Numerikus módszerek 1 jegyzet

Toffalini Leonardo, Havasi Ágnes

2024. május 8.

ELTE

Tartalomjegyzék

1. Bevezetés **1**

2. Numerikus modellezés **3**
2.1. Numerikus modellezés lépései . . . 3
2.2. Hibaforrások . . . 4
2.3. Hibafogalmak . . . 4
2.4. Az alapműveletek hibája . . . 5
2.5. Korrekt kitűzésű feladatok . . . 5

3. Normált terek **7**
3.1. Normált tér . . . 7
3.2. Fontos fogalmak normált terekben . . . 8
3.3. Mátrixnormák . . . 9
3.4. Kondíciószám . . . 10

4. Lineáris algebrai egyenletrendszerek megoldása **13**
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

5. Gradiens alapú módszerek **23**
5.1. Gradiens módszer . . . 25
5.2. Konjugált gradiens-módszer . . . 25

6. Általános algebrai egyenletek megoldása **27**
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

7. Interpolációs feladatok **35**
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

8. Közelítő integrálás **43**
8.1. Kvadratúra formulák 44
8.2. Interpolációs típusú kvadratúra formulák 45
8.3. Összetett kvadratúra formulák 47
8.4. Gauss kvadratúrák 49

9. Numerikus deriválás **53**
9.1. Az első derivált közelítése 53
9.2. A második derivált közelítése 54
9.3. A lépéstávolság dillemája 55

10. Közönséges Differenciál Egyenletek (KDE-k) megoldása **57**
10.1. Véges külömbséges módszerek 57
10.2. Explicit Euler-módszer (EE) 58
10.3. Implicit Euler-módszer (IE) 58
10.4. Runge-Kutta módszerek 59

Irodalomjegyzék **61**

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 zsákmanyhalak 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 differenciálhányadossal
   - integrált helyettesítünk egy véges összeggel
   - végtelent helyettesítünk egy tetszőlegesen nagy termlszetes 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 valóéletben nem szimbólikusan számolunk valós számokkal, hanem egy számítógépre hagyjuk a számításokat. A számítógépünk viszont csak egy véges részhalmazát 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 nevezzuk, 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 alapműveletek hibája

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

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

1. **Összedaás**

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

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 bezetünk 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. $|x| \geq 0 \quad \forall x \in \mathbb{R}$ es $|x| = 0 \iff x = 0$
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 $\|\cdot\| : X \to \mathbb{R}$ egy függvény a következő tulajdonságokkal:*

1. $\|x\| \geq 0 \quad \forall x \in X$ *es* $\|x\| = 0 \iff x = 0_X$ *(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 értjuk.*

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 $p \to \infty$ *akkor* $\|x\|_p \to \|x\|_\infty \quad \forall x \in X$
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)* $\|f\|_{\int} := \int_a^b|f(x)|\,dx$

3.2. Fontos fogalmak normált terekben

Most hogy már kiterjesztettük a hossz fogalmát normált terekre, így képesek vagyunk az előző fejezetekben bezetett 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 $\|a - \tilde{a}\| \le \Delta_a$
   - $\tilde{a}$ relatív hibája: $\frac{a-\tilde{a}}{\|\tilde{a}\|} \in X$
   - $\tilde{a}$ relatív hibakorlátja: $\frac{\|a-\tilde{a}\|}{\|\tilde{a}\|} \le \delta_a \in \mathbb{R}$
2. **Konvergencia**

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

3.3. Mátrixnormák

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

Kérdés 1 *Hogyan definiálható ezen a vektortéren norma?*

Definíció 3.3.1 *Legyen $\|\cdot\|_{\mathbb{R}^\times}$ 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:
   $$\begin{bmatrix} -2 & 1 \\ 0 & 3 \end{bmatrix} \implies \|A\|_1 = \max\{|-2| + |0|, |1| + |3|\} = 3$$

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 sorosszeg!
Például:

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

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

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

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. $\|A\|_F = \sqrt{\sum_{i,j=n}^{n} a_{ij}^2}$ *(Frobenius norma)*

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

3.4. Kondíciószám

Az előbb meggondoltuk, hogy egy lineáris egyenletrendszernek, $Ax = b$-nek, az A együtthatómátrixának egy elemét kicsit pertulbálva 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 pertulbált $\tilde{b}$ van a jobb oldalon:

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

Jelölje:

$$\Delta x = x - \tilde{x} \implies \tilde{x} = x - \Delta x$$
$$\Delta b = b - \tilde{b} \implies \tilde{b} = b - \Delta b$$

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 alkalmazzük 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 $\|A^{-1}\| \cdot \|A\|$ számat az A mátrix kondíció számának nevezzük és* $\mathrm{cond}(A)$*-val jelöljük.*

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

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.

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

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

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

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-eliminácó. Az iterációs módszereknek viszont a lényege az, hogy egy vektorsorozatot generálnak, melyek tartanak a pontos megoldáshoz.

4.1. Gauss-elimináció

Megoldandó egyenletrendszer: $Ax = b, \quad A \in \mathbb{R}^{m\times m}, \quad \det A \neq 0, \quad b \in \mathbb{R}^m$
Lineáris algebrából tudjuk, hogy ezek a feltételek mellet egyértelműen létezik megoláds, tehát korrekt kitűzésű a feladat és van értelme nekiállni megoldani.

A lineáris egyenletrendszer teljes anyakönyvezet nevén a következő:

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

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

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

2. lépés: 4.1 segitsegevel a másodiktól az m-edik egyenletekből eliminájuk x_1-et, kivonva belőlük a 4.1-nek az a_{i1}-szereset.

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

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

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

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

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

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

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

Kompaktabb mátrix formába átírva:

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

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.

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

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 szakaszaval es 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*):

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

Állítás 4.1.1 *Ha* $\Delta_j \neq 0, \ \forall j \in \{1, \ldots, m\}$*, 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 létezzen 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 $A = L_1U_1 = L_2U_2$

$$\begin{aligned} L_2^{-1}L_1U_1 &= U_2 \\ L_2^{-1}L_1 &= U_2U_1^{-1} \end{aligned}$$

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. 2×2 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* $\Delta_j \neq 0, \quad \forall j = 1, \ldots, m$ *teljesül, ha A szimmetrikus pozitív definit mátrix (szpd).*
2. *A* $\Delta_j \neq 0, \quad \forall j = 1, \ldots, m$ *teljesül, ha A szigorúan dominans főátlójú, tehát* $\forall i = 1, \ldots, n$*-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 pozitiv definit mátrix, akkor létezik egy másik felbontása is:* $A = G \cdot G^T$*, ahol G alsó háromszög mátrix, pozitív főátlóval. (Cholesky-felbontas)*

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 tartsuk 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 elemeit is cseréljük. Tehát ha egy P mátrixszal permutáljuk az oszlopait A-nak, akkor mikor visszaolvassuk x megoldást, akkor P^{-1}-el meg kell szorozni elötte.

4.3. Klasszikus iterációs módszerek

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

Definíció 4.3.2 *az $f : \mathbb{R}^m \to \mathbb{R}^m$ 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 $f : \mathbb{R}^n \to \mathbb{R}^n$ 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 $x_n \to x^*$.*
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?*

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

Tegyük fel, hogy 4.2 átírható a következő alakra:

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

Ekkor az $f(x) := Qx+r$ ejlöléssel a fealdat megoldása az $f : \mathbb{R}^m \to \mathbb{R}^m$ 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 vektornomában), és $x_n \to x^*$, 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.

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

$$Q_{GS} = -(L + D)^{-1}U, \quad r_{GS} = (L + D)^{-1}b$$

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:

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

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 említet 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:

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

Tegyük fel, hogy A szimmetrikus pozitív definit (szpd). Tehát $A=A^T$, $x^TAx>0$, ha $x\neq 0$. Másképpen, $\exists\delta>0:(Ax,x)\geq\delta\cdot\|x\|^2$.
Jelölje x^* a 3.1 egyenelet megoldasat, azaz $Ax^*=b$ és $e_n:=x^n-x^*$ (az n-edik iterácio hibáját).

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

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

Bizonyítás:

$$x^n=e_n+x^*,\quad x^{n+1}=e_{n+1}+x^*\rightsquigarrow \text{ (SI)-be beirva}$$
$$B\frac{e_{n+1}+x^*-e_n-x^*}{\tau}+Ae_n+Ax^*=b$$
$$B\frac{e_{n+1}-e_n}{\tau}+Ae_n=0 \qquad \text{(3) hibaegyenlet}$$

Fejezzük ki e_{n+1}-el

$$\begin{aligned}Be_{n+1}&=(B-\tau A)e_n\\ e_{n+1}&=(I-\tau B^{-1}A)e_n\\ Ae_{n+1}&=(A-\tau AB^{-1}A)e_n\\ \Longrightarrow\ (Ae_{n+1},e_{n+1})&=(Ae_n-\tau AB^{-1}Ae_n,e_n-\tau B^{-1}Ae_n)\end{aligned}$$

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

Tudjuk, hogy

$$(AB^{-1}Ae_n,e_n)=(B^{-1}Ae_n,A^Te_n)=(B^{-1}Ae_n,Ae_n)=(Ae_n,B^{-1}Ae_n)$$

Tehát

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

Jelölje $J_n=(Ae_n,e_n)$. Ezzel

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

Ezzel $By_n=Ae_n$

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

$$\rightsquigarrow J_{n+1} = J_n - 2\tau((B - 0.5\tau A)y_n, y_n) \tag{4.4}$$
$$\implies J_{n+1} \leq J_n$$

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, jelöles $J^* := \lim_{n\to\infty} J_n$
4.4-ban vegyuünk limeszt $\rightsquigarrow$

$$J^* = J^* - 2\tau \lim_{n\to\infty}((B - 0.5\tau A)y_n, y_n)$$

$$\implies \lim_{n\to\infty}((B - 0.5\tau A)y_n, y_n) = 0$$

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:

$$((B - 0.5\tau A)y_n, y_n) \geq \delta \cdot \|y_n\|^2 \geq 0$$

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

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

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

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

4.9. SOR-módszer konvergenciája

Kérdés 8 *Hogyan válasszuk meg ω paramétert, hogy konvergáljon?*

Észrevétel: ω választása erősen függ A-tól.

Állítás 4.9.1 *Tetszőleges $A \in \mathbb{R}^{n\times m}$ eséten 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 iterácio konvergens, mert a Gauss-Seidel iteráció pont a SOR-modszer $\omega = 1$-el.*

5. fejezet

Gradiens alapú módszerek

Tekintsük megint a következő egyenletet.

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

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

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

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

ez differenciálható $\mathbb{R}^m$-en.
Célunk, hogy a $\phi(x)$ függvényt minimalizáljuk, tehát néézük meg, hogy hol lesz 0 a gradiense.

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

Ekkor pont az $r := b - Ax$ *maradékvektor* -1 szeresét kapjuk.
Hol 0 a gradiens?

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

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 ϕ-nek.

Kérdés 9 *Hogy néz ki a ϕ függvény?*

Példa 4 *Tekintsünk egy két dimenziós példát, ahol már a következő egyenletrendszernél 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 ϕ a legkisebb értéket?

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

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

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

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

megvalásztás esetén veszi föl!

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

Kérdés 11 *Hogyan válasszuk meg $p_1, p_2, \ldots$ 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:

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

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 megyunk p_i irányban ameddig nem érintjuk 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 koveto iranyoknak!

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

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

5.2. Konjugált gradiens-módszer

Az előbb láttuk be, hogy a gradiens módszernél a p_1 kerésési 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 es $y \in \mathbb{R}^m$ vektorok A-konjugáltak/ortogonálisak, ha $(x, Ay) = 0$.*

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

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

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

Ezen β_1-et választva, a $p_2 = r_1 - \beta_1 \cdot p_1$ irányba lépve az x^* minimum helybe lépunk! 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 $f : \mathbb{R} \to \mathbb{R}$ 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 megvaltoztatasara?*

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-középérték tétél) 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 $\max\limits_{[a,b]}|f - \tilde{f}| < \varepsilon$. Alkalmazzuk a Lagrange-középérték tételt az $[x^*, \tilde{x}^*]$ intervallumon (feltéve, hogy $x^* < \tilde{x}^*$):

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

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

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

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

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

6.2. Konvergencia sebesség

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

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

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

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

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

Elsőrendű:

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

Másodrendű:

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

Állítás 6.2.1 *Tegyük fel, hogy* $|e_k| = c_k \cdot |e_{k-1}|, \quad k = 1, 2, \ldots$ *ahol* $0 < \underline{c} \leq c_k \leq \overline{c} < 1$. *Valamilyen* $\underline{c}$ *és* $\overline{c}$ *konstansokra.*
Ekkor $x_k \to x^*$ *monoton módon és elsőrendben.*

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

Konvergál, mivel $|e_k| = c_k \cdot |e_{k_1}| \leq \overline{c} \cdot |e_{k-1}| \leq \overline{c} \cdot \overline{c} \cdot |e_{k-2}| \leq \cdots \leq \overline{c}^k \cdot |e_0|$. Mivel $\overline{c} < 1$ ezért tényleg $\lim_{k\to\infty} |e_k| = 0$.

A feltételben lévő egyenletnek mindkét oldalán logaritmust véve:

$$\log|e_k| = \log c_k + \log|e_{k-1}|$$

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

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

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

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

Állítás 6.2.2 *Tegyük fel, hogy* $|e_k| = c_k \cdot |e_{k-1}|^p \quad k = 1, 2, \ldots$ *ahol* $p > 1$ *és* $0 < \underline{c} \leq c_k \leq \overline{c} < +\infty$. *Valamilyen* $\underline{c}$ *és* $\overline{c}$ *konstansokra. Továbbá* $\overline{c}^{1/p-1} \cdot |e_0| < 1$. *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* $\overline{c} < +\infty$*, és nem kell teljesülnie, hogy* $\overline{c} < 1$.

6.3. Intervallum felezés

Megoldandó feladat: $f(x) = 0$

Feltevések:

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

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 $c = \frac{a+b}{c}$ 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 intebrallumot választuk melyben az inteballum 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 aktuálisa 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 ε távolságra lesz a valós gyöktől.

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

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

$\operatorname{diam}(I_k) = \frac{b-a}{2^k} < \varepsilon$ ebből következik, hogy $k > \frac{\log\left(\frac{b-a}{\varepsilon}\right)}{\log(2)}$ 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.

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

ezen felsőkorlátok sorozata linárisan konvergens, mert $\operatorname{diam}(I_k)$ mindig feleződik és az előző fejezetben megbeszéltük, hogy ha a hibatag valahanyadrészére 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 $\varphi : \mathbb{R} \to \mathbb{R}$ valamilyen függvény. Ekkor f gyöke pontosan a φ fixpontja.

Érvényes a fixponttétel a következő változata:

Tétel 6.4.1 *Legyen $H \subset \mathbb{R}$ zárt halmaz, és $\varphi : H \to H$ 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 φ-nek fixpontja, azaz $\exists! x^*$ melyre $\varphi(x^*) = x^*$*
- *tetszőleges $x_0 \in H$ kezdőpontot választva a kovetkező 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 φ?*

Vegyük észre, hogy valamilyen módon a φ' abszolútértékétől függ, hogy kontrakció-e a φ.

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

Bizonyítás: Legyen $x, y \in I$ két tetszőleges pont, $x < y$ Alkalmazzuk φ-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 $\left[0, \frac{\pi}{2}\right]$ intervallumon? És ha igen mi a q kontrakciószám? $|\varphi'(x)| = \left|-\frac{1}{2}\sin x\right| \leq \frac{1}{2} < 1 \quad \forall x \in [0, \frac{\pi}{2}]$ (sőt $\forall x \in \mathbb{R}$) Tehát φ kontrakció és $q = \frac{1}{2}$ jó választás kontrakciószámra.*

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 φ beleképez $[a,b]$-be és φ 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 kovergencia rendjét belátni. Írjuk fel φ-nek x^* körüli $p-1$-ed fokú Taylor polinomjának a hibáját az x_k pontban $\exists \vartheta_k$ az x^* és x_k között

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

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: $0 < \underline{c} \leq c_k \leq \overline{c} < +\infty$ $\varphi \in C^p[a,b]$ és $\varphi^{(p)}(x^*) \neq 0$ ekkor $|\varphi^{(p)}(x)|$ x^* egy kis környezetében is pozitív. Ha k elég nagy, akkor mivel ϑ_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 $x_1, x_2, \cdots \to x^*$

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 metszéspontja:

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

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 tartalmlazó 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 φ_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 $0 < \underline{c} \leq c_k \leq \overline{c} < +\infty$. 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 előzőkben 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 nővő. Ekkor vegyük fel x_0-t balra x^*-tól ($x_0 < x^*$)
2. f konkáv és szigorúan monoton csökken. Ekkor vegyük fel x_0-t jobbra x^*-tól ($x_0 > x^*$)
3. f konvex és szigorúan monoton nő. Ekkor vegyük fel x_0-t jobbra x^*-tól ($x_0 > x^*$)
4. f konves és szigorúan monoton csökken. Ekkor vegyük fel x_0-t balra x^*-tól ($x_0 < x^*$)

6.6. Egyenletrendszerek megoldása

Feladat: $f : \mathbb{R}^m \to \mathbb{R}^m$ 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: $x_{t+1} = x_k - \frac{f(x_k)}{f'(x_k)}$
   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 minen k lépésenként újraszámolni a Jacobi-mátrix inverzét.

7. fejezet

Interpolációs feladatok

7.1. Interpolációs alapfeladat

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

Elnevezések:

- $x_0, x_1, \ldots, x_n$ - interpolációs alappontok, $(x_0 < x_1 < \cdots < x_n)$
- $f_k = f(x_k), \quad k = 0, \ldots, n$ - 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 illeszünk?*

Különösen kedvező tulajdonságúak a polinomok, tehát illesszünk polinomot.

Kérdés 20 *Hanyadfokú polinomot illesszünk?*

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

$n := 1$ eset:

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

$n := 2$ eset:

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

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

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 $(x - x_k), \quad k \neq m$ gyöktényezőket, vagyi sa következő alakúnak kell lennie $l_m(x)$-nek

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

   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:

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

   Ellenőrízzük, 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:

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

Legyen $d := p - q$ Ekkor $d \in P_n$ és $d(x_k) = 0 \quad \forall k = 1, \ldots, n$ Így d egy legfeljebb n-ed fokú polinom, melynek van $n + 1$ darab különböző zéruzhelye, 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 $f : I \to \mathbb{R}$ 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 $x_0, x_1, \ldots, x_n \in I$ 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 ξ 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 ξ-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.

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

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

$$g(x) = f(x) - p(x) - c \cdot \omega_n(x) = 0$$

$$\Longrightarrow \; c = \frac{f(x) - p(x)}{\omega_n(x)}$$

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

Rolle-tétel emlék: $f \in C[a,b] \cap D(a,b) \quad f(a) = f(b)$ 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 ξ

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:

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

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

7.3. Hermite-interpoláció

Eddig az $x_0, \ldots, x_n$ 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 írnunk a legmagasabb deriváltig bezárólag az összes többit. 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 $f_k, f_k^{(1)}, \ldots, f_k^{(m_k)}$ Összesen $N = n + 1 + m_0 + m_1 + \cdots + m_k$ 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 $\forall k = 0, \ldots, n$ $\forall i = 0, \ldots, m_k$ számokra $H^i(x_k) = f_k^{(i)}$*

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:

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

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

Kiküszöbölése a nemdifferenciálhatóságnak az, hogy magasabbfokú interplolációkat 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 írnunk 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 2π intervallum hosszal, és a következő pontokbab ismerjük a függvény értékeket a $[0, 2\pi]$ intervallumnak az összes $x_k = \frac{k}{n+1} \cdot 2\pi$ pontjában.*

Keressük azt a

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

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 illesztetni, 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 haladjonaz 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 inerpolációkkal 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 jelezzen.

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 $p(x) = c_0 + c_1 x$ alakú függvényt szeretnénk illeszteni. Hogyan kell megválasztani c_0 és c_2 együtthatókat, hogy a következő kifejezés minimális legyen:

$$\sum_{k=0}^{n}(f_k - c_0 - c_1 x_k)^2$$

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

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

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

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

$$\frac{\partial F}{\partial c_1} = \sum_{k=0}^{n} 2(f_k - c_0 - c_1 x_k) \cdot (-x_k) = 0$$

A következő egyenletrendszerre jutunk:

$$\sum_{k=0}^{n}(f_k - c_0 - c_1 x_k) = 0$$

$$\sum_{k=0}^{n}(f_k x_k - c_0 x_k - c_1 x_k^2) = 0$$

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

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

Ezt az alakot már könnyen meg lehet oldani számítógéppel és kapunk egy-egy értéket c_0 és c_1-re, amiből megkapjuk a $p = c_0 + c_1 x$ 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 valószínúségszámításban 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 kvanntilisét ($\Phi(0.05)$ értéket), akkor egy táblázatból kiolvassuk, ahol valamilyen numerikus módszerrel kiszámoltál 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, tdujuk 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: Haszáljuk 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 ötleből kapjuk az alapvető *kvadratúraformulákat.*

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

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

8.1. Kvadratúra formulák

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. $t(f) := h \cdot \frac{f(a)+f(b)}{2}$

Kérdés 23 *Hogyan tudjuk jellemezni, hogy egy közelítő integrál formula mennyire jó?*

Egy szempont: Mennyire jól viselkedik polinomokon.

Definíció 8.1.1 *Azt mondjuk, hogy a kvadratúra formula rendje n, ha a formula pontos $\forall f \in P_{n-1}$ polinomra, de létezik olyan n-ed rendű polinom amire már nem pontos.*

például:

- $t(f)$ azaz, a trapéz formula másodrendű
- $k(f)$ azaz, a középponti formula is másodrendű

Kérdés 24 *Hogyan kaphatunk magasabbrendű kvadratúra formulákat?*

Alkalmazzuk a $k(f)$ és a $t(f)$ formulákat az $f(x) = x^2$ függvényre a $[0, 1]$ intervallumon!

pontos értékek:

- $T = \int_0^1 x^2 \, dx = \frac{1}{3}$
- $k(f) = h \cdot f(c) = 1 \cdot \left(\frac{1}{2}\right)^2 = \frac{1}{4}$
- $t(f) = h \cdot \frac{f(0)+f(1)}{2} = 1 \cdot \frac{0^2+1^2}{2} = \frac{1}{2}$

Hibák:

- $T - k(f) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$
- $T - t(f) = \frac{1}{3} - \frac{1}{2} = -\frac{1}{6}$

Észrevehető, hogy a trapéz formula hibája -2 -szerese a középponti formula hibájának. Azaz, $T - t(f) = -2 \cdot (T - k(f))$ Kifejezve T értékét:

$$T = \frac{2k(f) + t(f)}{3}$$

Az így kapott formula az $f(x) = x^2$ függvényre a $[0, 1]$ intervallumon pontos.
Elnevezés: *Simpson formula* (vagy *parabola formula*):

$$s(f) = \frac{2}{3}k(f) + \frac{1}{3}t(f) = \frac{2}{3}h \cdot f(c) + \frac{1}{3} \cdot \frac{h}{2}(f(a) + f(b)) = \frac{h}{6} \cdot (f(a) + 4f(c) + f(b))$$

Figyelem: A rendjét a Simpson formulának még nem tudjuk, csak azért mert egy specifikus másodfokú polinomra pontos egy specifikus intervalliumon.

8.2. Interpolációs típusú kvadratúra formulák

Ötlet: Legyen $f : [a, b] \to \mathbb{R}$ és $x_0, x_1, \ldots, x_n \in [a, b]$ és p legyen az $(x_k, f(x_k))$ pontokra illesztett interpolációs polinom. Ekkor $\int_a^b f(x)\,dx$ értékét közelítsük $\int_a^b p(x)\,dx$ értékkel.

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

Tehát

$$\int_a^b f(x)\,dx \approx \int_a^b p(x)\,dx = \int_a^b \sum_{m=0}^{n} f_m \cdot l_m(x)\,dx$$

Azt keressük, hogy az így közelített érték mennyire tér el a pontos értéktől, azaz $\int_a^b f(x)\,dx - \int_a^b p(x)\,dx = ?$
Alkalmazzuk az interpolációs polinom hibájáról szóló tételt.
Tegyük fel, hogy $f \in C^{n+1}[a, b]$ legyen $x \in [a, b]$ tetszőleges pont. Ekkor $\exists \xi \in (\min\{x, x_0\}, \max\{x, x_n\})$

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

De minket az intergálok külömbsége érdekel, tehát integráljunk mindkét oldalon!

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

Fontos: Nem emelhetjük ki $f^{(n+1)}(\xi)$ számot a jobboldalon az integrandusból, mivel ξ értéke függ x értékétől.

Tekintsük az interpolációs kvadratúra formulát:

$$\int_a^b p(x)\,dx = \int_a^b \sum_{m=0}^{n} f_m \cdot l_m(x)\,dx = \sum_{m=0}^{n} \int_a^b f_m \cdot l_m(x)\,dx = \sum_{m=0}^{n} \left(\int_a^b l_m(x)\,dx \right) \cdot f_m$$

Az átalakítások után azt kapjuk, hogy az illesztett polinom integrálja az egy lineáris kombinációja az f_m értékeknek.

Definíció 8.2.1 *A $\sum_{m=0}^{n} A_m \cdot f_m$ alakú kvadratúra formulát, ahol $A_m \in \mathbb{R}$ lineáris kvadratúra formulának *nevezzük.*

Tehát az interpolációs kvadratúra formula is lineáris, ahol $A_m = \int_a^b l_m(x)\,dx, \quad m = 0, \ldots, n$

Felmerülhet a kérdés, hogy tudunk-e jobb lineáris kvadratúra formulát kitalálni, mint az interpolációs kvadratúra formula.

Tétel 8.2.1 *A $\sum_{m=0}^{n} A_m \cdot f_m$ lineáris kvadratúra formula akkor és csak akkor pontos $\forall f \in P_n$ polinomra, ha interpolációs típusú, azaz ha $A_m := \int_a^b l_m(x)\,dx, \quad m = 0, \ldots, n$*

Bizonyítás: Kezdjünk a könnyebb iránnyal, azaz lássuk be azt, hogy az interpolációs kvadratúra formula pontos minden n-ed rendű polinomra.
Ez az állítás könnyen következik abból a tényből, hogy $\forall f \in P_n$ polinomnak az $n+1$ darab pontra támaszkodó interpolációs polinomja saját maga, tehát a formula pontos.

A nehezebb irány az, hogy ha a formula pontos minden n-ed rendű polinomra, akkor az a formula interpolációs típusú.
Tegyük fel, hogy $\sum_{m=0}^{n} A_m \cdot f_m$ pontos $\forall f \in P_n$ polinomra. Be kell látnunk, hogy $A_j = \int_a^b l_j\,dx, \quad j = 0, \ldots, n$
Tudjuk, hogy $l_j \in P_n$ azaz a feltételünk szerint ezekre az l_j polinomokra pontos a formula. Ekkor l_j közelítő integrálja a következő:

$$\sum_{m=0}^{n} A_m \cdot l_j(x_m) = \int_a^b l_j(x)\,dx$$

Mivel az l_j polinomok úgy voltak definiálva, hogy $l_j(x_i) = \delta_{ij}$ azaz $l_j(x_j) = 1$ és $l_j(x_i) = 0$ ha $i \neq j$
Tehát a fenti egyenlet baloldalán a szumma majdnem minden tagja kiesik kivéve $l_j(x_j)$ tehát a fenti egyenlet a következőre redukálódik:

$$\begin{aligned} A_j \cdot l_j(x_j) &= \int_a^b l_j(x)\,dx \\ A_j \cdot 1 &= \int_a^b l_j(x)\,dx \\ A_j &= \int_a^b l_j(x)\,dx \end{aligned}$$

Definíció 8.2.2 *Az interpolációs kvadratúra formulát Newton-Cotes-formulának nevezzük, ha az alappontok egyenlő lépésközönként vannak felvéve.*
Ezen belül zárt Newton-Cotes-formulának nevezik azt amikor $x_0 = a$ és $x_n = b$ azaz az intervallum szélei is kontrolpontok.

Speciális esetek:

- $n := 1$ Ekkor csak úgy kaphatunk zárt Newton-Cotes-formulát, ha $x_0 = a$ és $x_1 = b$ Ekkor visszakapjuk a trapéz formulát. Azért jó, hogy mostmár tudjuk, hogy a trapéz formula valójában egy interpolációs típusú kvadratúra formula, mert ekkor már

van formulánk a formula hibájára.

Most $n = 1$ és az alappontok $x_0 = a$ és $x_1 = b$ és $\int_a^b p(x)\,dx = t(f)$. Ha $f \in C^2[a,b]$ akkor

$$\int_a^b f(x)\,dx - t(f) = \int_a^b \frac{1}{2!}(x-a)(x-b)\cdot f''(\xi(x))\,dx$$

Mivel $\frac{1}{2!}(x-a)(x-b)$ Riemann integrálható és végig nem pozitív, és $f''(\xi(x))$ folytonos, ezért az integrál-középérték tétel értelmében $\exists\kappa \in [a,b]$:

$$\int_a^b f(x)\,dx - t(f) = f''(\kappa)\cdot\int_a^b \frac{1}{2}(x-a)(x-b)\,dx = \cdots = -\frac{h^3}{12}\cdot f''(\kappa)$$

Ebből látszik, hogy $t(f)$ másodrendű formula, mert elsőrendű f függvény második deriváltja 0 lesz minden pontba, így a hiba is 0.

- $n := 2$ Ekkor az alappontok a következők: $x_0 = a,\ x_1 = c = \frac{a+b}{2},\ x_2 = b$ Számítsuk ki a $\sum_{m=0}^{2} A_m \cdot f_m$ formula A_0, A_1, A_2 együtthatóit.

$$A_0 = \int_a^b l_0(x)\,dx = \int_a^b \frac{(x-c)(x-b)}{(a-c)(a-b)}\,dx = \cdots = \frac{h}{6}$$
$$A_1 = \int_a^b l_1(x)\,dx = \cdots = \frac{2h}{3}$$
$$A_2 = \int_a^b l_2(x)\,dx = \cdots = \frac{h}{6}$$

A tagokat összeadva:

$$\sum_{m=0}^{2} A_m \cdot f_m = A_0\cdot f(x_0) + A_1\cdot f(x_1) + A_2\cdot f(x_2) = \frac{h}{6}(f(a) + 4f(c) + f(b))$$

Így látszik, hogy ez valójában a Simpson-formula! (Innen következik a másik elnevezése a Simpson-formulának, parabola-formula)

Képlethibája: Ha $f \in C^4[a,b]$, akkor $\exists\kappa \in [a,b]$ melyre:

$$\int_a^b f(x) - s(f)\,dx = -\frac{h^5}{2880}\cdot f^{(4)}(\kappa)$$

Innen már látszik, hogy a Simpson formula 4-ed rendű, mert minden 3 vagy kevesebb rendű polinomnak a negyedik deriváltja mindenhol 0.

Probléma még mindig ezekkel a módszerekkel, hogy nem tudjuk tetszőleges pontossággal meghatározni a valós integrál értékét, akkor is ha bármelyik paraméterrel tartuk valahova.

8.3. Összetett kvadratúra formulák

Ötlet: lkalmazzuk szakaszonként az előbbi formulákat.

1. **Összetett trapéz formula**
   Osszuk fel m darab egyenlő részre az $[a,b]$ intervallumot. Jelölje egy ilyen kis szakasz hosszát $\Delta h := \frac{b-a}{m} = \frac{h}{m}$

$$\int_a^b f(x)\,dx \sum_{i=1}^{m} \int_{x_i-1}^{x_i} f(x)\,dx \approx \sum_{i=1}^{m} \frac{\Delta h}{2}(f(x_{i-1}) + f(x_i)) =: t_m(f)$$

   Mit lehet ennek az összetett formulának a hibájáról? Ha $f \in C^2[a,b]$ akkor

$$\int_a^b f(x)\,dx - t_m(f) = \sum_{i=1}^{m} -\frac{(\Delta h)^3}{12} \cdot f''(\kappa_i)$$

   ahol $\kappa_i \in [x_{i-1}, x_i]$ A fenti képletet kicsit átalakítva úgy, hogy $\Delta h = \frac{h}{m}$-et írunk

$$= \sum_{i=1}^{m} -\frac{h^3}{12m^2} \cdot \frac{1}{m} \cdot f''(\kappa_i) = -\frac{h^3}{12m^2} \overbrace{\left(\frac{1}{m}\sum_{i=1}^{m} f''(\kappa_i)\right)}^{S} = -\frac{h^3}{12m^2} f''(\eta)$$

   f'' m darab függvény értékének az átlaga és $S \in (\min f'', \max f'')$ és ezt felveszi egy $\eta \in [a,b]$ helyen.

Következmény 4 *1. Ha $f \in P_1$ akkor a formula pontos mert elsőfokú polinom második deriváltja mindig nulla*

2. *Feltétel szerint $f \in C^2[a,b]$ ebből következik, hogy $f'' \in C[a,b]$ és ebből meg következik, hogy felülről korlátos, azaz $\exists K > 0$ úgy, hogy $|f''| \le K$. Ebből következik, hogy*

$$\left|\int_a^b f\,dx - t_m(f)\right| = \frac{h^3}{13m^2} \cdot |f''(\eta)| \le \frac{h^3}{12m^2} \cdot K$$

   A fenti képletbe becsempészve $\Delta h := \frac{h}{m}$ kifejezést:

$$= \frac{h}{12} K (\Delta h)^2$$

   Legyen $\frac{h}{12}K := \tilde{K}$ és így a fenti képlet végső alakja:

$$\left|\int_a^b f\,dx - t_m(f)\right| \le \tilde{K}(\Delta h)^2$$

   Tehát ha m elég nagy (azaz Δh eléggé kicsi), akkor a hiba tetszőlegesen kicsi lesz.

Kérdés 25 *Ha $\Delta h \to 0$, milyen gyorsan tart ez a hibakorlát 0-hoz?*

Definíció 8.3.1 *Tegyük fel, hogy $g : K(0) \to \mathbb{R}$ olyan nulla körül értelmezett függvény, amelyhez $\exists \tilde{p} \in \mathbb{N}^+$ és $K \in \mathbb{R}$, hogy a 0-hoz kellően közeli t pontokban igaz a következő:*

$$|g(t)| \le K \cdot |t|^{\tilde{p}}$$

Ekkor jelölje p a legnagyobb ilyen tulajdonságú $\tilde{p}$ számot. Ekkor ez a bizonyos g függvény p-ad rendben tart 0-hoz a 0-ban. Jelölés: $g(t) = O(t^p)$ ("ordó t^p ")

Az előző eredményre visszatérve, a fenti ordó jelöléssel azt mondhatjuk, hogy

$$\left| \int_a^b f\,dx - t_m(f) \right| = O((\Delta h)^2)$$

Δh	$\tilde{K}(\Delta h)^2$
$\frac{\Delta h}{2}$	$\tilde{K}\left(\frac{\Delta h}{2}\right)^2$

2. **Összetett Simpson-formula**
   Az $[a,b]$ intevallumot felbontjuk m kisebb intervallumra és minden m darab kicsi intervallumban felveszünk egy segédpontot (például a kisintervallum felezőpontját) és ebben a kisintervallumban lévő 3 pontra illesztünk parabolát amit integrálunk. A végén összeadjuk a sok kicsi intevallumon illesztett parabola integrálját és ez lesz a közelítés az $\int_a^b f\,dx$ értékre.
   Ha $f \in C^4[a,b]$,a akkor $\exists \eta \in [a,b]$ úgy, hogy

$$\left| \int_a^b f\,dx - s_m(f) \right| = \left| -\frac{h^5}{2880m^4} f^{(4)}(\eta) \right| \leq \cdots \leq \tilde{\tilde{K}} \cdot (\Delta h)^4 = O((\Delta h)^4)$$

8.4. Gauss kvadratúrák

Eddig az interpolációs kvadratúra formuláknál adottnak vettük az alappontokat: $x_0, x_1, \ldots, x_n$ és láttuk, hogy a $\sum_{m=0}^n A_m \cdot f_m$ interpolációs kvadratúra formula pontos $\forall f \in P_n$ függvényre. Tehát a formula alapból legalább $n+1$ rendű, azaz annyi a rendje ahány alappont van.

Kérdés 26 *Növelhető-e a rend az alappontok megfelelő megválasztásával?*

Emlék: A középponti formula $k(f)$ is interpolációs formula, ahol egy alappont van (az intervallum felezőpontja c) és konstans függvényt illesztünk. Itt láttuk, hogy a középponti formula pontos elsőfokú polinomokra is, de a tételünk azt mondja, hogy csak elsőrendű a formula. Tehát ez azt jelenti, hogy ha konstans függvénnyt illesztünk, akkor ha a középpontot választjuk alappontnak, akkor magasabb rendű módszert kapunk.

A Simpson-formulánál is hasonló eredmény jött ki. Csak három alappont van tehát azt várnánk, hogy harmadrendű legyen, de ha a középpontot választjuk harmadik alappontnak, akkor negyedrendet érünk el.

Tegyük fel, hogy 2 alappont van x_0 és x_1. Kérdés: hogyan válasszuk meg ezeket, hogy minnél nagyobb rendű legyen a kvadratúra formulánk? És legyen az egyszerűség kedvéért $[a,b] := [-1,1]$ Ekkor az interpolációs formula a következőképpen néz ki:

$$\sum_{m=0}^{1} A_m \cdot f_m = A_0 \cdot f(x_0) + A_1 \cdot f(x_1)$$

ahol

$$\begin{aligned} A_0 &= \int_{-1}^{1} l_0(x)\,dx \\ &= \int_{-1}^{1} \frac{x - x_1}{x_0 - x_1}\,dx \\ &= \frac{1}{x_0 - x_1} \cdot \left[\frac{x^2}{2} - x_1 \cdot x\right]_{-1}^{1} \\ &= \frac{1}{x_0 - x_1} \cdot \left(\frac{1}{2} - x_1 - \frac{1}{2} - x_1\right) \\ &= \frac{2x_1}{x_1 - x_0} \end{aligned}$$

$$A_1 = \int_{-1}^{1} l_1(x)\,dx = -\frac{2x_0}{x_1 - x_0}$$

Tehát tetszőleges x_0 és x_1 esetén ez a formula pontos $\forall f \in P_1$ függvényre.

Kérdés 27 *Meg lehet megválasztani x_0 és x_1-et úgy, hogy magasabb fokú polinomokra is pontos legyen?*

Nyilvánvalóan: Egy kvadratúra formula pontos $\forall f \in P_q$ polinomra $\iff$ pontos az $f(x) = x^0, x^1, x^2, \ldots x^q$ függvényekre.

Mikor lesz pontos $\forall f \in P_2$ függvényre a fenti formula?

$$A_0 \cdot x_0^2 + A_1 \cdot x_1^2 = \int_{-1}^{1} x^2\,dx = \frac{2}{3}$$

Behelyettesítve az előbb kiszámolt A_0 és A_1 értékeket a következő egyenletet kapjuk:

$$\begin{aligned} \frac{2x_1}{x_1 - x_0} \cdot x_0^2 - \frac{2x_0}{x_1 - x_0} \cdot x_1^2 &= \frac{2}{3} \\ \frac{x_0 x_1}{x_1 - x_0} \cdot (x_0 - x_1) &= \frac{1}{3} \\ x_0 x_1 &= -\frac{1}{3} \end{aligned}$$

Mikor lesz pontos $\forall f \in P_3$ polinomra? Ha pontos az $f(x) = x^3$ függvényre is.

$$\begin{aligned} \frac{2x_1}{x_1 - x_0} \cdot x_0^3 - \frac{2x_0}{x_1 - x_0} \cdot x_1^3 &= \int_{-1}^{1} x^3\,dx = 0 \\ \frac{x_0 x_1}{x_1 - x_0}(x_0^2 - x_1^2) &= 0 \\ \frac{x_0 x_1}{x_1 - x_0}(x_0 + x_1)(x_0 - x_1) &= 0 \end{aligned}$$

A fenti kifejezés csak akkor 0 ha $x_0 + x_1 = 0$, azaz $x_0 = -x_1$

A két egyenlet megoldása:

$$-x_1^2 = -\frac{1}{3} \rightsquigarrow x_1 = \frac{1}{\sqrt{3}} \implies x_0 = -\frac{1}{\sqrt{3}}$$

$A_0, A_1 = ?$

$$A_0 = \frac{2x_1}{x_1 - x_0} = \frac{\left(2 \cdot \frac{1}{\sqrt{3}}\right)}{\frac{2}{\sqrt{3}}} = 1$$
$$A_1 = \cdots = 1$$

Tehát a formula: $f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) \approx \int_{-1}^{1} f\,dx$ és ez pontos $\forall f \in P_3$ polinomra.

Megjegyzés 14 *$n+1$ alappont esetén a rend legfeljebb $(n+1)$-el növelhető meg.*

Megjegyzés 15 *Ha $n = 2$ (azaz 3 alappont)*

$$x_0 = -\sqrt{\frac{3}{5}}, \quad x_1 = 0, \quad x_2 = \sqrt{\frac{3}{5}}$$
$$A_0 = \frac{5}{9}, \quad A_1 = \frac{8}{9}, \quad A_2 = \frac{5}{9}$$

Megjegyzés 16 *Ha $[a,b] \neq [-1,1]$ akkor az $[a,b]$ intervallumot átranszformáljuk a $[-1,1]$ intervallumra és ott felvesszük a Gauss-kvadratúra alappontokat, majd visszatranszformáljuk a $[-1,1]$ intervallumot az $[a,b]$ intervallumra.*

9. fejezet

Numerikus deriválás

Legyen $f : I \to \mathbb{R}$ és $x_0, x_1, \ldots, x_n \in I$ és $x_{i+1} = x_i + h$ ahol $i = 0, 1, \ldots, n-1$

Kérdés 28 *$f'(x_i) \approx ?$ az $f(x_0), f(x_1), \ldots, f(x_n)$ függvényértékek segítségével $f'(x_i) \approx ?$ az $f(x_0), f(x_1), \ldots, f(x_n)$ függvényértékek segítségével*

9.1. Az első derivált közelítése

Derivált definíciója:

$$f'(x_i) = \lim_{x \to x_i} \frac{f(x) - f(x_i)}{x - x_i}$$

Tehát

$$f'(x_i) \approx \frac{f(x_i + h) + f(x_i)}{h} =: \Delta f_+$$

úgynevezett jobboldali differencia hányados.

Ha az x_i ponttol van balra is értékünk:

$$f'(x_i) \approx \frac{f(x_i) - f(x_i - h)}{h} =: \Delta f_-$$

úgynevezett baloldali differencia hányados.

Ha már ezt a kettőt bevezettük, akkor vizsgáljuk a számtani közepüket:

$$\Delta f_c := \frac{\Delta f_+ + \Delta f_-}{2} = \frac{f(x_i + h) - f(x_i - h)}{2h}$$

úgynevezett centrális (vagy központi) differenciahányados.

Kérdés 29 *Mennyire jók ezek a közelítések?*

Definíció 9.1.1 *Jelölje f valamelyik deriváltját az x_i pontban Df és annak közelítését $\Delta f(h)$ Azt mondjuk, hogy a közelítés rendhe p, ha*

$$|Df - \Delta f(h)| = O(h^p)$$

Állítás 9.1.1 *A bal- és a jobboldali differenciahányados is elsőrendben közelíti egy f függvény első deriváltját x_i-ben, ha $f \in C^2(I)$.*

Bizonyítás: Csak a jobboldali differenciahányadosra. Fejtsük Taylor-sorba f-et x_i körül: $\exists \eta \in [x_i, x_i + h]$ úgy, hogy

$$f(x_i + h) = f(x_i) + h \cdot f'(x_i) + \frac{h^2}{2} f''(\eta)$$

Innen a következőképpen alakíthatjuk az egyenletet:

$$\frac{f(x_i + h) - f(x_i)}{h} = f'(x_i) + \frac{h}{2} f''(\eta)$$

$$\implies |f'(x_i) - \Delta f_+| = \frac{h}{2} |f''(\eta)| \leq K \cdot h$$

ahol $K = \frac{1}{2}|f''(\eta)|$

Állítás 9.1.2 *A Δf_c centrális differenciahányados másodrendben közelíti $f'(x_i)$-t, ha $f \in C^3(I)$.*

Bizonyítás: $\exists \eta_1 \in [x_i, x_i + h]$ és $\eta_2 \in [x_i - h, x_i]$ úgy, hogy:

$$f(x_i + h) = f(x_i) + h \cdot f'(x_i) + \frac{h^2}{2} f''(x_i) + \frac{h^3}{6} f'''(\eta_1)$$
$$f(x_i - h) = f(x_i) - h \cdot f'(x_i) + \frac{h^2}{2} f''(x_i) - \frac{h^3}{6} f'''(\eta_2)$$

$$\implies \Delta f_c = \frac{f(x_i + h) - f(x_i - h)}{2h} = f'(x_i) + \frac{h^3}{6} \cdot \frac{f'''(\eta_1) + f'''(\eta_2)}{2h}$$

Mivel $f''' \in C(i)$ ezért $\frac{f'''(\eta_1) + f'''(\eta_2)}{2} = f'''(\eta)$ ahol $\eta \in [x_i - h, x_i + h]$

Tehát

$$|f'(x_i - \Delta f_c)| = \frac{h^2}{6} \cdot |f'''(\eta)|$$

tehát a centrális differenciahányados valóban másodrendű közelítés.

9.2. A második derivált közelítése

Definícióból a következőt jelenti a második derivált:

$$f''(x_i) = \lim_{x \to x_i} \frac{f'(x) - f'(x_i)}{x - x_i}$$

Kis h esetén a második deriváltat a következőképpen tudjuk jól közelíteni:

$$f''(x_i) \approx \frac{f'(x_i+h) - f'(x_i)}{h}$$

Általában, amikor a második deriváltat kívánjuk közelíteni, az első deriváltat sem ismerjük. Így az első deriváltat is valahogyan közelítenünk kell. Szerencsére az előző fejezetben az első derivált közelítésével foglalkoztunk.

Közelítsük az első deriváltakat az előbb tanult módszerekkel.

Például mindkettőt a baloldali differenciahányadossal közelítjük

Így kapjuk a következő közelítést a második deriváltra:

$$f''(x_i) \approx \frac{\frac{f(x_i+h)-f(x_i)}{h} - \frac{f(x_i)-f(x_i-h)}{h}}{h} = \frac{f(x_i+h) - 2f(x_i) + f(x_i-h)}{h^2} =: \Delta^2 f_c$$

A fenti jelölésben a c alső index a centrális szóbol ered.

Kérdés 30 *Mi mondható a fenti közelítés rendjéről?*

Állítás 9.2.1 *Ez a bizonyos séma másodrendben közelíti $f''(x_i)$-t, ha eléggé sima függvény, azaz $f \in C^4(I)$*

(Nem bizonyítjuk)

9.3. A lépéstávolság dillemája

A következőkben arra a kérdésre próbálunk választ adni, hogy hogyan érdemes megválasztani a lépéstávolságokat, hogy minnél jobb közelítést érjünk el.

Hibaképletekből következik, hogy kisebb h esetén kisebb hibát várunk.

Az $f(x_i)$ értéket gyakorlatban általában hibával terheltek (pl.: számábrázolási hiba, mérési hiba, stb.)

Mi következik ebből?

Tekintsük a Δf_c központi sémát:

$$\Delta f_c := \frac{f(x_i+h) - f(x_i-h)}{2h}$$

Tegyük fel, hogy $f(x_i+h)$ és $f(x_i-h)$ értéket valamilyen hibával terhelve ismerjük csak, ezért $\tilde{f}_{i-1}$ és $\tilde{f}_{i+1}$ értékekkel számolunk helyettük.

$$\Delta \tilde{f}_c = \frac{\tilde{f}_{i+1} - \tilde{f}_{i-1}}{2h}$$

és itt $f(x_i + h) = \tilde{f}_{i+1} + \varepsilon_{i+1}$ és $f(x_i-) = \tilde{f}_{i-1} + \varepsilon_{i-1}$

Továbbá, tegyük fel, hogy $|\varepsilon_{i+1}|, |\varepsilon_{i-1}| \leq \varepsilon > 0$

Ekkor a következő összefüggést írhatjuk fel:

$$\Delta f_c = \frac{f(x_i + h) - f(x_i - h)}{2h} = \frac{\tilde{f}_{i+1} + \varepsilon_{i+1} - \tilde{f}_{i-1} - \varepsilon_{i-1}}{2h} = \frac{\overbrace{\tilde{f}_{i+1} - \tilde{f}_{i-1}}^{\Delta \tilde{f}_c}}{2h} + \frac{\varepsilon_{i+1} - \varepsilon_{i-1}}{2h}$$

$$f'(x_i) + \frac{h^2}{6} f''(\eta) = \Delta \tilde{f}_c + \frac{\varepsilon_{i+1} - \varepsilon_{i-1}}{2h}$$

$$\implies |f'(x_i) - \Delta \tilde{f}_c| \leq \overbrace{\frac{h^2}{6} M_3 + \frac{2\varepsilon}{2h}}^{g(h)}$$

ahol $M_3 := \sup|f''|$

Látható, hogy a hibára felső korlátot adó $g(h)$ függvényben a második tag nő, ha csökken h. Tehát, h nem választható sem túl negynak, sem túl kicsinek.

Kérdés 31 *Mi lesz az optimális h lépésköz?*

$g(h)$-nak ott lehet minimuma, ahol a deriváltja 0, azaz $g'(h) = 0$

$$g'(h) = \frac{1}{3} h M_3 - \frac{\varepsilon}{h^2} = 0$$

$$h^3 = \frac{3\varepsilon}{M_3} \implies h_{\text{opt}} = \left(\frac{3\varepsilon}{M_3} \right)^{1/3}$$

10. fejezet

Közönséges Differenciál Egyenletek (KDE-k) megoldása

Megoldandó:

$$y'(t) = f(t, y(t)), \quad t \in [t_0, T] \tag{10.1}$$

$$y(t_0) = y_0 \tag{10.2}$$

A fenti egy úgynevezett Cauchy-feladat.

Felmerül a probléma, hogy a pontos megoldást nem tudjuk mindig zárt alakban felírni

10.1. Véges külömbséges módszerek

A 10.1 és 10.2 egyenleteket diszkretizáljuk. Azaz, a $[t_0, T]$ intervallumon definiáljuk a következő rácshálót:

Legyen $N \in \mathbb{N}^+$ és jelöljük ω_τ-val a következő halmazt:

$$\omega_\tau := \left\{ t_n = t_0 + n \cdot \tau : \quad n = 0, 1, \ldots, N, \quad \tau = \frac{T - t_0}{N} \right\}$$

Egy folytonos függvény helyett egy olyan függvényt fogunk válaszul adni, mely csak ezeken a rácspontokon értelmezett. Gyakorlatilag egy vektort fogunk kapni, ahol az i-edik koordináta az i-edik rácsponton felvett függvényértéket jelképezi.

A megoldást csak az ω_τ pontjaiban közelítjük. Tehát a numerikus megyoldás egy rácspontfüggvény lesz.

Jelölje a t_n rácspontban a numerikus megoldást y_n. Szeretnénk, hogy ez köztel legyen a pontos megoldáshoz, azaz $y(t_n)$ értékhez.

Kérdés 32 *Hogyan konstruálhatunk egy, a megoldást közelítő rácspontfüggvényt?*

10.2. Explicit Euler-módszer (EE)

Az elejéről építjük fel a rácshálót.

t_0-ban $y(t_0) = y_0$ adva van.

$$y_1 = ?$$

Fejtsük sorba az y pontos megoldást a t_0 körül. Ha $y \in C^2[t_0, T]$ akkor a következőt állíthatjuk:

$$y(\overbrace{t_0 + \tau}^{t_1}) = y(t_0) + y'(t_0) \cdot \tau + O(\tau^2) = y_0 + f(t_0, y(t_0)) \cdot \tau + O(\tau^2) = y_0 + f(t_0, y_0) \cdot \tau + O(\tau^2)$$

Tehát, ha τ kicsi, akkor $O(\tau^2)$ tag elhanyagolható és a következő közelítést kapjuk:

$$y(t_1) \approx y_0 + f(t_0, y_0) \cdot \tau =: y_1$$

Ezután iteráljuk az eljárást: $y_1 \rightsquigarrow y_2 \rightsquigarrow \cdots \rightsquigarrow y_N$

$$y_{n+1} = y_n + f(t_n, y_n) \cdot \tau$$

Megjegyzés: Ez úgy is megkapható, hogy az (1)-es egyenletben az y'-t jobboldali differenciahányadossal közelítjük:

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n)$$

Ezt átrendezve valóban az előbb kapott képletet visszakapjuk.

Grafikusan az implicit Euler-módszer azt jelenti, hogy az y_n pontból az y_{n+1} pontba úgy jutunk el, hogy a t_n pontban az y pontos megoldásnak a meredekségével lépünk előre τ hosszút az y_n ponttól. Azaz a megoldás iránymező alapján lépkedünk mindig τ hosszúakat a mező irányába.

Ezt a módszert azért hívják *explicit* módszernek, mert eggyoldalra rendezve ki lehet fejezni y_{n+1} értéket.

10.3. Implicit Euler-módszer (IE)

Az EE-módszer képletében f-et ne (t_n, y_n)-ben, hanem (t_{n+1}, y_{n+1})-ben értékeljük ki:

$$y_{n+1} = y_n + f(t_{n+1}, y_{n+1}) \cdot \tau$$

A fenti képletben felmerül a probléma, hogy y_{n+1} kiszámításához kéne tudni y_{n+1} értéket, mert szerepel a jobboldalon f függvény hasában.

Tekinstünk a fenti képletre mint egy egyenletre és oldjuk meg y_{n+1}-re.

Jelölje $x := y_{n+1}$. Megoldandó: x-re a következő egyenlet:

$$x = y_n + f(t_{n+1}, x) \cdot \tau$$

Ezt az egyenletet midnen időlépésben meg kell oldanunk.

Az egyenletet meg tudjuk oldani előző fejezetekben tanult iterációs módszerekkel, például Newton-iterációval.

Rendezzünk 0-re az egyenletet:

$$\overbrace{x - y_n - f(t_{n+1}, x) \cdot \tau}^{F(x)} = 0$$

Válasszunk x_0 kezdőértéket

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

Érdemes meggondolni, hogy $F'(x)$ hogyan írható fel.

$$F'(x) = 1 - \partial_x f(t_{n+1}, x) \cdot \tau$$

10.4. Runge-Kutta módszerek

Egy fokkal bonyolultabb módszer: Azt csináljuk, hogy az $y(t_n)$ pontban megnézzük a meredekséget és csak $t_n + \frac{\tau}{2}$ ponttig megyünk el és megnézzük itt is a meredekséget és miután ezt kiszámoltuk, ezzel a meredeksséggel fogunk y_n-ből lépni.

Jelölje $k_1 = f(t_n, y_n)$ $y_n + k_1 \cdot \frac{\tau}{2}$ Így $k_2 = f\left(t_n + \frac{\tau}{2}, y_n + k_1 \frac{\tau}{2}\right)$ $y_{n+1} = y_n + k_2\tau$

Összegezve:

$$y_{n+1} = y_n + f\left(t_n + \frac{\tau}{2}, y_n + f(t_n, y_n) \cdot \frac{\tau}{2}\right) \cdot \tau$$

Ez egy kétlépcsős Runge-Kutta módszer. Lehet általánosabban s lépcsős Runga-Kutta módszert is definiálni.

$[t_n, t_{n+1}]$ intervallumon s pontban számítjuk ki a meredekséget és ezeket súlyozzuk és végül ezen súlyozott meredekségek összegével lépünk y_n-ből y_{n+1}-be.

A legelterjettebb Runge-Kutta módszer az RK4 módszer, amely 4 lépcsős.

A következő a képlete:

$$\begin{aligned}
k_1 &= f(t_n, y_n) \\
k_2 &= f\left(t_n + \frac{\tau}{2}, y_n + k_1 \frac{\tau}{2}\right) \\
k_3 &= f\left(t_n + \frac{\tau}{2}, y_n + k_2 \frac{\tau}{2}\right) \\
k_4 &= f(t_n + \tau, y_n + k_3\tau) \\
y_{n+1} &= y_n + \tau\left(\frac{1}{6}k_1 + \frac{1}{3}k_2 + \frac{1}{3}k_3 + \frac{1}{6}k_4\right)
\end{aligned}$$

Irodalomjegyzék

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