1 - x_0|$

**Kérdés 14** Mikor kontrakció  $\varphi$ ?

Vegyük észre, hogy valamilyen módon a  $\varphi'$  abszolútértékétől függ, hogy kontrakció-e a  $\varphi$ .

**Állítás 6.4.1** Tegyük fel, hogy  $\varphi \in C(I)$  és  $\varphi \in D(\text{int}(I))$  tehát folytonos az intervallumon és differenciálható a belsejében. Ha  $\exists q \in [0, 1)$ , amely mellett  $|\varphi'(x)| \leq q \quad \forall x \in \text{int}(I)$ , akkor  $\varphi$  kontrakció  $I$ -n a  $q$  kontrakciószámmal.

*Bizonyítás:* Legyen  $x, y \in I$  két tetszőleges pont,  $x < y$  Alkalmazzuk  $\varphi$ -ra  $[x, y]$  intervallumon a Lagrange-középérték-tételt: Létezik  $c \in (x, y)$  melyre

$$\varphi(y) - \varphi(x) = \varphi'(c) \cdot (y - x)$$

Vegyünk mindkét oldalt abszolút értéket:

$$|\varphi(x) - \varphi(y)| = |\varphi'(c)| \cdot |x - y| \leq q \cdot |x - y| \quad \forall x, y \in I$$

Az egyenlőtlenség a feltétel miatt áll.

**Példa 6**  $\varphi(x) = \frac{1}{2} \cos(x)$  kontrakció-e a  $[0, \frac{\pi}{2}]$  intervallumon? És ha igen mi a  $q$  kontrakciósám?  $|\varphi'(x)| = \left| -\frac{1}{2} \sin x \right| \leq \frac{1}{2} < 1 \quad \forall x \in [0, \frac{\pi}{2}]$  (sőt  $\forall x \in \mathbb{R}$ ) Tehát  $\varphi$  kontrakció és  $q = \frac{1}{2}$  jó választás kontrakciószámmra.

**Kérdés 15** Mi a konvergencia rendje?

**Állítás 6.4.2** Tegyük fel, hogy  $\varphi \in C^p[a, b]$ , azaz  $p$ -szer folytonosan deriválható, és  $\varphi$  beleképez  $[a, b]$ -be és  $\varphi$  kontrakció  $[a, b]$ -n. Ha az  $x^*$  fixpontban a következő igazak:

$$\begin{aligned}\varphi'(x^*) &= 0 \\ \varphi''(x^*) &= 0 \\ \varphi'''(x^*) &= 0 \\ \varphi^{(4)}(x^*) &= 0 \\ &\vdots \\ \varphi^{(p-1)}(x^*) &= 0 \\ \varphi^{(p)}(x^*) &\neq 0\end{aligned}$$

Ekkor tetszőleges  $x_0 \in [a, b]$  pontból indítva a fixpont iterációt  $p$ -ed rendben konvergesn.

*Bizonyítás:* A konvergenciát biztosítja a fixpont tétel, tehát elég a konvergencia rendjét belátni. Írjuk fel  $\varphi$ -nek  $x^*$  körüli  $p - 1$ -ed fokú Taylor polinomjának a hibáját az  $x_k$  pontban  $\exists \vartheta_k$  az  $x^*$  és  $x_k$  között

$$\begin{aligned}\varphi(x_k) - T_{p-1}(\varphi(x_k), x^*) &= \frac{\varphi^{(p)}(\vartheta_k)}{p!} (x_k - x^*)^p \\ \varphi(x_k) + \varphi(x^*) + 0 + 0 + \dots + 0 &= \frac{\varphi^{(p)}(\vartheta_k)}{p!} (x_k - x^*)^p \\ x_{k+1} - x^* &= \frac{\varphi^{(p)}(\vartheta_k)}{p!} (x_k - x^*)^p\end{aligned}$$

Vegyük ezt abszolút értékben és vizsgáljuk így a konvergencia rendjét

$$\begin{aligned}|x_{k+1} - x^*| &= \left| \frac{\varphi^{(p)}(\vartheta_k)}{p!} (x_k - x^*)^p \right| \\ |x_{k+1} - x^*| &= \left| \frac{\varphi^{(p)}(\vartheta_k)}{p!} \right| \cdot |e_k|^p \\ |x_{k+1} - x^*| &= c_k \cdot |e_k|^p\end{aligned}$$

Kell még:  $0 < \underline{c} \leq c_k \leq \bar{c} < +\infty$   $\varphi \in C^p[a, b]$  és  $\varphi^{(p)}(x^*) \neq 0$  ekkor  $|\varphi^{(p)}(x)|$   $x^*$  egy kis környezetében is pozitív. Ha  $k$  elég nagy, akkor mivel  $\vartheta_k$   $x_k$  és  $x^*$  között van

$$\left| \frac{\varphi^{(p)}(\vartheta_k)}{p!} \right|$$

beszorítható két pozitív konstans közé.

## 6.5. Newton módszer (érintő módszer)

Megoldandó feladat:  $f(x) = 0$

Alapötlet:

1. Tegyük fel, hogy  $f$  differenciálható
2. Vegyünk fel egy tetszőleges  $x_0 \in D(f)$  kezdőpontot.
3. Húzzuk itt meg  $f$  érintőjét.
4. Ennek  $x$  tengellyel való metszéspontja legyen  $x_1$
5. Folytassuk  $x_1$ -el az iterációt

Megfelelő feltételekkel  $x_1, x_2, \dots \rightarrow x^*$

**Kérdés 16** *Mindig működik ez az eljárás?*

A módszer képlete:  $x_k$ -beli érintő:  $x = f'(x_k)(x - x_k) + f(x_k)$   $x$  tengellyel metszéső pontja:

$$\begin{aligned}
 0 &= f'(x_k)(x - x_k) + f(x_k) \\
 -f(x_k) &= f'(x_k)(x - x_k) \\
 -\frac{f(x_k)}{f'(x_k)} &= x - x_k \\
 x_{k+1} &= x_k - \frac{f(x_k)}{f'(x_k)}
 \end{aligned}$$

**Kérdés 17** *Mit lehet mondani a Newton módszer konvergencia rendjéről?*

**Állítás 6.5.1** *Tegyük fel, hogy az  $x^*$  gyököt és az egész  $(x_k)$  sorozatot tartalmazó valamely  $I$  intervallumban  $f \in C^2(I)$ , továbbá  $\exists m_1, M_1, m_2, M_2 > 0$  konstansok, amelyekkel*

$$m_1 \leq |f'(x)| \leq M_1$$

és

$$m_2 \leq |f''(x)| \leq M_2$$

Ekkor  $\frac{M_2}{2m_1}|e_0| < 1$  esetén a Newton módszer másodrendben konvergens.

*Bizonyítás:* Írjuk fel az  $f$  függvény  $x_k$  körüli elsőfokú Taylor polinomjának hibáját az  $x^*$  pontban!

$$f(x^*) - f(x_k) - f'(x_k)(x^* - x_k) = \frac{f''(\vartheta_k)}{2!}(x^* - x_k)^2$$

ahol  $\vartheta_k$  valamely pont az  $x_k$  és  $x^*$  között. Utána osszunk  $f'(x_k)$ -val mindkét oldalt

$$\begin{aligned}
 0 - \frac{f(x_k)}{f'(x_k)} - (x^* - x_k) &= \frac{f''(\vartheta_k)}{2f'(x_k)}(x^* - x_k)^2 \\
 x_{k+1} - x^* &= \frac{f''(\vartheta_k)}{2f'(x_k)}(x^* - x_k)^2 \\
 |e_{k+1}| &= \left| \frac{f''(\vartheta_k)}{2f'(x_k)} \right| \cdot |e_k|^2 \\
 |e_{k+1}| &= c_k \cdot |e_k|^2
 \end{aligned}$$

Kell még, hogy  $0 < \underline{c} \leq c_k \leq \bar{c} < +\infty$ . A feltétel szerint

$$0 < \frac{m_2}{2M_1} \leq c_k \leq \frac{M_2}{2m_1} < +\infty$$

$$\frac{M_2^{1/2-1}}{2m_1} \cdot |e_0| = \frac{M_2}{2m_1} < 1$$

esetén másodrendű a konvergencia.

**Kérdés 18** Mikor teljesül, hogy  $(x_k) \subset I$ ?

Fontos tárgyalnunk a fenti kérdést, mivel az előzőekben megfogalmazott Newton-módszer csak akkor teljesülnek, ha nem lépünk ki az intervallumból. Erre a kérdésre adunk négy esetet, mikor az intervallumban maradunk és működik a Newton-módszer.

1.  $f$  konkáv és szigorúan monoton növvő. Ekkor vegyük fel  $x_0$ -t balra  $x^*$ -től ( $x_0 < x^*$ )
2.  $f$  konkáv és szigorúan monoton csökken. Ekkor vegyük fel  $x_0$ -t jobbra  $x^*$ -től ( $x_0 > x^*$ )
3.  $f$  konvex és szigorúan monoton nő. Ekkor vegyük fel  $x_0$ -t jobbra  $x^*$ -től ( $x_0 > x^*$ )
4.  $f$  konvex és szigorúan monoton csökken. Ekkor vegyük fel  $x_0$ -t balra  $x^*$ -től ( $x_0 < x^*$ )

## 6.6. Egyenletrendszerek megoldása

**Feladat:**  $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$  és  $f(x) = 0$

1. Egyszerű iterációt alkalmazzuk

$$f(x) = 0 \implies F(x) = x$$

Fixpont-iterációt alkalmazzuk:

$$x_{t+1} = F(x_t)$$

ahol  $x_0$  valamilyen kezdővektor.

2. Newton-módszer Skaláris esetben:  $x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$

Ennek analógiájára felírható a következő:

$$x_{t+1} = x_t - J_f^{-1}(x_t) \cdot f(x_t)$$

Megfelelő feltételek esetén másodrendben konvergál, ahogyan láttuk azt az egyváltozós esetben.

Hátránya ennek a módszernek, hogy az invertálás miatt nagyon költséges tud lenni. Ennek kiküszöbölésére lehet ezt a költséget csökkenteni, ha nem minden lépésben számoljuk újra a Jacobi-mátrix inverzét. Szokás például azt használni, hogy minden lépésben az  $x_0$ -beli Jacobi-mátrix inverzét használjuk. Ilyenkor csak elsőrendű konvergencia áll fent! Vagy lehet minden  $k$  lépésenként újraszámolni a Jacobi-mátrix inverzét.

## 7. fejezet

# Interpolációs feladatok

### 7.1. Interpolációs alapfeladat

Adott  $f : \mathbb{R} \rightarrow \mathbb{R}$  függvényt csak diszkrét pontokban ismerjük. Például csak diszkrét pontokban vannak méréseink egy adatról.

Elnevezések:

- $x_0, x_1, \dots, x_n$  - interpolációs alappontok,  $(x_0 < x_1 < \dots < x_n)$
- $f_k = f(x_k)$ ,  $k = 0, \dots, n$  - interpolált értékek

Cél: olyan folytonos függvényt keresünk, amely átmegy az összes ponton. (Megj.: Azért keresünk folytonos függvényt, mert a legtöbb analízisbeli tétel és állítás folytonos függvényekre szól.)

**Kérdés 19** Milyen típusú függvényt illesztünk?

Különösen kedvező tulajdonságúak a polinomok, tehát illesztünk polinomot.

**Kérdés 20** Hányadfokú polinomot illesztünk?

Legfeljebb  $n$ -ed fokú polinomot illesztünk. *Jelölés:*  $P_n$  jelöli a legfeljebb  $n$ -ed fokú polinomok halmazát.

$n := 1$  eset:

$$\exists! p \in P_1 \text{ melyre } p(x_0) = f_0 \text{ és } p(x_1) = f_1$$

$n := 2$  eset:

$$\exists p \in P_2 \text{ melyre } p(x_0) = f_0 \text{ és } p(x_1) = f_1 \text{ és } p(x_2) = f_2$$

**Tétel 7.1.1**  $\exists! p \in P_n$  melyre  $p(x_k) = f_k \quad \forall k = 0, 1, \dots, n$

*Bizonyítás:*

1. Létezés (konstruktívan)

Keressünk először olyan  $l_m \in P_n$  függvényt, amelyre

$$l_m(x_k) = \begin{cases} 1, & \text{ha } k = m \\ 0, & \text{ha } k \neq m \end{cases}$$

Mivel  $x_m$ -en kívül az alappontokban el kell tűnnie, tartalmaznia kell az  $(x - x_k)$ ,  $k \neq m$  gyöktényezőket, vagyis sa következő alakúnak kell lennie  $l_m(x)$ -nek

$$l_m(x) = c_m \cdot (x - x_0)(x - x_1) \cdot \dots \cdot (x - x_n) \cdot \frac{1}{x - x_m} = c_m \cdot \prod_{\substack{k=0 \\ k \neq m}}^n (x - x_k)$$

Már csak tudnunk kéne, hogy  $c_m$  micsoda.  $l_m(x_m) = 1$  tehát legyen

$$c_m \cdot \prod_{\substack{k=0 \\ k \neq m}}^n (x_m - x_k) = 1 \implies c_m = \frac{1}{\prod_{\substack{k=0 \\ k \neq m}}^n (x_m - x_k)}$$

$$\implies l_m(x) = \frac{1}{\prod_{\substack{k=0 \\ k \neq m}}^n (x_m - x_k)} \cdot \prod_{\substack{k=0 \\ k \neq m}}^n (x - x_k)$$

$$l_m(x) = \prod_{\substack{k=0 \\ k \neq m}}^n \frac{x - x_k}{x_m - x_k}$$

Mindegyik  $x_m$  alapponthoz tartozik egy ilyen  $l_m$  függvény. Ezekből készítsük el a következő  $p$  függvényt:

$$p(x) := \sum_{m=0}^n f_m \cdot l_m(x)$$

Ellenőrizzük, hogy ezen  $p$  függvény tényleg az amit kerestünk:

$$p(x_k) = \sum_{m=0}^n f_m \cdot l_m(x_k) = f_k \cdot l_k(x_k) = f_k \cdot 1 = f_k$$

Kell még, hogy  $p \in P_n$  Ez igaz, mert  $l_m \in P_n$  és  $P_n$  vektortér, tehát a lineáris kombinációjuk is eleme  $P_n$ -nek

2. Egyértelműség Tegyük fel, hogy  $p$  és  $q$  függvények is teljesítik a kívánt tulajdonságokat, azaz:

- $p, q \in P_n$
- $p(x_k) = f_k = q(x_k), \quad k = 0, \dots, n$



Legyen  $d := p - q$  Ekkor  $d \in P_n$  és  $d(x_k) = 0 \quad \forall k = 1, \dots, n$  Így  $d$  egy legfeljebb  $n$ -ed fokú polinom, melynek van  $n + 1$  darab különböző zérushelye, tehát  $d \equiv 0$ . Mert egy legfeljebb  $n$ -edfokú polinomnak legfeljebb  $n$  zérushelye lehet, kivéve ha az azonosan a 0 függvény.

*Elnevezések:*

- $l_m$  függvényeket \*Lagrange-féle interpolációs alappolinomok\*-nak nevezzük
- $p$  függvényt \*interpolációs polinomnak\* nevezzük
- $\sum f_m \cdot l_m$  alakot az interpolációs polinomnak a \*Lagrange-féle alak\*-jának nevezzük

Az egyik hátránya a Lagrange-interpolációnak, hogy ha új adatpont ékezik, akkor az összes eddigi munkánk megy a kukába és újra kell kezdeni az interpolációt.

Kiküszöblése ennek a hátránynak megoldható Newton-féle alakkal (ezt az interpolációt már láttuk első félévben algebrából és gyakorlaton is tárgyalni fogjuk).

## 7.2. Függvény approximáció interpolációval

Tegyük fel, hogy egy  $f : I \rightarrow \mathbb{R}$  folytonos függvényt az egész  $I$  intervallumon ismerjük. Szeretnénk polinommal közelíteni ezt a függvényt, hogy könnyebben tudjunk vele számolni.

*Ötlet:* Vegyünk fel adatpontokat ezen a függvényen és illesszünk interpolációs polinomot a felvett adatpontokra.

**Kérdés 21** *Mennyire halad közel az interpolációs polinom az eredeti függvényhez?*

**Tétel 7.2.1** *Legyen  $f \in C^{n+1}(I)$ , és  $x_0, x_1, \dots, x_n \in I$  alappontok,  $p$  pedig az  $(x_k, f(x_k))$  pontokon átmenő interpolációs polinom. Ekkor az  $x \in I$  pontot és az összes  $x_k$  alappontot tartalmazó legszűkebb intervallumban van olyan  $\xi$  pont, amelyre*

$$f(x) - p(x) = \frac{1}{(n+1)!} \omega_n(x) \cdot f^{(n+1)}(\xi)$$

ahol  $\omega_n(x) = \prod (x - x_k)$  az úgynevezett \*alappont polinom\*.

*Bizonyítás:* Két eset van:

- Ha  $x = x_k$  (valamelyik alappontra).
- Ha  $x \neq x_k$  (bármelyik alappontra).

Az első esetben nincs mit bizonyítani, mert ekkor mindkét oldalon 0 van és bármilyen  $\xi$ -re fenn áll az egyenlőség.

A második esetben tekintsük a következő segédfüggvényt:

$$g(t) = f(t) - p(t) - c \cdot \omega_n(t)$$

ahol  $c$  egy tetszőleges állandó.

$$g(x_k) = f(x_k) - p(x_k) - c \cdot \omega_n(x_k) = 0 \quad \forall k = 0, 1, \dots, n$$

Válasszuk meg a  $c$  konstans úgy, hogy  $g(x) = 0$  legyen.

$$\begin{aligned} g(x) &= f(x) - p(x) - c \cdot \omega_n(x) = 0 \\ \implies c &= \frac{f(x) - p(x)}{\omega_n(x)} \end{aligned}$$

Ezen  $c$  mellett  $g$ -nek van legalább  $n + 2$  zérushelye ( $x_0, x_1, \dots, x_n$  és  $x$ )

*Rolle-tétel emlék:*  $f \in C[a, b] \cap D(a, b)$   $f(a) = f(b)$  ekkor  $\exists c \in (a, b)$  melyre  $f'(c) = 0$

Rolle-tétel értelmében  $g'$ -nek van legalább  $n + 2 - 1 = n + 1$  darab zérushelye. Hasonló módon  $g^{(n+1)}$ -nek van legalább  $n + 1 - n = 1$  darab zérushelye.

Jelölje az egyik ilyen zérushelyet  $\xi$

Deriváljuk  $g(t)$  függvényt  $(n + 1)$ -szer, ekkor a következőt kapjuk:

$$g^{(n+1)}(t) = f^{(n+1)}(t) - p^{(n+1)}(t) - c \cdot (n+1)! = f^{(n+1)}(t) - 0 - c \cdot (n+1)! = f^{(n+1)}(t) - c \cdot (n+1)!$$

$t = \xi$  pontban a derivált:

$$\begin{aligned} 0 &= f^{(n+1)}(\xi) - \frac{f(x) - p(x)}{\omega_n(x)} \cdot (n+1)! \\ \implies f(x) - p(x) &= \frac{1}{(n+1)!} \omega_n(x) \cdot f^{(n+1)}(\xi) \end{aligned}$$

### 7.3. Hermite-interpoláció

Eddig az  $x_0, \dots, x_n$  pontokban csak a függvény értéket írtuk elő. Lehetséges általánosítása ennek a feladatnak, ha nem csak a függvény értékeket írjuk elő, hanem a pontbeli deriváltakat is. A legegyszerűbb formája ennek a feladatnak, amikor csak minden pontban az első deriváltat írjuk elő, de lehet ezt általánosabban is. Előírhatjuk minden pontban a magasabbrendű deriváltakat is, de ilyenkor elő kell írunk a legmagasabb deriváltig bezárólag az összes többi. Továbbá, nem feltétlenül kell minden pontban megadni az összes deriváltat. Tegyük fel, hogy az  $x_k$  alappontban az  $m_k$ -adikig írjuk elő az összes deriváltat, azaz adottak  $f_k, f_k^{(1)}, \dots, f_k^{(m_k)}$ . Összesen  $N = n + 1 + m_0 + m_1 + \dots + m_k$  feltétel van. Ezek alapján azt várjuk, hogy  $N - 1$ -ed fokú polinom egyértelműen illeszthető a pontokra. Ez így is van!

**Állítás 7.3.1** Egyértelműen létezik olyan  $H \in P_{N-1}$  polinom amelyre  $\forall k = 0, \dots, n \quad \forall i = 0, \dots, m_k \quad H^i(x_k) = f_k^{(i)}$

Ezt a polinomot Hermite-féle interpolációs polinomnak nevezzük. Speciálisan, ha  $m_k = 1$  minden  $k$ -ra, akkor úgy nevezik, hogy Hermite-Fejér-interpoláció. Ekkor az interpolált polinom fokszáma:  $N = 2n + 2 - 1 = 2n + 1$

## 7.4. Spline-interpoláció

Eddig a pontokra egyetlen polinomot illesztettünk. Hátránya ennek a megközelítésnek az, hogy ha sok pont van, akkor az illesztett polinom magas fokszámú lesz. Ekkor a deriváltja is magas fokszámú és ennek a szintén magas fokszámú polinomnak sok zérushelye van, ami azt mondja az eredeti polinomról, hogy sok helyen vízszintes a deriváltja. Tehát nagyon hullámos lesz az illesztett polinom sok pont esetén.

Egy megoldás erre a problémára lehet az, hogy ne egy polinomot próbáljunk illesztetni az összes pontra, hanem szakaszonként más-más polinomot illesszünk és ezeket az alacsony fokszámú polinomokat ragasszuk össze. Ebből az ötletből jön a Spline-interpoláció.

Legegyszerűbb, amikor szakaszonként lineáris-interpolációt alkalmazunk és így kapunk egy töröttvonalat, ami összeköti az adatpontjainkat.

## 7.5. Lineáris spline-interpoláció

Az  $[x_{k-1}, x_k]$  szakaszon

$$s_k(x) = f_{k-1} \cdot \frac{x - x_k}{x_{k-1} - x_k} + f_k \cdot \frac{x - x_{k-1}}{x_k - x_{k-1}}$$

a fenti képlet adja meg, hogy milyen értékeket fog felvenni a spline függvény az  $x \in [x_{k-1}, x_k]$  pontokban.

A teljes spline-függvény viszont a következőképpen néz ki:

$$s(x) = \begin{cases} s_1(x) & x \in [x_0, x_1] \\ s_2(x) & x \in [x_1, x_2] \\ \vdots & \\ s_n(x) & x \in [x_{n-1}, x_n] \end{cases}$$

Nyilvánvalóan az  $s$  spline-függvény folytonos az  $[x_0, x_n]$  intervallumon, viszont az alap-pontokban csúcsos, tehát nem differenciálható ezekben a pontokban.

Kiküszöbölése a nemdifferenciálhatóságnak az, hogy magasabbfokú interpolációkat használunk szakaszonként.

## 7.6. Kvadratikus spline-interpoláció

Az  $[x_{k-1}, x_k]$  szakaszon  $s_k \in P_2$  legyen. Ahhoz, hogy szakaszonként másodfokú polinomot illesszünk, elő kell írunk a függvényértékeken túl a pontbeli deriváltakat is ( $s'(x_0) = d_0$ )

1. Alkalmazzunk Hermite-interpolációt az  $[x_0, x_1]$  szakaszon:

$$\begin{aligned} s_1 \in P_1 : \quad & s_1(x_0) = f_0 \\ & s_1'(x_0) = d_0 \\ & s_1(x_1) = f_1 \end{aligned}$$

2. Hermite-interpoláció az  $[x_1, x_2]$  intervallumon:

$$\begin{aligned} s_2 \in P_2 : \quad & s_2(x_1) = f_1 \\ & s_2(x_2) = f_2 \\ & s_2'(x_1) := s_1'(x_1) \end{aligned}$$

3. Hasonlóképpen folytatjuk a további  $[x_{k-1}, x_k]$  intervallumokon.

Ezzel a módszerrel  $s$  folytonosan differenciálható lesz.

**Megjegyzés 9** Létezik trigonometrikus-interpoláció is. Tegyük fel, hogy  $f$  periodikus függvény  $2\pi$  intervallum hosszal, és a következő pontokban ismerjük a függvény értékeit a  $[0, 2\pi]$  intervallumnak az összes  $x_k = \frac{k}{n+1} \cdot 2\pi$  pontjában.

Keressük azt a

$$t_m(x) = a_0 + \sum_{j=1}^m a_j \cdot \cos(jx) + b_j \cdot \sin(jx)$$

trigonometrikus polinomot, amelyre az igaz, hogy  $t_m(x_k) = f_k$  minden  $k$ -ra. Az  $a_0, a_k, b_k$  együtthatókat diszkrét Fourier-együtthatóknak nevezzük.

## 7.7. Legkisebb négyzetek módszere

Felmerülhet az a baj, ha sok pont egy kupacban van és nagyjából egy alakban és pontosan szeretnénk ezekre egy polinomot illeszteni, akkor feleslegesen magasfokú lesz az illesztett polinom és nem is fogja megragadni a pontok alakjának a lényegét. Ennek orvosolására próbáljuk meg nem pontosan illeszteni polinomot a pontokra, hanem csak legyen az a célunk, hogy minden ponthoz a legközelebb haladjon az illesztett polinomunk.

Legyenek adva az  $(x_k, f_k)$  pontok. Olyan  $P$  polinomot keresünk, amely:

- Adott fokszámú (ezt mi döntjük el).

- Globálisan az összes ponthoz a legközelebb halad.

**Megjegyzés 10** *Figyelem, ezt nem nevezzük interpolációnak, mert nem megy át minden alapponton az illesztett polinom! Mégis ezt a módszert az interpolációkkal együtt tárgyaljuk, mert ezekhez a módszerekhez áll a legközelebb a tematikában.*

**Kérdés 22** *Hogyan mérjük a költséget?*

Lehetne azt csinálni, hogy a költség a következő:  $\sum |f_k - P(x_k)|$ . Ezzel a megközelítéssel az lesz a probléma, hogy nem deriválható, továbbá szeretnénk azt is, hogy nagy eltérés nagy hibát jelezzon.

Helyette legyen a költség függvényünk a következő:  $\sum (f_k - P(x_k))^2$

A fenti költség függvény már könnyen deriválható és nagy eltérésre sokkal nagyobb hibát jelez, mint kicsi hibára.

Tegyük fel, hogy  $p \in P_1$  polinomot akarunk illeszteni, tehát  $p(x) = c_0 + c_1x$  alakú függvényt szeretnénk illeszteni. Hogyan kell megválasztani  $c_0$  és  $c_1$  együtthatókat, hogy a következő kifejezés minimális legyen:

$$\sum_{k=0}^n (f_k - c_0 - c_1x_k)^2$$

Itt az  $x_k, f_k$  adottak, és a  $(c_0, c_1)$  számoktól függő

$$F(c_0, c_1) := \sum_{k=0}^n (f_k - c_0 - c_1x_k)^2$$

$\mathbb{R}^2 \rightarrow \mathbb{R}$  függvény minimum helyét keressük. Nézzük meg, hogy hol nulla a gradiens vektora.

$$\begin{aligned} \frac{\partial F}{\partial c_0} &= \sum_{k=0}^n 2(f_k - c_0 - c_1x_k) \cdot (-1) = 0 \\ \frac{\partial F}{\partial c_1} &= \sum_{k=0}^n 2(f_k - c_0 - c_1x_k) \cdot (-x_k) = 0 \end{aligned}$$

A következő egyenletrendszerre jutunk:

$$\begin{aligned} \sum_{k=0}^n (f_k - c_0 - c_1x_k) &= 0 \\ \sum_{k=0}^n (f_kx_k - c_0x_k - c_1x_k^2) &= 0 \end{aligned}$$

Ez egy lineáris algebrai egyenletrendszer  $(c_0, c_1)$ -re. Mátrix alakban:

$$\begin{bmatrix} n+1 & \sum x_k \\ \sum x_k & \sum x_k^2 \end{bmatrix} \cdot \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} \sum f_k \\ \sum f_kx_k \end{bmatrix}$$

Ezt az alakot már könnyen meg lehet oldani számítógéppel és kapunk egy-egy értéket  $c_0$  és  $c_1$ -re, amiből megkapjuk a  $p = c_0 + c_1x$  illesztett polinomot.

**Megjegyzés 11** *Ha  $n \geq 1$ , akkor egyértelműen létezik megoldás és ez tényleg minimum hely lesz.*

**Megjegyzés 12** *Ha  $N$ -ed fokú polinomot szeretnénk illeszteni, akkor  $(N + 1)$  ismeretlenes lineáris algebrai egyenletrendszert kapunk, aminek a megoldása megadja az illesztett polinom együtthatóit.*

## 8. fejezet

# Közelítő integrálás

Ebben a fejezetben arra az alapvető feladatra próbálunk megoldásokat adni, amikor egy függvényt integrálni szeretnénk de vagy túl költséges kiszámítani a primitív függvényt vagy nem is létezik. Például sokszor kell statisztikában és valószínűségyszámításban a standard normális eloszlás kvantiliseit számolni, azaz ki kell számolni a következő integrált:

$$\int_{-\infty}^x e^{-t^2} dt$$

Ennek az integrálnak nem tudjuk kiszámolni az értékét analitikusan, mivel köztudott, hogy  $e^{-t^2}$  függvénynek nem létezik primitív függvénye. Tehát, amikor meg kell mondanunk a standard normális eloszlás 0.05-ös kvantilisét ( $\Phi(0.05)$  értéket), akkor egy táblázatból kiolvassuk, ahol valamilyen numerikus módszerrel kiszámolták adott diszkrét értékekre.

**Feladat:** Adott egy  $f \in R[a, b]$  függvény, melynek szeretnénk az integrálját meghatározni, azaz  $\int_a^b f = ?$

Nyilván fel kell tennünk, hogy egyáltalán integrálható a függvény a megadott intervallumon. Továbbá, tudjuk ha  $f$ -nek létezik  $F$  primitív függvénye, akkor  $\int_a^b f = F(b) - F(a)$  a Newton-Leibniz szabály alapján.

Gyakran  $F$ -et nem tudjuk meghatározni, ekkor felmerül a megoldás, hogy hogyan tudjuk közelítőleg integrálni a függvényt.

**Ötlet:** Használjuk ki a  $\int_a^b f$  kifejezés geometriai jelentését, azaz azt, hogy az integrál a görbe alatti előjeles területet számolja.

Közelítsük a görbe alatti területet egyszerűbb alakzat területével. Ebből az ötletből kapjuk az alapvető *kvadrátúraformulákat*.

**Megjegyzés 13** A kvadrátúra elnevezés onnan ered, hogy a legalapvetőbb módja a görbe alatti terület kiszámításának az, hogy kellően sűrű négyzethálóra bontjuk a síkot és megszámláljuk a görbe alatti négyzeteket. Innen a négyzet szóból ered a kvadrátúra formula elnevezés.

Vezessük be az intervallum hosszára következő változót:  $h = b - a$

A következő pár alapvető kvadratúra formula:

1. **Középponti formula:**

$c := \frac{a+b}{2}$  pontban megnézzük a függvényértéket.  $k(f) := h \cdot f(c)$

2. **Trapéz formula:**

Megnézzük a  $(a, b, f(b), f(a))$  pontok által meghatározott trapéz területét.  $t(f) := h \cdot \frac{f(a)+f(b)}{2}$



# Irodalomjegyzék

---

- [1] Faragó István, H.R.: Numerikus módszerek. Typotex (2016)

