1. fejezet

A numerikus módszerek alapjai

Ebben a fejezetben megfogalmazzuk a Numerikus módszerek I. kurzus célját, definiáljuk a legfontosabb alapfogalmakat és összefoglaljuk a szükséges előismereteket.

1.1. A numerikus modellezés szerepe

A valóságban végbemenő fizikai, biológiai, társadalmi és egyéb folyamatok számos aspektusa leírható számszerű jellemzőkkel. Ezek tanulmányozásában hasznos eszköz a numerikus modellezés. Tekintsük át a numerikus modellezés legfőbb lépéseit!

1) **Valódi probléma.** Először megfogalmazzuk a megoldandó valódi problémát. Például tegyük fel, hogy kíváncsiak vagyunk egy tóban élő zsákmányhalak és ragadozó halak állományának az időbeli alakulására.

2) **Tudományos modell.** Következő lépésként felderítjük mindazon tényezőket, amelyek ismereteink szerint fontosak az adott jelenség lefolyásában. Példánkban azt kell számba vennünk, hogy mi minden befolyásolja a halpopulációk méretét. Úgy gondolhatjuk, hogy leginkább a következő tényezők:

1. a természetes szaporulat;

2. a zsákmányhalak ragadozók általi pusztítása;

3. a ragadozó halak természetes pusztulása...

3) **Matematikai modell.** Feltevéseinket matematikai formába öntjük. Pl. ha $x(t)$ jelöli a zsákmányhalak és $y(t)$ a ragadozó halak állományának a méretét (pl. össztömeg) a t

időpillanatban, akkor felállíthatjuk a következő differenciálegyenlet-rendszert:

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

valamely $[0, T]$ időintervallumon. (Ez a híres Lotka–Volterra-féle egyenletrendszer.) Az x_0 és y_0 a két halpopuláció kezdeti mérete, az egyenletekben szereplő a, b, c és d együtthatókat pedig pozitív konstansnak feltételezzük. Az a pl. a zsákmányhalak szaporodási rátája stb.

4) **Numerikus modell.** A matematikai modell pontos megoldása igen gyakran nem ismeretes, így közelítő megoldási módszereket kell alkalmaznunk. Pl. a fenti közönséges differenciálegyenlet-rendszernek sem tudjuk zárt alakban előállítani a megoldását, ezért valamilyen numerikus módszert kell alkalmaznunk, és ezzel eljutunk a numerikus modellhez.

5) **Számítógépes modell.** A numerikus modell által meghatározott számításokat tipikusan számítógépen végezzük el. A kiválasztott programozási nyelven megírt gépi kódot futtatjuk, és ennek megoldása szolgáltatja a valódi problémára adott választ.

A Numerikus módszerek I. előadáson a numerikus modellek felépítéséhez szükséges eljárásokkal ismerkedünk meg. A gyakorlaton a számítógépes realizáció kérdései is nagy hangsúlyt kapnak.

1.2. Hibaforrások

A fent vázolt folyamat során bőven van lehetőség hibák elkövetésére. Vegyük sorra a legfontosabb hibaforrásokat!

1. **Modellhiba.** A tudományos modell felállítása során nem tudunk minden fontos tényezőt figyelembe venni. A halas példában pl. nem vettük figyelembe a halászatot, az élőhely véges eltartóképességét stb. Ezért a felállított modell hibás eredményre vezethet. Mindig legyünk tudatában annak, hogy a modell sohasem a valóság!

2. **Képlethiba.** Amikor egy képletet kezelhetőségének érdekében úgy egyszerűsítünk, hogy bizonyos részeit elhagyjuk, vagy egyszerűbbel helyettesítjük, képlethiba kelet-

kezik. Pl. az exp(2) pontos értéke a következő sorösszeg:

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

A gép azonban nem tud sorösszeget (határérték!) számítani, ezért a fenti képlet helyett másikat használunk, pl. a végtelen sor összege helyett valamely részletösszegét számítjuk ki:

$$\sum_{k=0}^{\infty} \frac{2^k}{k!} \quad \text{helyett} \quad \sum_{k=0}^{n} \frac{2^k}{k!},$$

ahol n valamely nagy index. Ez azonban csak közelítése lesz az exp(2) számnak. A különbség oka a pontos képletnek egy másik képlettel való felcserélése. A képlethibához sorolható a diszkretizációs hiba is, amely abból ered, hogy pl. a deriváltat differenciahányadossal, az integrált részletösszeggel és általában a folytonos függvényeket ún. rácsfüggvényekkel közelítjük a numerikus modell felépítése során. A Lotka–Volterra-féle rendszerre is numerikus módszert kell alkalmaznunk, amelynél elkerülhetetlenül fellép ilyen típusú hiba.

3. **A bemenő adatok hibája.** A matematikai modellnek meg kell adnunk bizonyos bemenő adatokat, pl. fent a két halpopuláció kezdeti méretét. A bemenő adatok azonban sokszor hibával terheltek. A példában a kezdeti halpopulációkat csak becsülni tudjuk jól-rosszul. Gyakran mért adatokat használunk, amelyek pontossága szintén korlátozott.

4. **A számábrázolásból eredő hiba.** Ez a hiba abból adódik, hogy a számítógépen a valós számoknak csak egy véges részhalmazát tudjuk ábrázolni. Ennek tárgyalásával az előadáson nem foglalkozunk.

1.3. Hibafogalmak

Pontosítsuk, hogy mit értünk a továbbiakban hibán. Látni fogjuk, hogy többféle hibafogalom is használatos. Jelöljük a-val valaminek a pontos értékét, amit szeretnénk kiszámítani. Legyen ez most az egyszerűség kedvéért egy valós szám, például a zsákmányhalak vagy a ragadozó halak össztömege a T időpillanatban, de gondolhatunk akár exp(2) pontos értékére vagy egy algebrai egyenlet pontos megoldására is. Felállítottuk számítógépes modellünket, és jelölje számításunk eredményét $\tilde{a}$. Az $\tilde{a}$ valós szám többnyire nem egyezik meg pontosan az a-val, és a hibáját a legegyszerűbben a két érték különbségével jellemezhetjük.

1.3.1 Definíció. *A*

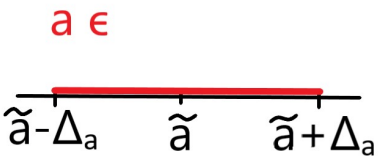
$$\Delta a := a - \tilde{a}$$

számot az $\tilde{a}$ közelítő érték **abszolút hibá***jának nevezzük.*

Ha ez abszolút értékben elég kicsi, akkor az $\tilde{a} \approx a$ jelölést alkalmazzuk. (Kiolvasva: $\tilde{a}$ közelíti a-t.) Természetesen az a pontos értékét általában nem ismerjük, így az abszolút hibát sem. Elvárható azonban tudnunk, hogy legfeljebb mekkora ennek a hibának a nagysága.

1.3.2 Definíció. *A $\Delta_a \geq 0$ számot az $\tilde{a}$ közelítő érték* **abszolút hibakorlát***jának nevezzük, ha*

$$|a - \tilde{a}| = |\Delta a| \leq \Delta_a.$$



Erre az $a = \tilde{a} \pm \Delta_a$ jelölést is alkalmazzuk.
Nyilvánvalóan egy Δ_a-nál nagyobb szám is abszolút hibakorlát. Az abszolút hiba önmagában nem elegendő a közelítés jóságának a jellemzéséhez, fontos lenne valamihez viszonyítani.

1.3.3 Definíció. *A $\delta a := \Delta a/|\tilde{a}|$ számot az $\tilde{a}$ közelítő érték* **relatív hibá***jának nevezzük.*

Ennek természetesen nincs értelme, ha az $\tilde{a}$ közelítő érték nulla, de használata akkor is problémás lehet, ha $\tilde{a}$ csak közel van a nullához, ekkor kis abszolút hiba esetén is nagy relatív hibát kaphatunk. Célszerű tehát az abszolút és a relatív hibát együtt vizsgálni.

1.3.4 Definíció. *A δ_a számot az $\tilde{a}$ közelítő érték egy* **relatív hibakorlát***jának nevezzük, ha $|\delta a| \leq \delta_a$.*

Nyilvánvalóan, ha Δ_a abszolút hibakorlát, akkor $\frac{\Delta_a}{|\tilde{a}|}$ egy relatív hibakorlát lesz. A relatív hibát ill. hibakorlátot az a pontos értékhez viszonyítva is szokásos értelmezni. A gyakorlatban azonban a többnyire nem ismert, ezért ezt nem mindig tudjuk kiszámítani. (Ha $\tilde{a}$ elég közel van a-hoz, akkor persze a kétféle módon definiált relatív hiba is közel egyenlő.)

1.3.1. Az alapműveletek hibája

Érdemes megvizsgálni a valós számokkal végzett alapműveletek során keletkező hibát. Tegyük fel, hogy az x és y valós számok helyett a hibával terhelt $\tilde{x}$ és $\tilde{y}$ értékekkel számolunk. Hogyan hat ki ez a művelet eredményére?

1. Összeadás:

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

   Ez azt jelenti, hogy az összeadandók abszolút hibakorlátjait összeadva becsülhetjük az összeg abszolút hibáját. (Ezt röviden úgy szokás mondani, hogy összeadásnál a hibák összeadódnak.)

2. Kivonás:

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

   A hibák tehát itt is összeadódnak, nem pedig kivonódnak. A különbséggel osztva látható, hogy közeli számok kivonásával nagy lehet az elkövetett relatív hiba.

3. Szorzás:

$$|x\cdot y-\tilde{x}\cdot\tilde{y}| = |x\cdot y - x\cdot\tilde{y}+x\cdot\tilde{y}-\tilde{x}\cdot\tilde{y}| \leq |x|\cdot|y-\tilde{y}|+|\tilde{y}|\cdot|x-\tilde{x}|$$

$$\approx |\tilde{x}|\cdot|y-\tilde{y}|+|\tilde{y}|\cdot|x-\tilde{x}| \leq |\tilde{x}|\cdot\Delta_y+|\tilde{y}|\cdot\Delta_x =: \Delta_{xy}$$

4. Osztás (önállóan):

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

   Az utóbbi jelzi, hogy kis számmal való osztásnál nagy lehet az elkövetett hiba.

1.3.2. A feladat korrekt kitűzése

Leszögezzük, hogy csak olyan feladat megoldásával érdemes foglalkozni, amely eleget tesz három alapkövetelménynek:

a.) létezik megoldása (egzisztencia);

b.) csak egy megoldása létezik (unicitás);

c.) a bemenő adatokban elkövetett kis hiba csak kicsit változtat a megoldás értékén.[1]

Az első kettő magától értetődik, de nézzünk egy példát a c.) fontosságának illusztrálására.

1.3.5 Példa. *Tekintsük az $Ax = b$ lineáris algebrai egyenletrendszert, ahol*

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1,01 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 2,01 \end{bmatrix}$$

Ezt meg tudjuk oldani pontosan, pl. a Cramer-szabállyal, a megoldás:

$$x_1 = \frac{\begin{vmatrix} 2 & 1 \\ 2,01 & 1,01 \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ 1 & 1,01 \end{vmatrix}} = \frac{0,01}{0,01} = 1; \quad x_2 = \frac{\begin{vmatrix} 1 & 2 \\ 1 & 2,01 \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ 1 & 1,01 \end{vmatrix}} = \frac{0,01}{0,01} = 1.$$

Most perturbáljuk a jobb oldalt úgy, hogy a b vektor második elemét egy századdal megnöveljük:

$$\tilde{b} = \begin{bmatrix} 2 \\ 2,02 \end{bmatrix}.$$

Az $A\tilde{x} = \tilde{b}$ perturbált feladat $\tilde{x}$ pontos megoldása már

$$\tilde{x}_1 = \frac{\begin{vmatrix} 2 & 1 \\ 2,02 & 1,01 \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ 1 & 1,01 \end{vmatrix}} = 0; \quad \tilde{x}_2 = \frac{\begin{vmatrix} 1 & 2 \\ 1 & 2,02 \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ 1 & 1,01 \end{vmatrix}} = \frac{0,02}{0,01} = 2.$$

Szembetűnő a nagy eltérés, és mivel a számolás pontos volt, ez az eltérés magának a feladatnak a rossz tulajdonságából ered. (Később választ kapunk arra, hogy a lineáris egyenletrendszer (mátrixának) milyen tulajdonsága mutatja meg azt, ha a megoldás ilyen rendkívüli módon érzékeny a szabad tagokra.) Könnyen látható, hogy ekkor nincs sok remény a feladat megfelelő pontosságú megoldására, mert – számítógéppel számolva – a jobb oldali vektor elemeit legalábbis számábrázolási (de akár még mérési vagy egyéb hiba is) terheli, ezért az eredeti feladat helyett valójában egy perturbált feladatot oldunk meg.

[1] Ezt kissé pontatlanul úgy is szokás mondani, hogy a megoldás folytonosan függ a bemenő adatoktól, ami valójában a következőt jelenti. Jelöljön b egy bemenő adatot, és x a hozzá tartozó megoldást, ekkor minden $\varepsilon > 0$-hoz létezik olyan $\delta > 0$, hogy ha a $\tilde{b}$ bemenő adat esetén a megoldás $\tilde{x}$, és $|b - \tilde{b}| < \delta$, akkor $|x - \tilde{x}| < \varepsilon$. Ez teljesül például ha van olyan $K > 0$ konstans, amelyre $|x - \tilde{x}| \leq K|b - \tilde{b}|$. Ugyanakkor, ha itt K nagy, akkor hiába van folytonos függés, a bemenő adatot kicsit megváltoztatva a megoldás nagyot is változhat. Ezért a korrekt kitűzés c) kritériumát helyesebb úgy megfogalmazni, hogy a bemenő adatban elkövetett kis hiba csak kicsit változtatja meg a megoldást. Tehát a korrekt kitűzéshez nem elég, hogy $|x - \tilde{x}| \leq K|b - \tilde{b}|$, hanem itt K-nak elég kicsinek is kell lennie.

1.4. Általánosítás, előismeretek

Ebben a fejezetben általánosabb keretet adunk a hibaanalízishez, és összegyűjtjük a jegyzet további részének megértéséhez szükséges előismereteket.

1.4.1. Normált tér

Eddig feltettük, hogy $a, \tilde{a} \in \mathbb{R}$. A gyakorlatban azonban nem mindig ez a helyzet, pl. ha a egy n-ismeretlenes lineáris algebrai egyenletrendszer megoldása, akkor ez egy $\mathbb{R}^n$-beli elem. Kérdés: mit értsünk ekkor a számított $\tilde{a} \in \mathbb{R}^n$ vektor abszolút/relatív hibáján? Hogyan értsük továbbá ezeket a hibákat valamely egyéb X halmazon?

$\mathbb{R}$-ben $a-\tilde{a}$ értelmes, és az abszolút hibakorlát az $|a-\tilde{a}|$ kifejezésre adott felső korlátot jelentette. Ennek általánosításához tehát az kellene, hogy értelmezve legyen
i) az elemek különbsége. Ezért tegyük fel, hogy az X halmazon van összeadás és skalárral való szorzás, azaz X vektortér;
ii) az abszolút érték függvényhez hasonló tulajdonságú függvény, amely a vektortér elemeinek a „nagyságát” jellemzi.

A ii) tulajdonság biztosításához vegyük szemügyre, hogy milyen tulajdonságai vannak az abszolút érték függvénynek! Ez a függvény $\mathbb{R} \to \mathbb{R}$ képez, és a következő három tulajdonsága révén alkalmas a valós számok nagyságának a kifejezésére:

1. $|x| \geq 0 \quad \forall x \in \mathbb{R}$, és $|x| = 0 \Leftrightarrow x = 0$;
2. $|\lambda x| = |\lambda| \cdot |x| \quad \forall x \in \mathbb{R}, \forall \lambda \in \mathbb{R}$;
3. $|x + y| \leq |x| + |y| \quad \forall x, y \in \mathbb{R}$.

A valós számok nagyságának a fogalmát így kézenfekvő a következőképpen kiterjeszteni tetszőleges vektortérre.

1.4.1 Definíció. *Legyen X egy vektortér. Ekkor egy $\|\cdot\| : X \to \mathbb{R}$ függvényt normának nevezünk, ha igaz rá a következő három tulajdonság:*

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

Azt a számot, amelyet ez a függvény egy $x \in X$ vektorhoz hozzárendel, az x vektor normájának nevezzük, és $\|x\|$-vel jelöljük. Az $(X, \|\cdot\|)$ rendezett párt normált térnek nevezzük.

Példák normált térre

1.) $X = \mathbb{R}$, $\|x\| = |x|$.
2.) $X = \mathbb{R}^n$-en három alapvető norma használatos:

- $\|x\|_1 := |x_1| + |x_2| + \ldots + |x_n|$ (egyes norma);
- $\|x\|_2 := \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$ (kettes vagy euklideszi norma);
- $\|x\|_\infty := \max\{|x_1|, |x_2|, \ldots, |x_n|\}$ (maximumnorma)

Pl. Határozzuk meg az $x = (1, 2, -3) \in \mathbb{R}^3$ vektor mindhárom fenti normáját!

- $\|x\|_1 = 1 + 2 + 3 = 6$;
- $\|x\|_2 = \sqrt{1 + 2^2 + 3^2} = \sqrt{14}$;
- $\|x\|_\infty = \max\{1, 2, 3\} = 3$.

Az első kettő az alábbi, ún. p-norma speciális esete $p = 1$ ill. $p = 2$ mellett:

$$\|x\|_p := (|x_1|^p + |x_2|^p + \ldots + |x_n|^p)^{\frac{1}{p}},$$

ahol $p \geq 1$. Továbbá megmutatható, hogy $\lim_{p\to\infty} \|x\|_p = \|x\|_\infty$ minden $x \in \mathbb{R}^n$ esetén.

Az egyes vektornormáról lássuk is be, hogy valóban norma $\mathbb{R}^n$-en!

1. Nyilvánvalóan $\sum_{k=1}^n |x_k| \geq 0$, és nemnegatív számok összege csak akkor lehet nulla, ha minden tag nulla, azaz ha $|x_k| = 0\ \forall k = 1, 2, \ldots, n$. Ez pedig pontosan akkor áll fenn, ha $x_k = 0\ \forall k = 1, 2, \ldots, n$.
2. $\|\lambda x\|_1 = \sum_{k=1}^n |\lambda x_k| = \sum_{k=1}^n (|\lambda| \cdot |x_k|) = |\lambda| \cdot \sum_{k=1}^n |x_k| = |\lambda| \cdot \|x\|_1$.
3. $\|x+y\|_1 = \sum_{k=1}^n |x_k+y_k| \leq \sum_{k=1}^n (|x_k|+|y_k|) = \sum_{k=1}^n |x_k| + \sum_{k=1}^n |y_k| = \|x\|_1 + \|y\|_1$.

3.) $X = C[a, b]$-n a két legalapvetőbb norma:

- $\|f\|_\infty = \max_{x\in[a,b]} |f(x)|$ (maximumnorma);
- $\|f\|_\int := \int_a^b |f|$ (integrálnorma).

Egy X normált térben az $x \in X$ és $y \in X$ elem távolságán az $\|x - y\|$ számot értjük, és azt mondjuk, hogy x közel van y-hoz, ha $\|x-y\|$ kicsi. Ez alapján az abszolút és relatív hibát / hibakorlátot a következőképpen értelmezhetjük. Legyen $a \in X$ a pontos érték, $\tilde{a} \in X$ pedig a számításunk eredménye. Itt is a $\Delta a := a - \tilde{a}$ különbséget (X-beli vektor)

hívjuk az $\tilde{a}$ közelítés abszolút hibájának. A $\Delta_a \geq 0$ számot az $\tilde{a}$ közelítő érték abszolút hibakorlátjának nevezzük, ha $\|a - \tilde{a}\| = \|\Delta a\| \leq \Delta_a$. A relatív hiba ill. hibakorlát pedig ezek $\|\tilde{a}\|$-val valós osztásával értelmezhető.

Minden normált térben értelmes egy sorozat konvergenciájának a fogalma.

1.4.2 Definíció. *Az $(x_n) \subset X$ sorozatot konvergensnek nevezzük, ha létezik $x \in X$, amely mellett $\|x - x_n\| \to 0$. Ezt az x elemet az (x_n) sorozat határértékének nevezzük.*

Láttuk: ugyanazon vektortéren több norma is megadható, és ezekben ugyanazon vektor normája különbözhet. Ugyanakkor bizonyos szempontból mindegy, melyiket használjuk, amennyiben a normák ekvivalensek.

1.4.3 Definíció. *Legyen X vektortér, $\|\cdot\|_a$ és $\|\cdot\|_b$ pedig két norma X-en. Azt mondjuk, hogy $\|\cdot\|_a$ ekvivalens $\|\cdot\|_b$-vel, ha léteznek olyan $0 < m \leq M$ konstansok, amelyek mellett*

$$m\|x\|_a \leq \|x\|_b \leq M\|x\|_a \quad \forall x \in X.$$

Jelölése: $\|\cdot\|_a \sim \|\cdot\|_b$

Ekkor pl. ha $x_n \to x$ $\|\cdot\|_a$-ban, akkor $\|\cdot\|_b$-ben is, és fordítva. Megjegyezzük, hogy véges dimenziós vektortérben (pl. $\mathbb{R}^n$) bármely két norma mindig ekvivalens.

1.4.2. Mátrixnormák

Ismeretes, hogy $\mathbb{R}^{n\times n}$ (az n-szer n-es mátrixok halmaza) a mátrixok összeadásával és skalárral való szorzásával szintén vektorteret alkot. Kérdés: hogyan definiáljunk rajta normát?

Jelölje $\|\cdot\|_{\mathbb{R}^n}$ bármelyik $\mathbb{R}^n$-beli vektornormát, és 0_n az $\mathbb{R}^n$ nullvektorát. Belátható, hogy

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

norma $\mathbb{R}^{n\times n}$-en. Ezt a normát az $\|\cdot\|_{\mathbb{R}^n}$ vektornorma által indukált mátrixnormának nevezzük.

Mit fejez ez ki szemléletesen? Az $\frac{\|Ax\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}}$ hányados azt mutatja meg, hogy az A mátrix mint lineáris leképezés hányszorosára nyújtja meg az $x \in \mathbb{R}^n$ vektort. Ha az összes x-re képezzük ezt a hányadost, majd ezen hányadosok szupremumát, akkor ez azt mutatja meg, hogy A legfeljebb hányszorosára nyújt meg egy vektort.

Belátható, hogy a fenti szupremum ekvivalens a következőkkel:

$$\sup_{x\in\mathbb{R}^n, x\neq 0_n} \frac{\|Ax\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}} = \sup_{x\in\mathbb{R}^n, \|x\|_{\mathbb{R}^n}=1} \frac{\|Ax\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}} = \sup_{x\in\mathbb{R}^n, \|x\|_{\mathbb{R}^n}=1} \|Ax\|_{\mathbb{R}^n} =$$

$$= \max_{x \in \mathbb{R}^n, \|x\|_{\mathbb{R}^n}=1} \|Ax\|_{\mathbb{R}^n}$$

Megjegyezzük, hogy más-más $\mathbb{R}^n$-beli vektornorma más mátrixnormát indukál, tehát egy mátrixnak nem feltétlenül egyezik meg pl. a maximumnorma és az euklideszi norma által indukált normája. A szupremumot nem mindig könnyű kiszámítani, ezért érdemes tudnunk, hogy az indukált mátrixnormákat a mátrix elemeiből az alábbi módokon számíthatjuk ki:

1. Első norma (másképpen oszlopnorma): a leképezés mátrixának minden oszlopában összeadjuk az elemek abszolút értékét, és a legnagyobb ilyen oszlopösszeget választjuk.

$$\|A\|_1 := \max_{j \in \{1,2,\ldots,n\}} \sum_{i=1}^{n} |a_{ij}|$$

2. Második norma (spektrálnorma):

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

ahol A^T az A mátrix transzponáltja, $\lambda_{\max}(A^T A)$ pedig az $A^T A$ mátrix legnagyobb abszolút értékű sajátértéke.

3. Maximumnorma (sornorma): a legnagyobb sorösszeg.

$$\|A\|_{\max} := \max_{i \in \{1,2,\ldots,n\}} \sum_{j=1}^{n} |a_{ij}|$$

A kettes mátrixnormához több fontos megjegyzést fűzünk. Vajon a $\sqrt{\lambda_{\max}(A^T A)}$ kifejezés mindig értelmes-e? Igen, ugyanis
i) $A^T A$ mindig szimmetrikus mátrix, így a sajátértékei valósak.
ii) $A^T A$ pozitív szemidefinit is ($\langle A^T Ax, x\rangle \geq 0 \ \forall x \in \mathbb{R}^n$), ugyanis

$$\langle A^T Ax, x\rangle = \langle Ax, Ax\rangle = \|Ax\|_2^2 \geq 0,$$

ezért a sajátértékei nemnegatívak.

1.4.4 Definíció. *Legyen $A \in \mathbb{R}^{n\times n}$ sajátértékei $\lambda_1, \lambda_2, \ldots, \lambda_n$. Ekkor a $\rho(A) := \max_i |\lambda_i|$ számot az A mátrix spektrálsugarának nevezzük.*

1.4.5 Állítás. *Ha A szimmetrikus, akkor $\|A\|_2 = \rho(A)$.*

Biz.: Tegyük fel, hogy A szimmetrikus. Ekkor $A = A^T$, így $A^TA = A^2$. Az A^2 mátrix sajátértékei az A sajátértékeinek a négyzetei, így

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)} = \sqrt{\lambda_{\max}(A^2)} = \sqrt{(\lambda_{\max}(A))^2} = |\lambda_{\max}(A)| = \rho(A).$$

Továbbá, tetszőleges $A \in \mathbb{R}^{n\times n}$ mátrix minden indukált mátrixnormájára fennáll a

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

egyenlőtlenség, azaz a spektrálsugárnál kisebb indukált norma nem létezik. Ez könnyen meggondolható, ugyanis legyen v egy $\lambda_{\max}$-hoz tartozó sajátvektor. Ekkor $Av = \lambda_{\max}v$, ezért $\|Av\| = |\lambda_{\max}| \cdot \|v\|$. Így a v vektorra

$$\frac{\|Av\|}{\|v\|} = |\lambda_{\max}| \Rightarrow \sup_{v\in\mathbb{R}^n, v\neq 0} \frac{\|Av\|}{\|v\|} \geq |\lambda_{\max}|,$$

ami a belátandó állítást jelenti.

Az indukált mátrixnormák rendelkeznek a következő három fontos tulajdonsággal:

i) Az Ax mátrix-vektor szorzatra mindig igaz, hogy $\|Ax\| \leq \|A\| \cdot \|x\|$ (itt a bal oldali és a jobb oldali második norma valamelyik vektornorma, és az A mátrix normája az ezen vektornorma által indukált mátrixnorma!) Ez azonnal következik az indukált mátrixnorma definíciójából.

ii) Az identitásmátrix normája 1, azaz $\|I\| = 1$.
Biz.:
$$\|I\| = \sup_{x\in\mathbb{R}^n, x\neq 0_n} \frac{\|Ix\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}} = \sup_{x\in\mathbb{R}^n, x\neq 0_n} \frac{\|x\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}} = \sup\{1\} = 1.$$

iii) Két mátrix szorzatának normája nem nagyobb, mint a tényezők normájának a szorzata (szubmultiplikativitás): $\|AB\| \leq \|A\| \cdot \|B\|$.
Biz.:
$$\|A \cdot B\| = \sup_{x\in\mathbb{R}^n, x\neq 0_n} \frac{\|(AB)x\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}} \leq \sup_{x\in\mathbb{R}^n, x\neq 0_n} \frac{\|A\| \cdot \|B\| \cdot \|x\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^n}} = \|A\| \cdot \|B\|.$$

A mátrixok vektorterében további, nem indukált normák is használatosak. Ezek a mátrixnormák nem írhatók fel egy adott vektornormához tartozó szupremumnormaként. Nem indukált mátrixnormát kapunk pl., ha kiválasztjuk a mátrix lehető legnagyobb abszolút értékű elemét:

$$\|A\|_* := \max_{i,j} |a_{ij}|;$$

vagy ha összeadjuk az összes mátrixelemet abszolút értékben:

$$\|A\|_{**} := \sum_{i,j=1}^{n} |a_{ij}|.$$

Ismeretes még az ún. Frobenius-norma:

$$\|A\|_F := \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2}.$$

Ezek tehát mind rendelkeznek a norma három tulajdonságával, de fontos tudni, hogy nem mindig rendelkeznek az i)-iii) tulajdonságokkal. Ez az oka annak, hogy a nem indukált mátrixnormákat ritkán használjuk.

1.4.3. Kondíciószám

Most visszatérünk arra a kérdésre, hogy a korábban látott lineáris egyenletrendszer megoldása miért lehetett annyira érzékeny a szabad tagok vektorának kis megváltoztatására. A mátrixnorma segítségével bevezethetünk egy mérőszámot, amely tájékoztat arról, hogy egy adott lineáris algebrai egyenletrendszer megoldása során kell-e számítanunk erre a problémára.

Legyen $A \in \mathbb{R}^{n\times n}$ egy invertálható mátrix, és $b \in \mathbb{R}^n \setminus \{0\}$. Tegyük fel, hogy a b jobb oldal Δb hibával terhelten ismert. Jelölje Δx a megoldás abszolút hibáját, ezzel

$$A(x - \Delta x) = b - \Delta b$$

a perturbált egyenletrendszer. Ebből $Ax = b$ miatt

$$A \cdot \Delta x = \Delta b$$

következik, azaz

$$\Delta x = A^{-1}\Delta b.$$

Ezt valamely indukált mátrixnormában véve és az i) tulajdonságot alkalmazva

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

adódik. Másrészt

$$\|b\| = \|Ax\| \leq \|A\| \cdot \|x\| \Rightarrow \frac{1}{\|x\|} \leq \|A\| \cdot \frac{1}{\|b\|}.$$

Ebből a

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

becsléshez jutunk. Látható: a megoldás relatív hibája felülről becsülhető a jobb oldal relatív hibájának az $\|A^{-1}\| \cdot \|A\|$ számszorosával. Ezért igen hasznos ez a szorzat, és külön nevet is kap:

1.4.6 Definíció. *Az $\|A^{-1}\| \cdot \|A\|$ számot az $A \in \mathbb{R}^{n\times n}$ invertálható mátrix kondíciószámának nevezzük. Jelölése:* cond$A$.

Nyilvánvalóan, cond$A$ függ a választott normától. A lineáris egyenletrendszert rosszul kondicionáltnak nevezzük, ha cond$A$ jóval nagyobb 1-nél.

Térjünk vissza a 1.3.5.példára. Ha az 1-es vektornorma által indukált mátrixnormában dolgozunk, akkor cond$A = 404,01 >> 1$, így a lineáris egyenletrendszer valóban rosszul kondicionált.

1.4.7 Megjegyzés. *Levezethető, hogy ha nem a jobb oldali vektor, hanem az A mátrix hibás, akkor is* cond$A$ *a releváns indikátor a megoldás relatív hibájára nézve.*

1.4.4. Fixponttétel $\mathbb{R}^n$-ben

A későbbiekben szükségünk lesz egy fontos tételre $\mathbb{R}^n$-ben. Ehhez először bevezetjük a kontrakció fogalmát.

1.4.8 Definíció. *Legyen $f : \mathbb{R}^n \to \mathbb{R}^n$. Azt mondjuk, hogy f kontrakció valamely $\|\cdot\|$ $\mathbb{R}^n$-beli normában, ha $\exists q \in [0, 1)$ állandó, amely mellett*

$$\|f(x) - f(y)\| \leq q\|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

1.4.9 Állítás. *Ha $f : \mathbb{R}^n \to \mathbb{R}^n$ lineáris leképezés, azaz $f(x) = Ax$, ahol $A \in \mathbb{R}^{n\times n}$ egy mátrix, és $\|A\| < 1$ valamely indukált mátrixnormában, akkor f kontrakció.*

Biz.: Abban a vektornormában, amelyhez az indukált mátrixnorma tartozik, az i) tulajdonság miatt

$$\|f(x) - f(y)\| = \|Ax - Ay\| = \|A(x - y)\| \leq \|A\| \cdot \|x - y\|.$$

1.4.10 Következmény. *Ha f kontrakció, akkor folytonos.*

Biz.: Tegyük fel, hogy $x_m \to x$, ekkor $\|f(x_m) - f(x)\| \leq q\|x_m - x\| \to 0$.

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

1.4.12 Tétel. *(Fixponttétel $\mathbb{R}^n$-ben)*
Tegyük fel, hogy $f : \mathbb{R}^n \to \mathbb{R}^n$ az egész $\mathbb{R}^n$-en értelmezett kontrakció a $q \in [0, 1)$ állandóval. Ekkor

1. *f-nek létezik fixpontja;*
2. *ez a fixpont egyetlen (jelölje $x^* \in \mathbb{R}^n$);*
3. *az $x^{k+1} = f(x^k),\ k = 0, 1, \ldots$ módon előállított (x^k) $\mathbb{R}^n$-beli sorozat tetszőleges x^0 esetén konvergens;*
4. $\lim x^k = x^*$;
5. $\|x^k - x^*\| \leq \frac{q^k}{1-q}\|x^1 - x^0\|$.

Biz.: Mutassuk meg, hogy $\|x^k - x^{k-1}\| \leq q^{k-1}\|x^1 - x^0\|$. Teljes indukcióval:

- $k = 1$ esetén $\|x^1 - x^0\| \leq \|x^1 - x^0\|$.
- Tegyük fel hogy $k - 1$-re igaz: $\|x^{k-1} - x^{k-2}\| \leq q^{k-2}\|x^1 - x^0\|$. Ekkor

$$\|x^k - x^{k-1}\| = \|f(x^{k-1}) - f(x^{k-2})\| \leq q\|x^{k-1} - x^{k-2}\| \leq q \cdot q^{k-2}\|x^1 - x^0\| = q^{k-1}\|x^1 - x^0\|.$$

Legyen $m > k$. Ekkor

$$\|x^m - x^k\| = \|x^m - x^{m-1} + x^{m-1} - x^{m-2} + \ldots + x^{k+1} - x^k\|$$

$$\leq \|x^m - x^{m-1}\| + \|x^{m-1} - x^{m-2}\| + \ldots + \|x^{k+1} - x^k\| \leq (q^{m-1} + q^{m-2} + \ldots q^k)\|x^1 - x^0\|$$

$$q^k\|x^1 - x^0\|(1 + q + \ldots + q^{m-1-k}) \leq q^k\|x^1 - x^0\| \sum_{m=0}^{\infty} q^m = q^k\|x^1 - x^0\|\frac{1}{1-q}. \qquad (1.1)$$

Így $k \to \infty$ esetén $\|x^m - x^k\| \to 0$ $(m > k)$. Ezért $\forall \varepsilon > 0\ \exists N \in \mathbb{N} : \|x^m - x^k\| < \varepsilon\ \forall m > k > N$. Így (x^k) Cauchy-sorozat $\mathbb{R}^n$-ben, azaz konvergens is, tehát létezik $\lim_{k\to\infty} x^k =: x^*$.

Mivel f folytonos is, ezért az $x^k = f(x^{k-1})$ összefüggésből

$$x^* = \lim_{k\to\infty} x^k = \lim_{k\to\infty} f(x^{k-1}) = f(\lim_{k\to\infty} x^{k-1}) = f(x^*),$$

azaz x^* fixpont. Az (1.1) becslésből $m \to \infty$ esetén

$$\|x^* - x^k\| \leq \frac{q^k}{1-q}\|x^1 - x^0\|$$

Az unicitás belátásához indirekt tegyük fel, hogy $\exists x^*, y^* \in \mathbb{R}^n,\ x^* \neq y^*$, amelyek fixpontok, azaz $x^* = f(x^*)$ és $y^* = f(y^*)$. Ekkor $\|x^* - y^*\| = \|f(x^*) - f(y^*)\| \leq q\|x^* - y^*\| \Rightarrow$

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

ami csak $\|x^* - y^*\| = 0$ esetén lehetséges $\Rightarrow x^* = y^*$, ami ellentmondás.

1.4.13 Megjegyzés. *A fixponttétel bizonyítása automatikusan átvihető tetszőleges X normált térre, ha igaz, hogy minden X-beli Cauchy-sorozat egyben konvergens is. Az ilyen tulajdonságú tereket teljes normált térnek, más szóval Banach-térnek nevezzük. Példa: $(C[a,b], \|\cdot\|_\infty)$. De minden véges dimenziós normált tér is Banach-tér!*

2. fejezet

Lineáris algebrai egyenletrendszerek megoldása

A gyakorlati feladatoknál sűrűn előfordul, hogy lineáris egyenletrendszereket kell megoldanunk. Többek között a lineáris parciális differenciálegyenletek véges különbséges módszerrel történő megoldása is lineáris algebrai egyenletrendszerek megoldására vezet.

A lineáris egyenletrendszerek megoldására szolgáló módszerek két nagy csoportba oszthatók. A **direkt módszer**ekkel véges számú lépéssel meghatározhatjuk az egyenletrendszer pontos megoldását (ha minden lépésben pontosan számolunk). Az **iterációs módszer**ek egy iterációs vektorsorozatot generálnak, amely a pontos megoldásvektorhoz tart. A sorozat elég nagy indexű tagja a megoldást tetszőlegesen megközelíti.

Leszögezzük, hogy itt csak olyan egyenletrendszerek megoldásával foglalkozunk, amelyek mátrixa négyzetes. Az általános alak tehát

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots a_{1m}x_m &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots a_{2m}x_m &= b_2 \\ &\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots a_{mm}x_m &= b_m \end{aligned} \tag{2.1}$$

amely az

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mm} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \tag{2.2}$$

jelölések bevezetésével rövidítve

$$Ax = b \tag{2.3}$$

alakban is felírható. A továbbiakban mindvégig feltesszük, hogy $\det A \neq 0$, azaz A reguláris. Ismeretes, hogy ekkor az egyenletrendszernek létezik egyetlen $x^* \in \mathbb{R}^m$ megoldása.

2.1. Direkt módszerek

2.1.1. Gauss-elimináció

Az eliminációs vagy kiküszöbölési eljárások azon az észrevételen alapulnak, hogy egyes speciális egyenletrendszerek könnyen megoldhatók. A legegyszerűbb eset az, amikor az A együtthatómátrix diagonális. Tekintsük például az

$$\begin{aligned} 5x_1 &= 10 \\ 2x_2 &= 4 \\ 3x_3 &= -1 \end{aligned}$$

egyenletrendszert. Ebben az esetben

$$A = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \quad \text{és} \quad b = \begin{bmatrix} 10 \\ 4 \\ -1 \end{bmatrix},$$

és az ismeretlenek az egyenletekből közvetlenül kifejezhetők:

$$\begin{aligned} x_1 &= 2 \\ x_2 &= 2 \\ x_3 &= -\tfrac{1}{3}. \end{aligned}$$

Könnyű dolgunk van háromszögmátrixú egyenletrendszerek esetén is. Az

$$\begin{aligned} x_1 \quad +2x_2 \quad +2x_3 &= 7 \\ 4x_2 \quad +5x_3 &= 17 \\ 9x_3 &= 27 \end{aligned}$$

egyenletrendszer mátrixa felsőháromszög-mátrix. Az utolsó egyenletből kifejezhető az x_3 ismeretlen: $x_3 = 27/9 = 3$. Ezt behelyettesítjük a második egyenletbe, és kifejezzük x_2-t: $x_2 = 1/2$. Végül x_1 az első egyenletből adódik x_2 és x_3 behelyettesítésével: $x_1 = 7 - 2 \cdot 3 - 2 \cdot \frac{1}{2} = 0$.

A módszer általános felírásához induljunk ki az egyenletrendszer részletes (2.1) alakjából! Fokozatosan eliminálunk az $x_1, x_2, \ldots, x_m$ változókat az egyenletekből.

1. lépés: Tegyük fel, hogy $a_{11} \neq 0$. Ekkor (2.1) első egyenlete felírható

$$x_1 + c_{12}x_2 + \ldots + c_{1m}x_m = y_1 \tag{2.4}$$

alakban, ahol

$$c_{1j} = \frac{a_{1j}}{a_{11}},\ j = 2, 3, \ldots, m. \tag{2.5}$$

és

$$y_1 = \frac{b_1}{a_{11}}.$$

2. lépés: A (2.1) maradék $m-1$ egyenlete

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{im}x_m = b_i,\ i = 2, 3, \ldots, m \tag{2.6}$$

Így (2.6)-ből kivonva a (2.4) egyenlet a_{i1}-szeresét ($i = 2, \ldots, m$), x_1 eliminálható a $2., 3., \ldots m.$ egyenletből. Ezzel az egyenletrendszer:

$$\begin{aligned} x_1 + c_{12}x_2 + \ldots + c_{1m}x_m &= y_1 \\ a_{22}^{(1)}x_2 + \ldots + a_{2m}^{(1)}x_m &= b_2^{(1)} \\ &\ldots \\ a_{m2}^{(1)}x_2 + \ldots + a_{mm}^{(1)}x_m &= b_m^{(1)}, \end{aligned} \tag{2.7}$$

ahol $a_{ij}^{(1)} = a_{ij} - c_{1j}a_{i1}$, $b_i^{(1)} = b_i - a_{i1}y_1$, $i, j = 2, \ldots, m$.

3. lépés: Tegyük fel, hogy a 2. egyenletben $a_{22}^{(1)} \neq 0$. Ekkor ez az egyenlet felírható

$$x_2 + c_{23}x_3 + \ldots + c_{2m}x_m = y_2 \tag{2.8}$$

alakban, ahol

$$c_{2j} = \frac{a_{2j}^{(1)}}{a_{22}^{(1)}},\ j = 3, 4, \ldots, m. \tag{2.9}$$

4. lépés: (2.8) felhasználásával a maradék $m-2$ egyenletből az x_2 ismeretlen eliminálható, és így tovább.

Mátrixos felírás: Az eljárás során az egyes lépésekben nyert lineáris egyenletrendszerek

együtthatómátrixának a struktúrája a következőképpen alakul:

$$\begin{bmatrix} x & x & x & \dots & x \\ x & x & x & \dots & x \\ x & x & x & & x \\ \vdots & \vdots & \vdots & & x \\ x & x & x & \dots & x \end{bmatrix} \rightarrow \begin{bmatrix} 1 & c_{12} & c_{13} & \dots & c_{1m} \\ 0 & x & x & \dots & x \\ 0 & x & x & \dots & x \\ \vdots & \vdots & \vdots & & x \\ 0 & x & x & \dots & x \end{bmatrix} \rightarrow \begin{bmatrix} 1 & c_{12} & c_{13} & \dots & c_{1m} \\ 0 & 1 & c_{23} & \dots & c_{2m} \\ 0 & x & x & \dots & x \\ \vdots & \vdots & \vdots & & x \\ 0 & x & x & \dots & x \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} 1 & c_{12} & c_{13} & \dots & c_{1m} \\ 0 & 1 & c_{23} & \dots & c_{2m} \\ 0 & 0 & 1 & \dots & c_{3m} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Tehát az algoritmus a megoldandó lineáris egyenletrendszert olyan lineáris egyenletrendszerré transzformálja ekvivalens átalakításokkal, amelynek a mátrixa felső háromszögű, a főátlóban 1-esekkel. Ezt nevezzük a Gauss-elimináció direkt irányának. A második ütemben a direkt irány után kiszámítható az x_m ismeretlen, és ebből fokozatosan meghatározzuk az $x_{m-1}, x_{m-2}, \dots, x_1$ ismeretleneket a következő algoritmussal:

$$\begin{aligned} x_m &= y_m \\ x_i &= y_i - \sum_{j=i+1}^{m} c_{ij} x_j, \ i = m-1, m-2, \dots, 1. \end{aligned}$$

Példa. Oldjuk meg a

$$\begin{aligned} 2x_1 \quad +x_2 \quad +x_3 &= 4 \\ x_1 \quad +3x_2 \quad +2x_3 &= 6 \\ x_1 \quad +2x_2 \quad +2x_3 &= 5 \end{aligned}$$

egyenletrendszert! Készítsük el az együtthatók és a szabad tagok táblázatát, az átalakításokat ezen a táblázaton követhetjük nyomon (hiszen szükségtelen az ismeretleneket minden lépésben kiírni).

2	1	1	4
1	3	2	6
1	2	2	5

Lépések:

1. Először az első sor minden elemét osztjuk 2-vel:

1	1/2	1/2	2
1	3	2	6
1	2	2	5

2. A 2. és a 3. egyenletből kivonjuk az 1. egyenletet:

1	1/2	1/2	2
0	5/2	3/2	4
0	3/2	3/2	3

3. A 2. sort végigszorozzuk 2/5-del:

1	1/2	1/2	2
0	1	3/5	8/5
0	3/2	3/2	3

4. A 3. sorból kivonjuk a 2. sor 3/2-szeresét.

1	1/2	1/2	2
0	1	3/5	8/5
0	0	3/5	3/5

5. A 3. sort végigszorozzuk 5/3-dal:

1	1/2	1/2	2
0	1	3/5	8/5
0	0	1	1

Tehát a megoldandó egyenletrendszer ekvivalens az

$$\begin{aligned} x_1 \quad +\tfrac{1}{2}x_2 \quad +\tfrac{1}{2}x_3 \quad &= \quad 2 \\ x_2 \quad +\tfrac{3}{5}x_3 \quad &= \quad \tfrac{8}{5} \\ x_3 \quad &= \quad 1 \end{aligned}$$

egyenletrendszerrel. Ebből az ismeretleneket alulról fölfelé kiküszöbölve az

$$x_1 = 1, \quad x_2 = 1, \quad x_3 = 1$$

megoldáshoz jutunk.

2.1.2. A Gauss-elimináció végrehajthatósága

Kérdés: mikor hajtható végre a Gauss-elimináció? Az eljárás során a (2.3) lineáris egyenletrendszert ekvivalens átalakításokkal

$$Ux = y$$

alakra hoztuk, ahol U felsőháromszög-mátrix, és $\operatorname{diag} U = I$. Mi a kapcsolat b és y között?

Mint láttuk,

$$y_1 = \frac{b_1}{a_{11}} \Rightarrow b_1 = a_{11}y_1.$$

Határozzuk meg b_2 értékét!

$$(2.7) \;\Rightarrow\; a_{22}^{(1)}x_2 + \ldots + a_{2m}^{(1)}x_m = b_2^{(1)} = b_2 - a_{21}y_1 \;\Rightarrow$$

$$x_2 + c_{23}x_3 + \ldots + c_{2m}x_m = \frac{b_2 - a_{21}y_1}{a_{22}^{(1)}} =: y_2 \Rightarrow b_2 = a_{21}y_1 + a_{22}^{(1)}y_2.$$

Általában is igaz:

$$b_j = l_{j1}y_1 + l_{j2}y_2 + \ldots + l_{jj}y_j,\; j = 1, 2, \ldots, m,$$

ahol $l_{jj} = a_{jj}^{(j-1)}$.
Tehát b és y kapcsolata:

$$b = Ly,$$

ahol L egy alsóháromszög-mátrix, amelynek főátlójában $a_{jj}^{(j-1)}$ $(j = 1, 2, \ldots, m)$, $a_{11}^{(0)} = a_{11}$ elemek állnak. Mivel feltevésünk szerint $a_{jj}^{(j-1)} \neq 0$, $j = 1, 2, \ldots, m$, ezért létezik L^{-1}, azaz $y = L^{-1}b$. Így $Ux = L^{-1}b$, azaz

$$LUx = b. \tag{2.10}$$

Innen adódik a következő új módszer:

1. Állítsuk elő az A mátrixot $L \cdot U$ alakban, ahol L invertálható alsó háromszögű mátrix, U pedig felső háromszögű mátrix, a főátlóban 1-esekkel.

2. Oldjuk meg y-ra az $Ly = b$ lineáris egyenletrendszert.

3. Az előző lépésben kiszámított y vektorral mint jobb oldallal oldjuk meg az $Ux = y$ lineáris egyenletrendszert.

A fenti módszer neve: LU-felbontási módszer.

Kérdés: Az A mátrix többféleképpen is felbontható-e LU alakban, vagy az L és U mátrix a Gauss-eliminációnál látott mátrixok lehetnek csak?

2.1.1 Állítás. *Az LU-felbontás egyértelmű.*

Biz.: Indirekt: tegyük fel, hogy két felbontás is van:

$$A = L_1U_1 = L_2U_2.$$

Szorozzuk meg az egyenlőséget balról az L_1^{-1}, jobbról az U_2^{-1} mátrixszal:

$$U_1U_2^{-1} = L_1^{-1}L_2.$$

Mivel felső ill. alsó háromszögű mátrixok inverze és szorzata is felső ill. alsó háromszögű, ezért ez utóbbi egyenlőség csak úgy állhat fenn, ha mindkét oldalon diagonális mátrix van. Mivel $U_1U_2^{-1}$ diagonális elemei 1-esek, ezért

$$U_1U_2^{-1} = L_1^{-1}L_2 = I \Rightarrow U_1 = U_2 \text{ és } L_1 = L_2.$$

2.1.2 Következmény. *A Gauss-elimináció pontosan akkor hajtható végre, ha A-nak létezik LU-felbontása.*

Kérdés tehát: mikor létezik az LU-felbontás?

Legyen

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

2.1.3 Tétel. *Tegyük fel, hogy $\Delta_j \neq 0, \ j = 1, 2, \ldots, m$. Ekkor létezik az LU-felbontás.*

Biz.: Teljes indukcióval:

- $m = 2$ eset: Az
$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} 1 & u_{12} \\ 0 & 1 \end{bmatrix}$$
egyenlőség pontosan akkor állhat fenn, ha megoldható a következő egyenletrendszer:
$$\begin{aligned} l_{11} &= a_{11} \\ l_{11}u_{12} &= a_{12} \\ l_{21} &= a_{21} \\ l_{21}u_{12} + l_{22} &= a_{22} \end{aligned}$$

Ennek létezik egyetlen megoldása a $\Delta_1 = a_{11} \neq 0$ feltétel miatt:

$$\begin{aligned} l_{11} &= a_{11} \\ u_{12} &= \frac{a_{12}}{a_{11}} \\ l_{21} &= a_{21} \\ l_{22} &= a_{22} - l_{21}u_{12} = a_{22} - a_{21}\frac{a_{12}}{a_{11}} = \frac{\det A}{a_{11}}. \end{aligned}$$

Továbbá, $\det L \neq 0$, hiszen $l_{11} \neq 0$, és $\det A \neq 0$ miatt $l_{22} \neq 0$.

- Tegyük fel, hogy az állítás igaz $m = k-1$-re, és mutassuk meg, hogy akkor igaz $m = k$-ra is. Ahhoz, hogy ki tudjuk használni az indukciós feltevést, elkülönítjük A-ban a bal felső $(k-1) \times (k-1)$-es blokkot:

$$A =: A_k = \begin{bmatrix} A_{k-1} & a_{k-1} \\ b_{k-1} & a_{kk} \end{bmatrix} \in \mathbb{R}^{k\times k},$$

ahol

$$A_{k-1} = \begin{bmatrix} a_{11} & \dots & a_{1,k-1} \\ \vdots & & \vdots \\ a_{k-1,1} & \dots & a_{k-1,k-1} \end{bmatrix} \in \mathbb{R}^{(k-1)\times(k-1)},$$

ahol az

$$a_{k-1} = \begin{bmatrix} a_{1k} \\ \vdots \\ a_{k-1,k} \end{bmatrix} \in \mathbb{R}^{k-1}, \quad b_{k-1} = [a_{k1}, \dots, a_{k,k-1}] \in \mathbb{R}^{k-1}.$$

Mivel $\Delta_j \neq 0,\ j = 1, \dots, k-1$, ezért az indukciós feltevés alapján

$$A_{k-1} = L_{k-1}U_{k-1},$$

ahol L_{k-1} és U_{k-1} a felbontásnak megfelelő tulajdonságú mátrixok. Megmutatjuk, hogy A_k felbontható

$$A_k = \begin{bmatrix} L_{k-1} & 0 \\ l_{k-1} & l_{kk} \end{bmatrix} \begin{bmatrix} U_{k-1} & u_{k-1} \\ 0 & 1 \end{bmatrix} \tag{2.11}$$

alakban, ahol $l_{k-1}, u_{k-1} \in \mathbb{R}^{k-1}$ sor- ill. oszlopvektorok, és $l_{kk} \in \mathbb{R}$. Ehhez szükségesek:
$(C1)$ $L_{k-1}U_{k-1} = A_{k-1}$

$(C2)$ $L_{k-1}u_{k-1} = a_{k-1}$
$(C3)$ $l_{k-1}U_{k-1} = b_{k-1}$
$(C4)$ $l_{k-1}u_{k-1} + l_{kk} \cdot 1 = a_{kk}$
A (C1) egyenlőség következik az indukciós feltevésből. Kérdés, van-e megoldása l_{k-1}, u_{k-1}, l_{kk}-ra a (C2)–(C4) egyenletrendszernek?
Mivel létezik L_{k-1}^{-1} és U_{k-1}^{-1}, ezért
$(C2) \Rightarrow u_{k-1} = L_{k-1}^{-1}a_{k-1}$
$(C3) \Rightarrow l_{k-1} = b_{k-1}U_{k-1}^{-1}$
$(C4) \Rightarrow l_{kk} = a_{kk} - l_{k-1}u_{k-1} = a_{kk} - b_{k-1}U_{k-1}^{-1}L_{k-1}^{-1}a_{k-1} = a_{kk} - b_{k-1}A_{k-1}^{-1}a_{k-1}$.
Még be kell látnunk, hogy $l_{kk} \neq 0$!

$$(2.11) \Rightarrow \det A_k = \det L_{k-1} \cdot l_{kk} \cdot \det U_{k-1} \cdot 1 = \det L_{k-1} l_{kk}.$$

A feltevésünk alapján $\det A_k = \Delta_k \neq 0$. Így $l_{kk} \neq 0$. Tehát létezik az LU-felbontása A_k-nak.

2.1.4 Megjegyzés. *Ha valamelyik bal felső sarokaldeterminánsa 0, akkor nem létezik LU-felbontás. (Ez a* 2×2*-es esetben jól látható: tegyük fel, hogy* $\det A \neq 0$ *(ezt mindig feltesszük), és* $a_{11} = 0$*. Az*

$$\begin{bmatrix} 0 & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} 1 & u_{12} \\ 0 & 1 \end{bmatrix}$$

egyenlőség fennállásához szükséges, hogy $l_{11} \cdot 1 = 0$ *legyen, ekkor azonban* L *determinánsa nulla lenne, így nem lehetne invertálható.)*

Ezért a Gauss-elimináció pontosan akkor hajtható végre, ha A minden bal felső főminora nullától különbözik.

Kérdés, mit tegyünk ha ez a feltétel nem teljesül.

2.1.5 Definíció. *Az olyan* $P \in \mathbb{R}^{m \times m}$ *mátrixot, amelynek mindegyik sorában és oszlopában egy elem 1, a többi pedig nulla, permutáló mátrixnak nevezzük.*

Tetszőleges $A \in \mathbb{R}^{m \times m}$ mátrixot balról egy permutáló mátrixszal szorozva A sorainak a sorrendje változik meg.

2.1.6 Állítás. *Legyen* $A \in \mathbb{R}^{m \times m}$, $\det A \neq 0$*. Ekkor létezik olyan* P *permutáló mátrix, amelyre* $P \cdot A = L \cdot U$*.*

Ez az állítás biztosítja, hogy tetszőleges, reguláris mátrixú lineáris algebrai egyenletrendszer mindig áttranszformálható sorcserékkel olyan rendszerré, amelyre már alkalmazható a Gauss-elimináció.

Jó tudnunk, mikor nem kell PA előállítás, ez ugyanis nagy műveletigényű. A következő speciális esetekben nincs szükség permutáló mátrixra:

- A szigorúan domináns főátlójú[1];
- A szimmetrikus pozitív definit mátrix.

Az utóbbi esetben létezik egy másik felbontás is:

$$A = G \cdot G^T,$$

ahol G egy alsó háromszögű mátrix pozitív főátlóval. Ezt a felbontást Cholesky-felbontásnak nevezzük.

2.1.3. Gauss-elimináció főelem-kiválasztással

Ha kerekítve vagy számítógéppel számolnunk, akkor a Gauss-eliminációt érdemes az ún. **főelem-kiválasztás** módszerével végezni. Láthattuk, hogy a szabályos Gauss-eliminációs algoritmus j-edik lépésében, az aktuális táblázat j-edik oszlopában úgy tudjuk nullára transzformálni az $a_{jj}^{(j-1)}$ alatti elemeket, hogy a j-edik sort először elosztjuk $a_{jj}^{(j-1)}$-gyel, majd az így kapott sor megfelelő számszorosát kivonjuk a lejjebb lévő sorokból. Mivel kis számmal való osztásnál nagy lehet a kerekítési hiba hatása, ezért kedvezőtlen, ha az $a_{jj}^{(j-1)}$ elem kis abszolút értékű. Ennek elkerülésére szolgál a főelem-kiválasztás.

A **részleges főelem-kiválasztás** során megvizsgáljuk, hogy az adott oszlopban a főátló alatt van-e a főátlóbeli elemnél nagyobb abszolút értékű szám, és ha igen, akkor sorcserével a főátlóba hozzuk (lásd a 2.1. táblázatot).

Még jobban csökkenthető a számítási hiba a **teljes főelem-kiválasztás**sal. Ilyenkor a táblázatnak a főátlóbeli elemből jobbra és lefelé kiinduló legnagyobb négyzetes blokkjában keressük a legnagyobb abszolút értékű elemet (főelem). Ennek főátlóba hozásához esetleg sor- és oszlopcserét is végre kell hajtanunk. Az utóbbinál arra kell vigyázni, hogy megváltozik az oszlopok számozása. (Pl. ha a 4. oszlopot áthozzuk a 2. oszlopba, akkor onnantól kezdve a 2. oszlopban lévő számok az x_4 ismeretlen együtthatói lesznek!) Egy példát mutatunk be a 2.2. táblázatban.

[1] Az $A \in \mathrm{R}^{m\times m}$ mátrixot szigorúan domináns főátlójú mátrixnak nevezzük, ha $|a_{ii}| > \sum_{j=1,j\neq i} |a_{ij}|, \quad i = 1, 2, \ldots, m$ azaz amikor minden sorában a főátlóban lévő elem abszolút értékben nagyobb a többi elem abszolút értékének összegénél.

1	3	6	2	7
0	**2**	5	4	1
0	**50**	5	10	-15
0	**3**	2	5	-9
1	3	6	2	7
0	50	5	10	-15
0	2	5	4	1
0	3	2	5	-9

2.1. táblázat. Példa részleges főelem-kiválasztásra. Mielőtt nullára transzformálnánk a 2. oszlopban lévő, főátló alatti elemeket, felcseréljük a 2. és a 3. sort. Ezzel a főátlóba kerül az oszlop legnagyobb abszolút értékű főátló alatti eleme.

1	3	6	2	7
0	**1**	**5**	**4**	1
0	**5**	**5**	**10**	-15
0	**3**	**2**	**5**	-9
1	3	6	2	7
0	5	5	10	-15
0	1	5	4	1
0	3	2	5	-9
1	2	6	3	7
0	10	5	5	-15
0	4	5	1	1
0	5	2	3	-9

2.2. táblázat. Példa teljes főelem-kiválasztásra. Megkeressük a legnagyobb abszolút értékű elemet a vastagon szedett blokkban. Ezt úgy hozhatjuk a főátlóba, hogy felcseréljük a 2. és a 3. sort, majd a 2. és a 4. oszlopot. Mostantól a 2. oszlop tartalmazza az x_4 együtthatóit, és a 4. oszlop az x_2 együtthatóit.

Az alábbi példán bemutatjuk a főelem-kiválasztás hatását a megoldás pontosságára. Tegyük fel, hogy Gauss-eliminációval szeretnénk megoldani a

$$\begin{aligned} 0,00031x_1 \quad +x_2 \quad &= \quad 3 \\ x_1 \qquad\quad +x_2 \quad &= \quad 7 \end{aligned}$$

egyenletrendszert. (Eláruljuk, hogy ennek pontos megoldása (kerekítve) $x_1 = 4,001$, $x_2 = 2,999$.) Alkalmazzuk a Gauss-eliminációt először főelem-kiválasztása nélkül, és mindig kerekítsünk négy értékes jegyre:

$$\left[\begin{array}{cc|c} 0,00031 & 1 & 3 \\ 1 & 1 & 7 \end{array}\right] \rightarrow^{(1.)/0,00031,\ \text{majd}\ (2.)-(1.)} \rightarrow \left[\begin{array}{cc|c} 1 & 3226 & 9677 \\ 0 & -3225 & -9670 \end{array}\right]$$

Ezzel a

$$\begin{aligned} x_1 \quad +3226x_2 &= 9677 \\ 3225x_2 &= 9670 \end{aligned}$$

egyenletrendszerhez jutottunk. Ennek megoldása négy értékes jegyre kerekítve: $x_1 = 6,452,\ x_2 = 2,998$. Vegyük észre, hogy az első ismeretlen értékét nagyon nagy hibával kaptuk meg!
Ha főelem-kiválasztással számolunk, akkor a

$$\left[\begin{array}{cc|c} 1 & 1 & 7 \\ 0,00031 & 1 & 3 \end{array}\right]$$

táblázatból indulunk ki. Könnyen végigszámolható, hogy ekkor a megoldás kerekítve $x_1 = 4,002,\ x_2 = 2,998$, azaz most már x_1-et is megfelelő pontossággal kaptuk meg. Vegyük észre, hogy a pontosságot mindenféle pluszmunka nélkül tudtuk jelentősen megnövelni.

2.2. Iterációs módszerek

Ha a megoldandó lineáris algebrai egyenletrendszer mátrixa nagyméretű és ritka (azaz viszonylag sok nulla elemet tartalmaz), akkor a Gauss-elimináció ill. az LU-felbontás alkalmazása során fölöslegesen sok műveletet végzünk el, nem tudjuk jól kihasználni azt a tényt, hogy csak kevés nemnulla elem van a mátrixban. Ilyenkor érdemes lehet inkább valamilyen iterációs módszert alkalmazni. Az iterációs módszerek lényege az, hogy felépítünk egy, a megoldáshoz konvergáló vektorsorozatot.

2.2.1. Klasszikus iterációs módszerek

Az iterációs módszerek egyik része azon alapul, hogy az $Ax = b$ egyenletrendszert $x = Qx + r$ alakú egyenletrendszerré transzformáljuk, ahol $Q \in \mathbb{R}^{m\times m}$ és $r \in \mathbb{R}^m$. Vegyük észre, hogy ekkor a megoldás az $f : \mathbb{R}^m \to \mathbb{R}^m$, $f(x) = Qx + r$ leképezés fixpontja. Az egyenletrendszer iterációs megoldása során olyan vektorsorozatot konstruálunk, amely az f leképezés fixpontjához, azaz a megoldásvektorhoz tart.

Itt természetesen a már tanult 1.4.12. tételre támaszkodhatunk. Tekintsük most az $f(x) = Qx + r$ függvényt, ahol $Q \in \mathbb{R}^{m\times m}$ és $r \in \mathbb{R}^m$. Ekkor $f : \mathbb{R}^m \to \mathbb{R}^m$. Vizsgáljuk meg, hogy f mikor kontrakció!

Legyen x és y két tetszőleges $\mathbb{R}^m$-beli vektor. Írjuk fel az $f(x) - f(y)$ különbséget:

$$f(x) - f(y) = Qx + r - Qy - r = Q(x - y).$$

Vegyük mindkét oldal valamely $\|\cdot\|_{\mathbb{R}^m}$ vektornormáját!

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

ahol $\|Q\|$ a Q mátrix $\|\cdot\|_{\mathbb{R}^m}$ vektornorma által indukált mátrixnormája. Világos, hogy ha $\|Q\| < 1$, akkor az f függvény $q := \|Q\|$ mellett kontrakció (az adott $\|\cdot\|_{\mathbb{R}^m}$ vektornormában). Ezért a fixponttétel értelmében $x^n \to x^*$ (az adott $\|\cdot\|_{\mathbb{R}^m}$ vektornormában, és így minden más $\mathbb{R}^m$-beli vektornormában is a normák ekvivalenciája miatt). Tehát a konvergenciának elégséges feltétele, hogy valamelyik indukált mátrixnormában $\|Q\|$ kisebb legyen, mint 1.

Természetes módon vetődik fel a következő két kérdés:

1. Hogyan tudunk egy $Ax = b$ egyenletrendszert $x = Qx + r$ alakú rendszerré transzformálni?

2. Mikor teljesül, hogy az így kapott Q mátrix valamelyik indukált mátrixnormában kisebb, mint 1?

A továbbiakban néhány konkrét módszert ismertetünk.

Jacobi-iteráció

Bontsuk fel az A mátrixot $A = L + D + U$ alakban, ahol L és U alsó ill. felső háromszögű mátrix, a főátlóban nulla elemekkel, D pedig diagonálmátrix. Tegyük fel, hogy D főátlójában nincs nulla elem, azaz D invertálható. Ezzel az egyenletrendszer

$$(L + D + U)x = b.$$

Vigyük át a jobb oldalra az L és az U mátrixot tartalmazó tagokat:

$$Dx = -(L + U)x + b,$$

majd szorozzunk D inverzével:

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

Ezzel sikerült az egyenletrendszert $x = Qx + r$ alakra hoznunk, ahol speciálisan

$$Q = Q_J = -D^{-1}(L + U) \quad \text{és} \quad r = D^{-1}b.$$

A Jacobi-iteráció n-edik lépésében az

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

összefüggés szerint számolunk.
Megmutatható: a $Q = -D^{-1}(L+U)$ mátrix sornormája pontosan akkor kisebb 1-nél, amikor A szigorúan domináns főátlójú. Ezért ilyen mátrixú egyenletrendszerre a Jacobi-iteráció mindig konvergens. A fejezet későbbi részében általánosabb iterációs eljárás konvergenciáját is vizsgáljuk majd más úton.

Példa. Tekintsük a

$$\begin{array}{rrrcl} 4x_1 & -x_2 & & = & 0 \\ -x_1 & +4x_2 & -x_3 & = & 6 \\ & -x_2 & +4x_3 & = & 2 \end{array}$$

egyenletrendszert. Legyen a Jacobi-iteráció kiindulási vektora $x^0 := (0,0,0)$. Konvergens-e a Jacobi-iteráció? Ha igen, számítsuk ki az x^1 és x^2 közelítést. Adjuk meg az n-edik iteráció képlethibáját.
Megoldás: Az A mátrix szigorúan domináns főátlójú, ezért Q sornormája kisebb 1-nél, így az iteráció konvergens. Az A együtthatómátrix felbontásában szereplő mátrixok:

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

Ezekből

$$Q = -D^{-1}(L+U) = \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & 0 \end{bmatrix}, \quad r = D^{-1}b = \begin{bmatrix} 0 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix}.$$

Az első iteráció

$$x^1 = Qx^0 + r = r = \begin{bmatrix} 0 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix}.$$

A második iteráció

$$x^2 = Qx^1 + r = \begin{bmatrix} \frac{3}{8} \\ \frac{1}{8} \\ \frac{3}{8} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{8} \\ \frac{13}{8} \\ \frac{7}{8} \end{bmatrix}.$$

A sornormát, amelyben Q-ról tudjuk, hogy kisebb, mint 1, a maximumnorma indukálja. Így az n-edik iteráció képlethibája maximumnormában

$$\|x^n - x^*\| = \max_i |x_i^n - x_i^*| \le \frac{\|Q\|^n}{1-\|Q\|} \cdot \max_i |x_i^1 - x_i^0| = 3 \cdot \left(\frac{1}{2}\right)^n,$$

ahol felhasználtuk, hogy a Q mátrix sornormája $\|Q\| = \frac{1}{2}$.

Gauss–Seidel-iteráció

A Gauss–Seidel-iterációnál is $L + D + U$ alakban bontjuk fel az A mátrixot, de most az Ux tagot visszük át a jobb oldalra, és $(L + D)$ inverzével szorzunk. Így az

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

alakhoz jutunk. Azaz most $Q = Q_{GS} = -(L + D)^{-1}U$ és $r = (L + D)^{-1}b$.

Az egylépéses lineáris iterációk általános alakja

Nyilvánvalóan az előbbi iterációs módszerek felírhatók a következő alakban:

Jacobi-iteráció: $D(x^{n+1} - x^n) + Ax^n = b$

Gauss–Seidel-iteráció: $(D + L)(x^{n+1} - x^n) + Ax^n = b$.

Ezeket a Jacobi- ill. a Gauss–Seidel-iteráció kanonikus alakjának nevezzük. A két iterációt módosíthatjuk új paraméterek bevezetésével:

$$D\frac{x^{n+1} - x^n}{\tau_{n+1}} + Ax^n = b,$$

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

Itt a τ_n paraméterek megválasztása az iteráció konvergenciájával függ össze.

A fenti kanonikus alakokat szem előtt tartva felírhatunk egy általános módszert. Legyenek $B_{n+1} \in \mathbb{R}^{m\times m}$ tetszőleges, rögzített reguláris mátrixok, és τ_{n+1} adott paraméterek $(n = 1, 2, \ldots)$. Ekkor a

$$B_{n+1}\frac{x^{n+1} - x^n}{\tau_{n+1}} + Ax^n = b \tag{2.12}$$

iterációt általános alakú lineáris iterációnak nevezzük.

2.2.1 Megjegyzés. *A B_{n+1} mátrix regularitása fontos, hiszen nélküle nem lehetne x^n vektorból x^{n+1}-et kiszámítani.*

2.2.2 Megjegyzés. *Ha $B_n = I$ minden n-re, akkor (2.12) iteráció explicit iteráció, egyébként implicit.*

2.2.3 Megjegyzés. *Az implicit iteráció csak akkor gazdaságos, ha B_n^{-1} kiszámolása jóval olcsóbb, mint az A^{-1} inverz mátrixé.*

2.2.4 Definíció. *Ha $B_n = B$ és $\tau_n = \tau$ minden n esetén, akkor az iterációt stacionáriusnak nevezzük.*

Példa 1. $B_n = I$ és $\tau_n = \tau$:

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

Elnevezés: egyszerű iteráció.

Példa 2. $B_n = I$ és $\tau_n > 0$ változó:

$$\frac{x^{n+1} - x^n}{\tau_{n+1}} + Ax^n = b.$$

Elnevezés: Richardson-féle iteráció.

Példa 3. Legyen $\omega > 0$ egy tetszőleges szám, $B_n = D + \omega L$, $\tau_n = \omega$. Ekkor

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

elnevezése: túlrelaxációs módszer (angolul successive over-relaxation method - innen a SOR-módszer elnevezés). Az ω neve relaxációs paraméter, és a „túlrelaxációs" jelző arra utal, hogy ω optimális értéke rendszerint 1-nél nagyobb.

A stacionárius iterációk konvergenciája

A továbbiakban a

$$B\frac{x^{n+1} - x^n}{\tau} + Ax^n = b \tag{2.13}$$

iteráció konvergenciájával foglalkozunk, ahol B, τ a módszert meghatározó mátrix ill. paraméter.

A differenciálegyenletek numerikus megoldása során a leggyakrabban előforduló mátrixtípusok közé tartoznak a szimmetrikus pozitív definit mátrixok és a szigorúan domináns főátlójú mátrixok, és több állításunk is az ilyen mátrixú lineáris egyenletrendszerre vonatkozik. Először azt vizsgáljuk meg, amikor A szimmetrikus, pozitív definit mátrix. Emlékeztetőül: ha A pozitív definit (azaz $(Ax, x) > 0\ \forall x \neq 0$ esetén), akkor létezik $\delta > 0$ állandó, amelyre

$$(Ax, x) \geq \delta \|x\|^2.$$

Jelölje x^* a (2.3) egyenlet megoldását, és legyen

$$e_n := x^n - x^*,$$

az n-edik iteráció hibája. Mivel $Ax^* = b$, így

$$B\frac{e_{n+1} - e_n}{\tau} + Ae^n = 0,\ n = 0, 1, \ldots, \tag{2.14}$$

és $e_0 = x_0 - x^*$.

2.2.5 Definíció. *Azt mondjuk, hogy a (2.13) eljárás konvergens, ha* $\lim_{n\to\infty} e_n = 0$.

2.2.6 Állítás. *Tegyük fel, hogy A szimmetrikus, pozitív definit mátrix. Ha B invertálható, és $\tau > 0$ olyan paraméter, amely mellett a $B - 0,5\tau A$ mátrix szimmetrikus, pozitív definit, akkor a (2.13) iterációs eljárás konvergens.*

Biz.: Belátandó, hogy a (2.14) szerinti (e_n) sorozatra $\lim e_n = 0$.

$$(2.14) \Rightarrow Be_{n+1} = (B - \tau A)e_n$$

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

Az A mátrixszal balról szorozva:

$$Ae_{n+1} = (A - \tau AB^{-1}A)e_n$$

A két utóbbi egyenlőségből

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

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

Mivel A szimmetrikus, ezért

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

Így a $J_n = (Ae_n, e_n)$ jelöléssel

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

Vezessük be az $y_n := B^{-1}Ae_n$ jelölést! Ekkor $By_n = Ae_n$, és így

$$J_{n+1} = J_n - 2\tau(By_n, y_n) + \tau^2(Ay_n, y_n)$$

$$= J_n - 2\tau[(By_n, y_n) - \frac{\tau}{2}(Ay_n, y_n)] = J_n - 2\tau((B - 0,5\tau A)y_n, y_n). \tag{2.15}$$

Mivel $B-0,5\tau A$ szpd, ezért $((B-0,5\tau A)y_n, y_n) \geq 0$, és így $J_{n+1} \leq J_n$. Másrészt, $J_n \geq 0$, így (J_n) egy monoton csökkenő, alulról korlátos sorozat. Ezért létezik véges határértéke, amelyet jelöljön J^*. Áttérve (2.15)-ban a $\lim_{n\to\infty}$ határértékekre:

$$J^* = J^* - 2\tau \lim_{n\to\infty}((B - 0,5\tau A)y_n, y_n),$$

azaz

$$\lim_{n\to\infty}((B - 0,5\tau A)y_n, y_n) = 0.$$

Másrészt,

$$((B - 0,5\tau A)y_n, y_n) \geq \delta\|y_n\|^2,$$

hiszen $B - 0,5\tau A$ pozitív definit. Így a két utóbbi összefüggésből

$$\lim_{n\to\infty}\|y_n\| = 0.$$

Mivel $e_n = A^{-1}By_n$, ezért

$$0 \leq \|e_n\| \leq \|A^{-1}B\| \cdot \|y_n\|,$$

és mivel a jobb oldali sorozat nullához tart, a rendőrelv alapján $\lim_{n\to\infty}\|e_n\| = 0$, azaz $\lim e_n = 0$.

Részletesebben foglalkozzunk most a SOR-módszer konvergenciájával. Az ω paramétert úgy kellene megválasztani, hogy a módszer konvergens legyen, és minél gyorsabban konvergáljon. Ezen paraméter optimális értékére nincs explicit formula általános esetben, ez az érték függ az együtthatómátrix tulajdonságaitól. Itt csak néhány fontos tételt említünk.

2.2.7 Állítás. *Legyen $A \in \mathbb{R}^{m\times m}$ egy tetszőleges mátrix. Ekkor a SOR-módszer konvergenciájának*

$$\omega \in (0, 2) \tag{2.16}$$

szükséges feltétele.

Biz.: A hibaegyenlet a SOR-módszer esetén

$$(D + \omega L)\frac{e_{n+1} - e_n}{\omega} + Ae_n = 0$$

$$(D+\omega L)e_{n+1} = (D+\omega L-\omega A)e_n = (D+\omega L-\omega L-\omega D-\omega U)e_n = [(1-\omega)D-\omega U]e_n.$$

Tehát $e_{n+1} = B_\omega e_n$, ahol $B_\omega = (D + \omega L)^{-1}[(1 - \omega)D - \omega U]$.
A konvergencia feltétele $\rho(B_\omega) < 1$, ami azt jelenti, hogy B_ω összes λ_i sajátértékére $|\lambda_i| < 1$. Ismeretes, hogy $\prod_{i=1}^m |\lambda_i| = |\det B_\omega|$. Tehát konvergencia esetén, mivel ekkor

$|\lambda_i| < 1$, $|\det B_\omega| < 1$. Mivel

$$|\det B_\omega| = |\det(D+\omega L)^{-1}| \cdot |\det[(1-\omega)D - \omega U]| = \left|\prod_{i=1}^{m} \frac{1}{a_{ii}}\right| \cdot \left|\prod_{i=1}^{m}(1-\omega)a_{ii}\right| = |1-\omega|^m.$$

Tehát $|1-\omega| < 1$, azaz $-1 < 1-\omega < 1$, és ez éppen a belátandó állítást jelenti. Most megmutatjuk, hogy ha A szimmetrikus, pozitív definit mátrix, akkor (2.16) egyben elégséges feltétel is.

2.2.8 Állítás. *(Reich, Ostrowski) Tegyük fel, hogy A szimmetrikus, pozitív definit mátrix. Ekkor (2.16) esetén a SOR-módszer konvergens.*

Biz.: Alkalmazható a 2.2.6. tétel, hiszen A megfelelő tulajdonságú. A SOR-módszernél $B = D + \omega L$, $\tau = \omega$, azaz a feltétel a

$$B - 0,5\tau A = D + \omega L - 0,5\omega(L + D + U) = (1 - 0,5\omega)D + 0,5\omega(L - U)$$

mátrix pozitív definitsége. Felhasználva A szimmetrikusságát, $U^T = L$, és a következő igaz minden $x \in \mathbb{R}^m$ vektorra:

$$((1 - 0,5\omega)Dx, x) + (0,5\omega(L - U)x, x) = (1 - 0,5\omega)(Dx, x) + 0,5\omega[(Lx, x) - (Ux, x)]$$

$$= (1-0,5\omega)(Dx, x)+0,5\omega[(Lx, x)-(x, U^T x)] = (1-0,5\omega)(Dx, x)+0,5\omega[(Lx, x)-(x, Lx)]$$

$$= (1 - 0,5\omega)(Dx, x)$$

Mivel $(Dx, x) > 0$ $\forall x \neq 0 \in \mathbb{R}^m$ esetén ($a_{ii} > 0$), a feltétel teljesül, azaz a SOR-módszer konvergens. Összefoglalva, a (2.16) feltétel szimmetrikus, pozitív definit mátrixokra a konvergencia szükséges és elégséges feltétele.

2.2.9 Következmény. *A Gauss–Seidel-iteráció szimmetrikus, pozitív definit mátrixokra konvergens.*

Megmutatható, hogy a Gauss–Seidel-iteráció szigorúan domináns főátlójú mátrixokra is konvergens. (Tehát nem kell a szimmetrikusság.)

Hasonlítsuk össze a Jacobi- és a Gauss–Seidel-iterációt! Általában nem összemérhető, de a szigorúan domináns főátlójú mátrixok osztályán a Gauss–Seidel-iteráció nem lehet lassúbb, mint a Jacobi-iteráció. Van azonban olyan eset is, amikor a Jacobi-iteráció konvergál, a Gauss–Seidel azonban nem!

2.2.2. Gradiens alapú módszerek

Ebben a fejezetben újabb iterációs módszert mutatunk szimmetrikus, pozitív definit mátrixú lineáris egyenletrendszerek megoldására. Az alapötlet az, hogy megadunk egy többváltozós, valós értékű függvényt, melynek abszolút minimumhelye az egyenletrendszer megoldása. Ezt a minimumhelyet keressük meg egy megfelelő iterációs eljárással.

Megoldandó az $Ax = b$ lineáris algebrai egyenletrendszer, ahol legyen most $A \in \mathbb{R}^{m\times m}$ szimmetrikus, pozitív definit mátrix. Tekintsük a következő $\mathbb{R}^m \to \mathbb{R}$ függvényt:

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

Deriváljuk ezen φ függvényt, azaz írjuk fel a gradiensét!

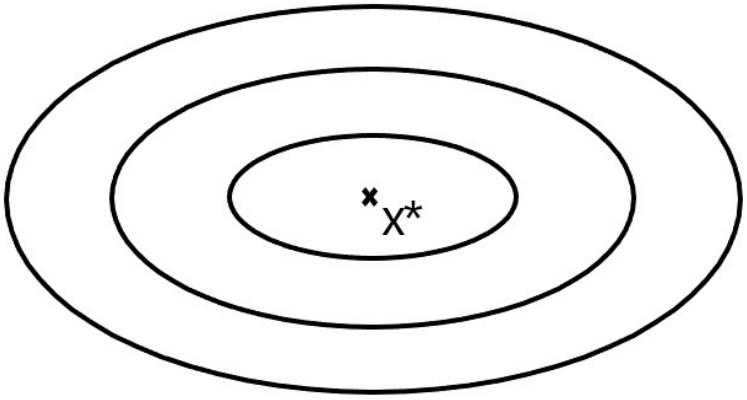
$$\varphi'(x) = \text{grad}\varphi(x) = Ax - b$$

Vegyük észre, hogy ez éppen az $r = b - Ax$ vektor (-1)-szerese. Továbbá, a gradiens egyetlen helyen nulla, és ez az $x^* = A^{-1}b$ vektor, a lineáris egyenletrendszer megoldása. Mivel a második derivált mátrixa

$$\varphi''(x) = A,$$

és A pozitív definit, így a megoldás a φ függvény abszolút minimumhelye.

Megjegyezzük, hogy $m = 2$ esetén a φ függvény szintvonalai ellipszisek, és a legmélyebb pontot keressük. Kérdés, adott A és b esetén hogyan tudjuk ezt megtalálni.



Ehhez először határozzuk meg, hogy adott x pontból adott p vektor mentén haladva hol van a legkisebb értéke φ-nek: vagyis mely α szám esetén lesz az $x+\alpha p$ pontban a legkisebb φ értéke.

2.2.10 Állítás. *Legyenek x és p adott vektorok, $p \neq 0$. A $g(\alpha) = \varphi(x + \alpha p)$ egyváltozós függvény egyértelmű minimumát az*

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

választás esetén veszi fel.

Biz.:

$$g(\alpha) = \varphi(x + \alpha p) = \frac{1}{2}(x + \alpha p, A(x + \alpha p)) - (x + \alpha p, b) =$$

$$\frac{1}{2}(x, Ax) - (x, b) + \frac{1}{2}\alpha(p, Ax) + \frac{1}{2}(x, A(\alpha p)) + \frac{1}{2}(\alpha p, A(\alpha p)) - \alpha(p, b)$$

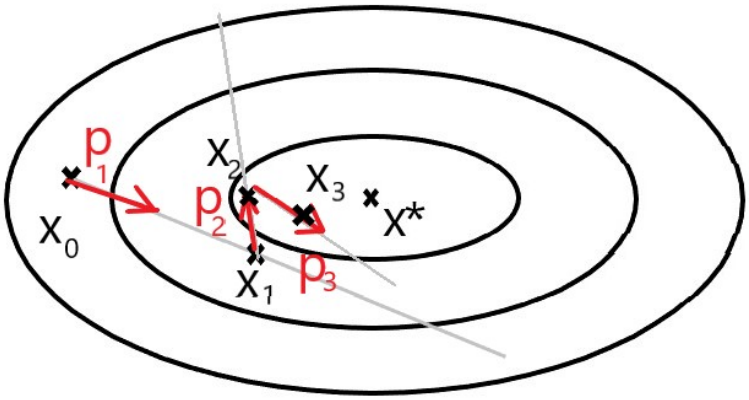
Felhasználva, hogy $b = r + Ax$, ez a következő alakra hozható:

$$g(\alpha) = \varphi(x) + \frac{1}{2}\alpha^2(p, Ap) - \alpha(p, r)$$

Ez egy felfelé álló parabola (ugyanis $(p, Ap) > 0$), így egyetlen minimumát ott veszi fel, ahol a deriváltja nulla:

$$g'(\alpha) = \alpha(p, Ap) - (p, r) = 0 \Rightarrow \alpha = \frac{(p, r)}{(p, Ap)}.$$

Ha tehát rögzítünk egy $x_0 \in \mathbb{R}^m$ vektort, akkor innen egy $p_1 \neq 0$ vektor irányában megkeresve a minimumhelyet kaphatunk egy jobb x_1 közelítést, onnan ismét egy p_2 vektor irányában keresve a minimumhelyet, egy még jobb x_2 közelítést, és így tovább. Kérdés, hogyan válasszuk meg a $p_1, p_2, \ldots$ keresési irányokat.



A gradiens-módszer

Mivel egy többváltozós, valós értékű függvény a gradiensével ellentétes irányban csökken a leggyorsabban, gradφ gradiense pedig éppen a maradékvektor (-1)-szerese, így a legjobb ötletnek az tűnik, hogy mindig az aktuális r maradékvektor irányában keressük a minimumhelyet. Tehát egy x pontból az

$$x + \frac{(r,r)}{(r,Ar)}r$$

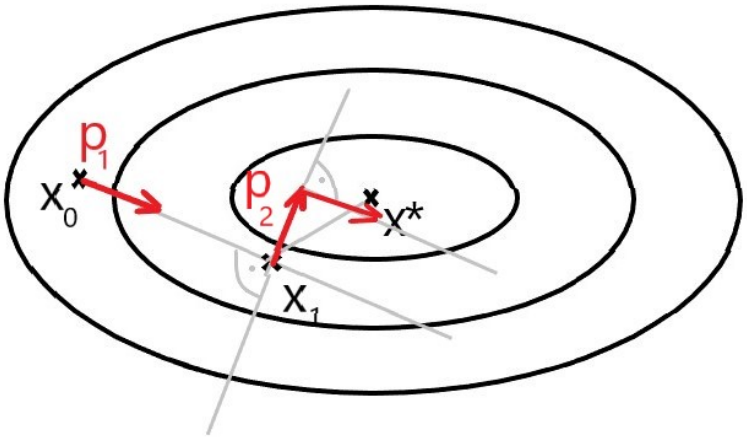
pontba lépünk. Ebben a pontban a maradékvektor

$$b - A(x + \frac{(r,r)}{(r,Ar)}r),$$

így ennek irányában keressük a következő lépésben a minimumhelyet. Vegyük észre, hogy

$$(r, b - A(x + \frac{(r,r)}{(r,Ar)}r)) = (r,r) - (r, A\frac{(r,r)}{(r,Ar)}r) = 0,$$

vagyis az egymás utáni keresési irányok merőlegesek egymásra.



Ez a módszer lassan konvergál, ha az A mátrix euklideszi normához tartozó $\text{cond}_2(A)$ kondíciószáma nagy. A következőkben bemutatjuk a gradiensmódszer olyan módosított változatát, amely a keresési irányok ügyes megválasztása révén gyorsabb konvergenciát biztosít.

A konjugált gradiens-módszer

Tekintsünk egy kétismeretlenes lineáris algebrai egyenletrendszert! Induljunk ki abból, hogy az x_0-beli keresési irány merőleges az x_1-beli maradékvektorra. Ekkor

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

Ebből a képletből látható, hogy az x_1 pontból az x^* megoldásba vezető vektor teljesíti a $(p, A(x^* - x_1)) = 0$ feltételt. Az egyszerűbb megfogalmazás kedvéért vezessük be az alábbi fogalmat.

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

Mivel az a legkedvezőbb, ha az adott iterációs lépésben éppen az x^* megoldás felé indulunk el, így a p_2 keresési irányt úgy kellene megválasztani, hogy az A-ortogonális legyen a p_1 vektorra. Mint látni fogjuk, ehhez nem kell ismernünk x^*-ot. Keressük ugyanis a második keresési irányt

$$p_2 = r_1 - \beta_1 p_1$$

alakban. Hogyan válasszuk meg β_1-et ahhoz, hogy $(p_1, Ap_2) = 0$ legyen?

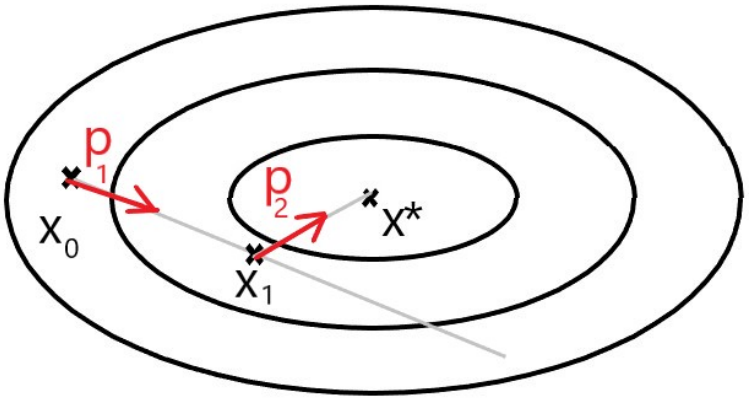
$$(p_1, A(r_1 - \beta_1 p_1)) = 0$$

$$(p_1, Ar_1) - \beta_1 (p_1, Ap_1) = 0$$

Ebből

$$\beta_1 = \frac{(p_1, Ar_1)}{(p_1, Ap_1)}.$$

Ezen β_1 megválasztással a $p_2 = r_1 - \beta_1 p_1$ keresési irányban keresve a minimumhelyet, az x^* megoldást kapjuk. Tehát a kétdimenziós esetben a konjugált gradiens-módszer két iterációs lépésben megadja a megoldást.



Megmutatható, hogy m-ismeretlenes lineáris algebrai egyenletrendszer esetén legfeljebb m lépésben véget ér a módszer. Ez azt jelenti, hogy a konjugált gradiens módszer lényegében direkt módszernek tekinthető. A gyakorlatban ugyanakkor a hibával terhelt számítások miatt iterációs módszerként alkalmazzuk. Kis kondíciószámú és kevés különböző sajátértékkel rendelkező A mátrixok esetén a módszer gyorsan konvergál. Előnyös, hogy nem kell optimális relaxációs paramétert keresni a gyors konvergenciához.

3. fejezet

Algebrai egyenletek közelítő megoldása

Az egyismeretlenes valós egyenletek egy része az egyenlet rendezésével könnyen megoldható. A legegyszerűbbek a lineáris egyenletek, de a másodfokúakat is gond nélkül meg tudjuk oldani a jól ismert megoldóképlet segítségével. A harmad- és negyedfokú egyenletekre is létezik megoldóképlet, de már jóval bonyolultabb. Magasabbfokú egyenletekre pedig már bizonyíthatóan nem lehet általános képletet felállítani. A logaritmikus, trigonometrikus és egyéb egyenletek megoldása esetenként szintén problémát jelenthet.

Világos, hogy minden egyismeretlenes valós egyenlet a tagok bal oldalra rendezésével felírható

$$f(x) = 0 \tag{3.1}$$

alakban, ahol $f : \mathbb{R} \to \mathbb{R}$ valamilyen függvény. A továbbiakban az ilyen alakú egyenletek megoldásával foglalkozunk. Ezen felírás mellett az egyenlet megoldása egyben az f függvény **zérushelye**, vagy más szóval **gyöke**.

A megoldandó egyenletről sokszor belátható, hogy létezik gyöke. Ha $f \in C[a, b]$, és $f(a) \cdot f(b) < 0$ (azaz f az a és a b pontban ellentétes előjelű), akkor Bolzano tétele értelmében (a, b)-ben található legalább egy olyan x^* pont, ahol $f(x^*) = 0$. Továbbá, a gyökök száma páratlan, és ha f szigorúan monoton, akkor egyetlen gyök van csak. Ugyanakkor $f(a) \cdot f(b) > 0$ esetén vagy nincs gyök, vagy páros számú gyök van. Mindenesetre, ha létezik gyök, akkor kijelölhető egy $[a, b]$ intervallum úgy, hogy csak egy gyök legyen rajta.

3.1. A gyökök stabilitása

Alkalmazásokban előfordulhat, hogy (3.1) helyett az

$$\tilde{f}(x) = 0 \tag{3.2}$$

egyenletet oldjuk meg, ahol az $\tilde{f}$ függvény nem egyezik pontosan meg az f függvénnyel. Vajon ez hogyan hat ki a gyökök értékére? Könnyen tudunk olyan példát adni, ahol már a gyökök száma is más lesz! Pl. az

$$f(x) = e^x - 0,01$$

függvénynek létezik egyetlen gyöke, és az $x^* = \ln 0,01$. Az

$$\tilde{f}(x) = e^x$$

függvénynek azonban nincs egy gyöke sem!
Az

$$f(x) = x^2 - 2x + 1$$

függvénynek egy kétszeres gyöke van: $x_1^* = x_2^* = 1$. A hozzá közel álló

$$f(x) = x^2 - 2x + 1,01$$

függvénynek ugyanakkor nincs valós gyöke, $x_{1,2}^* = 1 \pm 0,1i$. Megint egy kicsit változtatva a bal oldali függvényen, legyen most

$$f(x) = x^2 - 2x + 0,99,$$

ekkor két különböző valós gyököt kapunk.

Olyan példát is mutatunk, ahol a gyökök értéke jelentősen megváltozik. Cseréljük ki az előző példában x-et $x^{\frac{1}{6}}$-ra, ezzel az

$$f(x) = x^{\frac{1}{3}} - 2x^{\frac{1}{6}} + 1$$

függvényt kapjuk, amelyre $x_1^* = x_2^* = 1$. Ugyanakkor a tőle alig különböző

$$f(x) = x^{\frac{1}{3}} - 2x^{\frac{1}{6}} + 0,99$$

függvényre már $x_1^* = 1,771561$ és $x_2^* = 0,531441$, ami jelentős eltérés.

Adjunk feltételt arra, hogy ha a (3.1) és (3.2) egyenletnek létezik egyetlen gyöke, x^* és $\tilde{x}^*$, valamint f és $\tilde{f}$ közel van egymáshoz, akkor x^* és $\tilde{x}^*$ eltérése kicsi. Azaz

$$f(x^*) = 0;\ \tilde{f}(\tilde{x}^*) = 0,\ \max|f(x) - \tilde{f}(x)| < \varepsilon \Rightarrow |x^* - \tilde{x}^*| \text{ kicsi.}$$

Legyen $f \in C[a,b] \cap D(a,b)$, $x^*, \tilde{x}^* \in (a,b)$. Ekkor a Lagrange-középértéktétel[1] értelmében létezik olyan $\vartheta \in (x^*, \tilde{x}^*)$ amelyre

$$f(\tilde{x}^*) = f(x^*) + f'(\vartheta)(\tilde{x}^* - x^*).$$

Ha $f' \neq 0$ az $(x^*, \tilde{x}^*)$ intervallumon, akkor $f(x^*) = 0$ és $\tilde{f}(\tilde{x}^*) = 0$ miatt

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

3.1.1 Definíció. *Az*

$$M := \frac{1}{\min_{[a,b]} |f'|} \tag{3.3}$$

számot a (3.1) egyenlet kondicionáltsági számának nevezzük.

Így tehát $|\tilde{x}^* - x^*| \leq M \cdot \varepsilon$. Ha a kondicionáltsági szám nagy, akkor az egyenlet megoldása érzékeny a bal oldali függvény hibájára, és az egyenletet rosszul kondicionáltnak nevezzük.

A korábbi példánkban, ahol $f(x) = x^{\frac{1}{3}} - 2x^{\frac{1}{6}} + 1$, a deriváltfüggvény $f'(x) = \frac{1}{3}(x^{-\frac{4}{6}} - x^{-\frac{5}{6}})$, így

$$\min_{[0,1]} |f'| = 0 \Rightarrow M = +\infty,$$

vagyis az egyenlet rosszul kondicionált volt.

3.2. Konvergenciasebesség

A továbbiakban olyan módszerekről lesz szó a (3.1) feladat megoldására, amelyek lényege, hogy felépítünk egy, a gyökhöz konvergáló sorozatot (ún. iterációs módszerek). Fontos szempont lesz az, hogy a felépített számsorozat milyen gyorsan közelíti meg a határértékét, az egyenlet gyökét. Ezért mindenekelőtt arról tanulunk, hogy hogyan lehet egy sorozat konvergenciasebességét jellemezni.

Legyen (x_k) egy konvergens számsorozat, amelyre $\lim x_k = x^*$, és vezessük be az $e_k := x_k - x^*$ jelölést. (Fogalmaink tetszőleges normált térbeli (x_k) sorozatra is értelmesek, ezért mindenhol a norma jelet használjuk, valós sorozat esetén ez természetesen az abszolút értéket jelenti.)

3.2.1 Definíció. *Azt mondjuk, hogy az $x_k \to x^*$ sorozat konvergenciarendje $p \geq 1$, ha*

[1] A Lagrange-középértéktétel a következőképpen szól: Legyen $f \in C[a,b] \cap D(a,b)$. Ekkor létezik olyan $\vartheta \in (a,b)$ pont, amelyben f érintője párhuzamos az $(a, f(a))$ és $(b, f(b))$ pontokat összekötő húrral. Itt a Lagrange-tételt az $[x^*, \tilde{x}^*]$ intervallumon alkalmaztuk, amelyen szintén teljesülnek a tétel feltételei.

létezik a

$$\lim_{k\to\infty} \frac{\ln \|e_k\|}{\ln \|e_{k-1}\|}$$

határérték, és az egyenlő p-vel.

Ezen p rend a hiba lépésről lépésre történő logaritmikus relatív csökkenésének a mértékét fejezi ki. Minél nagyobb ez a p rend, annál jobban csökken a hiba. Ha $p = 1$, akkor lineáris, ha $p = 2$, akkor pedig kvadratikus konvergenciáról beszélünk. Kérdés: hogyan lehet a konvergenciarendet meghatározni?

3.2.2 Állítás. *Ha az (x_k) sorozatra és x^* számra teljesül az*

$$\|e_k\| = C_k \cdot \|e_{k-1}\| \quad k = 1, 2, \ldots \tag{3.4}$$

feltétel valamilyen $0 < \underline{C} \le C_k \le \overline{C} < 1$ konstansokkal, akkor (x_k) monoton módon és elsőrendben tart x^*-hoz.

Biz.: A monotonitás nyilvánvaló abból, hogy $0 < C_k < 1$ minden k-ra.

$(3.4) \Rightarrow \|e_k\| = C_k\|e_{k-1}\| \le \overline{C}\|e_{k-1}\| \le \ldots \le \overline{C}^k\|e_0\|$, és mivel $\overline{C} < 1$, ezért $\lim_{k\to\infty} \|e_k\| = 0$.

A (3.4) egyenlőség logaritmusát véve

$$\ln \|e_k\| = \ln C_k + \ln \|e_{k-1}\|.$$

Innen az $\ln \|e_{k-1}\| < 0$ tényezővel osztva (feltéve, hogy k elég nagy ahhoz, hogy a hiba már kisebb legyen, mint 1),

$$\frac{\ln \|e_k\|}{\ln \|e_{k-1}\|} = 1 + \frac{\ln C_k}{\ln \|e_{k-1}\|} \tag{3.5}$$

A $0 < \underline{C} \le C_k \le \overline{C} < 1$ feltétel miatt $\ln \underline{C} \le \ln C_k < 0$ (így $\ln C_k$ korlátos), és a konvergencia miatt $\|e_{k-1}\| \to 0$, azaz $\ln \|e_{k-1}\| \to -\infty$. Így (3.5) jobb oldala $k \to \infty$ esetén 1-hez tart, azaz a konvergenciarend valóban $p = 1$.

3.2.3 Állítás. *Ha*

$$\|e_k\| = C_k \cdot \|e_{k-1}\|^p \quad k = 1, 2, \ldots, \tag{3.6}$$

ahol $p > 1$ és $0 < \underline{C} \le C_k \le \overline{C} < +\infty$, és

$$\overline{C}^{\frac{1}{p-1}} \|e_0\| < 1, \tag{3.7}$$

akkor (x_k) konvergens, és a sorozat konvergenciarendje p.

Biz.: Legyen $\mathcal{E}_k := \overline{C}^{\frac{1}{p-1}} e_k$. Ekkor

$$\|\mathcal{E}_k\| = \overline{C}^{\frac{1}{p-1}} \|e_k\| \leq \overline{C}^{\frac{p}{p-1}} \|e_{k-1}\|^p = [\overline{C}^{\frac{1}{p-1}} \|e_{k-1}\|]^p = \|\mathcal{E}_{k-1}\|^p.$$

(3.6) $\Rightarrow$

$$\|\mathcal{E}_k\| \leq \|\mathcal{E}_0\|^{p^k}. \tag{3.8}$$

Mivel $\|\mathcal{E}_0\| = \overline{C}^{\frac{1}{p-1}} \|e_0\| < 1$, ezért (3.8) $\Rightarrow \lim \|\mathcal{E}_k\| = 0$, azaz $\lim \|e_k\| = 0$, tehát konvergens a sorozat. A konvergencia rendje:

$$\ln \|e_k\| = \ln C_k + p \ln \|e_{k-1}\|$$

$$\Rightarrow \frac{\ln \|e_k\|}{\ln \|e_{k-1}\|} = \frac{\ln C_k}{\ln \|e_{k-1}\|} + p \to p, \text{ ha } k \to \infty.$$

3.2.4 Megjegyzés. *A p-ed rendűség azt jelenti, hogy minden lépésben a pontosság nagyságrendje p-szeresre növekszik. (Tehát, ha $p = 2$, és egy közelítés hibája 10^{-3}, akkor a következő lépésben 10^{-6} a pontosság, a rákövetkezőben 10^{-12}, stb.)*

3.2.5 Megjegyzés. *Az állításból leolvasható, hogy $p > 1$ esetén a konvergencia csak akkor következik (3.6)-ból, ha (3.7) igaz, azaz, kellően közel választjuk a kezdeti közelítést. (Elsőrendű ($p = 1$) esetben nincs ilyen feltétel, viszont ott feltettük, hogy $\overline{C} < 1$.)*

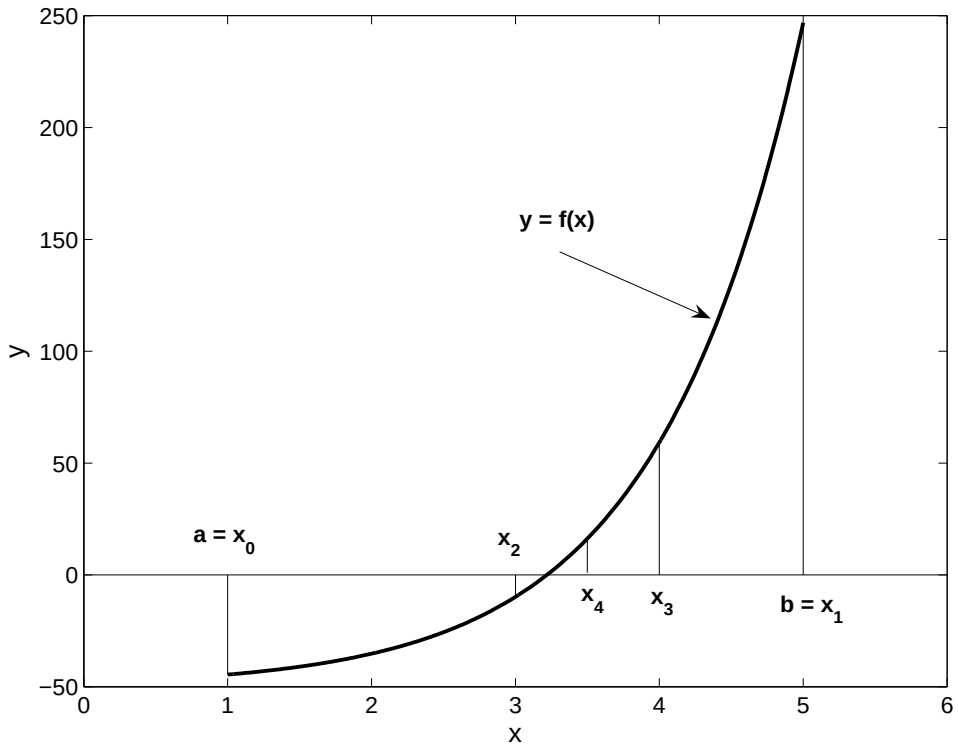
A következőkben néhány konkrét iterációs módszert ismertetünk algebrai egyenletek megoldására.

3.3. Intervallumfelezés

Legyen f folytonos az $[a, b]$ intervallumon, és tegyük fel, hogy $f(a) \cdot f(b) < 0$. Ekkor Bolzano tétele értelmében (a, b)-ben található legalább egy olyan x^* pont, ahol $f(x^*) = 0$.

Felezzük meg az $[a, b]$ intervallumot. Ha f értéke a felezőpontban nulla, akkor f valamelyik gyökét már meg is találtuk. Ha nem nulla, akkor válasszuk ki a két részintervallum közül azt, amelyiknek a végpontjaiban az f előjele ellentétes, és az eljárást ismételjük meg erre a részintervallumra, és így tovább. A megtartott intervallumok metszete egyetlen számot tartalmaz, és ez az f függvény valamelyik gyöke (lásd a 3.1. ábrát).

Az eljárás vagy véges sok lépésben befejeződik (amikor valamelyik felezőpontban az f értéke pontosan zérus), vagy elég sok lépés után tetszőlegesen kis intervallumba szorítjuk



3.1. ábra. Intervallumfelezés.

f valamelyik gyökét. A k-adik lépésben nyert intervallum hossza

$$\frac{b-a}{2^k},$$

így a részintervallum bármelyik végpontja adott $\varepsilon > 0$ hibakorlátnál pontosabban fogja közelíteni f $[a,b]$-beli egyik gyökét, ha

$$\frac{b-a}{2^k} < \varepsilon,$$

vagyis ha $k > k^*$, ahol

$$k^* = \frac{\ln \frac{b-a}{\varepsilon}}{\ln 2}.$$

A hiba felső becsléseinek sorozata elsőrendben konvergens, ugyanis

$$|e_k| \leq \tilde{e}_k := 2^{-k}(b-a)$$

$$|e_{k-1}| \leq \tilde{e}_{k-1} = 2^{-(k-1)}(b-a)$$

$$\frac{\ln \tilde{e}_k}{\ln \tilde{e}_{k-1}} = \frac{-k\ln 2 + \ln(b-a)}{-(k-1)\ln 2 + \ln(b-a)} \to 1 \ (k \to \infty).$$

3.4. Egyszerű iteráció

Ez a módszer a fixpont-tételen alapul. Ehhez a megoldandó $f(x) = 0$ egyenletet először $\varphi(x) = x$ alakra hozzuk. Ekkor a megoldás φ fixpontja.

Érvényes a fixpont-tétel következő változata:

3.4.1 Tétel. *Legyen $H \subset \mathbb{R}$ zárt halmaz, és $\varphi : H \to H$ kontrakció (azaz létezik olyan $q \in [0, 1)$ szám, amely mellett $|\varphi(x) - \varphi(y)| \leq q|x - y|$ $\forall x, y \in H$). Ekkor*

- *létezik egyetlen x^* fixpontja φ-nek;*

- *$x_0 \in H$ tetszőleges esetén az*

$$x_{k+1} = \varphi(x_k), \;\; k = 0, 1, \ldots$$

  rekurzióval definiált sorozat konvergens és $x_k \to x^*$, továbbá

- *a k-edik lépésben kapott érték hibájára igaz:*

$$|x_k - x^*| \leq \frac{q^k}{1 - q}|x_1 - x_0|.$$

Vegyük észre, hogy $\varphi : H \to H$, tehát feltettük, hogy φ beleképez H-ba.

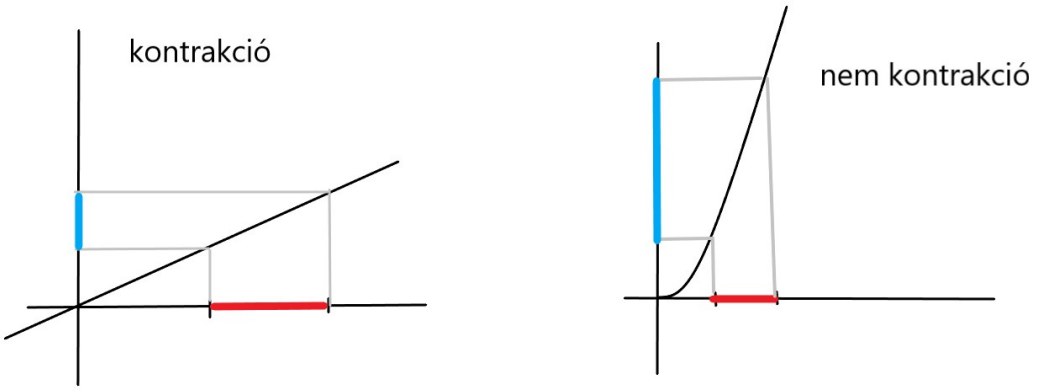
Mindenekelőtt tudunk kell, hogyan dönthető el egy függvényről, hogy kontrakció-e. A kontrakció zsugorítást jelent, a definíciója durván azt fejezi ki, hogy bármely két pontot választva a $H \subset D(\varphi)$ halmazból, a képpontok közelebb vannak egymáshoz mint az alappontok. Tekintsük pl. az $\varphi(x) = \frac{1}{2}x$ függvényt! Kontrakció-e $\mathbb{R}$-en?

$$|\varphi(x) - \varphi(y)| = \left|\frac{1}{2}x - \frac{1}{2}y\right| = \frac{1}{2}|x - y| \;\forall x, y \in \mathbb{R}.$$

Az utolsó egyenlőségjel helyett $\leq$ is írható. Tehát kontrakció $q = \frac{1}{2}$ mellett. Ugyanakkor a $\varphi(x) = x^2$ függvény a $H = [1, 2]$ halmazon nem kontrakció, ugyanis legyen pl. $x = 2$, $y = 1$. Ekkor

$$|\varphi(x) - \varphi(y)| = 3 \leq q \cdot |2 - 1|$$

semmilyen $q \in [0, 1)$ esetén nem áll fenn.



Vajon adhatunk-e általános módszert a kontrakció eldöntésére? A kontrakció jelentése alapján sejthetjük, hogy amennyiben differenciálható a függvény, akkor a meredeksége számíthat. Az $\frac{1}{2}x$ függvény grafikonja elég lapos, az x^2-é viszont túl meredek a kérdéses szakaszon. A választóvonal a meredekség 1-es értékénél húzódik.

3.4.2 Tétel. *Tegyük fel, hogy φ folytonos az I intervallumon, differenciálható I belső pontjaiban, és létezik $q \in [0,1)$, amelyre $|\varphi'(x)| \leq q$ I belső pontjaiban. Ekkor φ kontrakció I-n q kontrakciószámmal.*

Biz.: Tetszőleges $x, y \in I$ $(x < y)$ esetén alkalmazzuk a Lagrange-középértéktételt:

$$\exists \xi \in (x,y) : \frac{\varphi(y) - \varphi(x)}{y - x} = \varphi'(\xi).$$

Átrendezve és abszolút értéket véve

$$|\varphi(x) - \varphi(y)| = |\varphi'(\xi)| \cdot |x - y| \leq q \cdot |x - y|.$$

Pl. a $\varphi(x) = \frac{1}{2}\cos x$ függvény kontrakció-e az $[0, \frac{\pi}{2}]$ intervallumon? $q = ?$

$$|\varphi'(x)| = \left| -\frac{1}{2}\sin x \right| \leq \frac{1}{2} \quad \forall x \in [0, \frac{\pi}{2}].$$

Így $q = \frac{1}{2}$ mellett φ kontrakció a $[0, \frac{\pi}{2}]$ intervallumon.

3.4.3 Megjegyzés. *Nyilvánvalóan, ha $I = [a,b]$ (zárt), és $\varphi \in C^1[a,b]$ (φ' folytonos $[a,b]$-n), akkor a kontrakcióhoz elegendő, hogy $|\varphi'(x)| < 1$ legyen minden $x \in [a,b]$ pontban, hiszen ekkor létezik $\max_{x \in [a,b]} |\varphi(x)| =: q < 1$.)*

A következőkben az egyszerű iteráció konvergenciarendjét tanulmányozzuk. Mint látni fogjuk, a konvergencia általában elsőrendű.

3.4.4 Állítás. *Tegyük fel, hogy* $\varphi : [a,b] \to [a,b]$, $\varphi \in C^1[a,b]$, $|\varphi'(x)| < 1\ \forall x \in [a,b]$, *továbbá* $\varphi(x^*) \neq 0$. *Ekkor az egyszerű iteráció elsőrendben konvergens tetszőleges* $x_0 \in [a,b]$ *kezdőértékből indítva.*

Biz.: A feltételek teljesülése esetén φ kontrakció valamely $q \in [0,1)$ kontrakciószámmal, így a fixponttétel értelmében az iteráció konvergens, és elég az elsőrendet belátnunk.

A Lagrange-középértéktétel értelmében $\exists \vartheta_k$ az x_k és x^* között, amelyre

$$\varphi(x_k) - \varphi(x^*) = \varphi'(\vartheta_k)(x_k - x^*).$$

A bal oldalon $\varphi(x_k) = x_{k+1}$ és $\varphi(x^*) = x^*$, így

$$|x_{k+1} - x^*| = |\varphi'(\vartheta_k)| \cdot |x_k - x^*|.$$

Tehát a $c_k := |\varphi'(\vartheta_k)|$ jelöléssel az

$$|e_{k+1}| = C_k |e_k|,\ k = 0, 1, \ldots$$

összefüggésre jutottunk. A 3.2.2 állítás szerint itt még belátandó, hogy C_k közrefogható egy $\underline{C} > 0$ és $\overline{C} < 1$ konstanssal. Nyilván $\overline{C}$ felső korlátnak választható a q kontrakciószám. A pozitív alsó korlát megléte pedig az alábbi módon igazolható: Mivel az $x \mapsto |\varphi'(x)|$ függvény folytonos $[a,b]$-n, és $\varphi'(x^*) \neq 0$, így az x^* körüli valamely $J := [x^* - \delta, x^* + \delta]$ intervallumra $|\varphi'(x)| > 0\ \forall x \in J$. Továbbá, $x_k \to x^*$ miatt elég nagy k indekre $x_k \in J$, így $\vartheta_k \in J$. Ezért pozitív alsó becslésnek választható a $\underline{C} := \min_{x \in J} |\varphi'(x)|$ konstans. Kérdés, mi történik, ha az x^* pontban a φ kontrakció nullát vesz fel. Ekkor, ha elég sima ez a függvény, magasabb rend is elérhető a következő tétel szerint.

3.4.5 Állítás. *Legyen* $\varphi \in C^p[a,b]$ *(ahol* $p \geq 2$*) egy* $[a,b]$*-be képező kontrakció. Ha az* x^* *fixpontban*

$$\varphi'(x^*) = \ldots = \varphi^{(p-1)}(x^*) = 0 \text{ és } \varphi^{(p)}(x^*) \neq 0,$$

akkor a fixpont-iteráció p*-ed rendben konvergens tetszőleges* $x_0 \in [a,b]$ *pontból indítva.*

Biz.: A konvergencia ténye következik a fixponttételből, így csak a p-ed rendet kell igazolni. Fejtsük Taylor-sorba φ-t x^* körül:

$$\varphi(x_k) = \varphi(x^*) + \frac{\varphi^{(p)}(\vartheta_k)}{p!}(x_k - x^*)^p$$

$$\varphi(x_k) - x^* = \frac{\varphi^{(p)}(\vartheta_k)}{p!}(x_k - x^*)^p.$$

A bal oldalon $\varphi(x_k) = x_{k+1}$, így a $C_k := |\frac{\varphi^{(p)}(\vartheta_k)}{p!}|$ megválasztással

$$|x_{k+1} - x^*| = C_k|x_k - x^*|^p,$$

azaz

$$|e_{k+1}| = C_k|e_k|^p,$$

Mivel $\varphi \in C^p[a,b]$, és így $\varphi^{(p)}$ folytonos $[a,b]$-n, továbbá $\varphi^{(p)}$ az x^* pontban nem nulla, így x^* egy környezetében $|\varphi^{(p)}|$ pozitív. Ezért a fenti C_k szorzó elég nagy k-ra beszorítható valamely $\underline{C}$ és $\overline{C}$ pozitív konstansok közé. Ezzel beláttuk, hogy a konvergencia rendje p. Vagyis a rendet az határozza meg, hogy φ-nek hányadik az első olyan deriváltja, amely már nem nulla az x^* pontban.

3.5. Newton-módszer (érintőmódszer)

Az előző módszer alkalmazásához az x^* gyökhelyet először be kellett szorítanunk egy intervallumba. Newton módszerénél elég egyetlen elegendően pontos x_0 közelítést ismernünk. Az f függvénynek azonban deriválhatónak kell lennie, sőt, mint látni fogjuk, a módszer konvergenciájának bizonyításához a második derivált létezését is feltesszük.

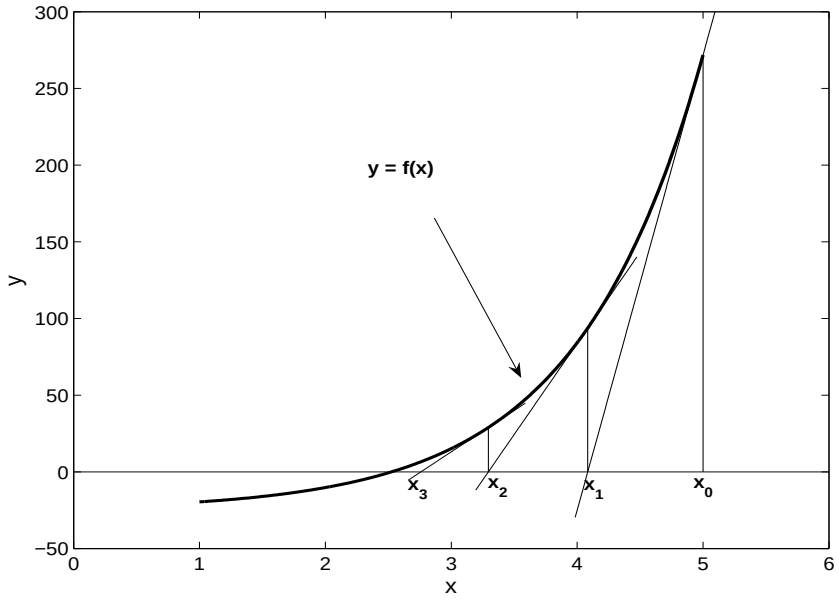
A módszer geometriailag a következőképpen interpretálható. Húzzuk meg az f függvény x_0-beli érintőjét, és jelölje x_1 az érintő x tengellyel alkotott metszéspontját. Az x_1 pontban megint behúzzuk f érintőjét, és annak x-tengellyel alkotott metszéspontja lesz x_2, és így tovább. Az eljárás folytatásával egy $x_0, x_1, x_2, \ldots$ sorozatot nyerünk, lásd a 3.2. ábrát. Írjuk fel a módszer formuláját! Az f függvény x_k-pontbeli érintőjének egyenlete

$$y = f'(x_k)(x - x_k) + f(x_k), \tag{3.9}$$

amelyből $y = 0$ helyettesítéssel kapjuk a Newton-módszer formuláját:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \tag{3.10}$$

Vizsgáljuk meg a módszer konvergenciáját! Tegyük fel, hogy az x^* megoldást és az (x_k) sorozatot tartalmazó valamely I intervallumban f kétszer folytonosan differenciálható, továbbá léteznek olyan m_1, M_1, m_2 és M_2 pozitív konstansok, amelyekkel $0 < m_1 \leq |f'(x)| \leq M_1 < \infty$ és $0 < m_2 \leq |f''(x)| \leq M_2 < \infty$ minden $x \in I$ pontban. A Taylor-



3.2. ábra. A Newton-módszer.

formula értelmében van olyan ϑ_k pont az x_k és az x^* között, amelyre

$$0 = f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{f''(\vartheta_k)}{2}(x^* - x_k)^2, \tag{3.11}$$

Innen, mivel f' sehol sem nulla, az

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

összefüggés adódik. A módszer (3.10) képletét felhasználva

$$x_{k+1} - x^* = \frac{f''(\vartheta_k)}{2f'(x_k)}(x_k - x^*)^2. \tag{3.12}$$

Így a (3.6) egyenlőség $C_k = |\frac{f''(\vartheta_k)}{2f'(x_k)}|$ állandóval és $p = 2$ értékkel fennáll, tehát

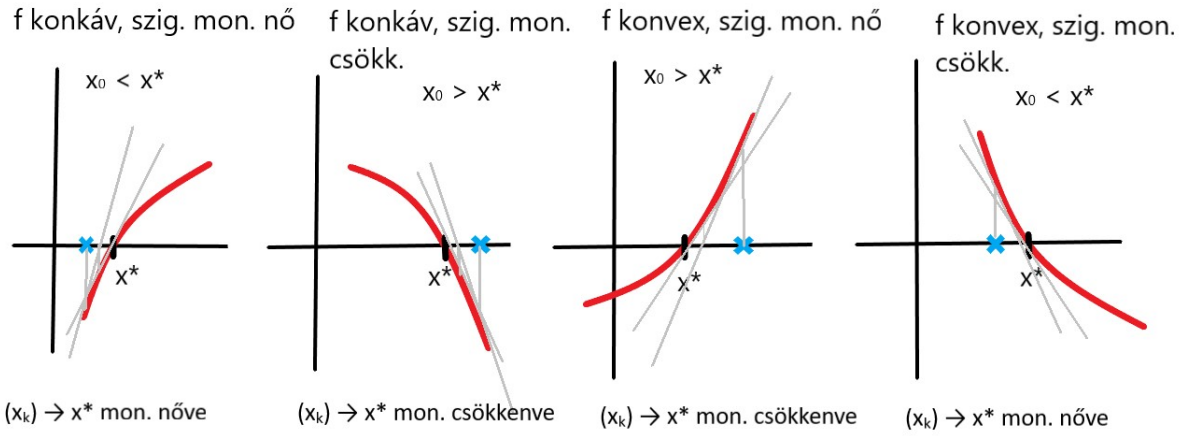
$$\overline{C} := \frac{M_2}{2m_1}$$

megválasztással az erre vonatkozó állítás alkalmazható, vagyis ha az $e_0 = x_0 - x^*$ kezdeti hibára

$$\frac{M_2}{2m_1}|e_0| < 1$$

teljesül, akkor az iteráció másodrendben konvergens.

3.5.1 Megjegyzés. *Kérdés, hogy mikor garantálható, hogy az egész (x_k) sorozat benne van abban az I intervallumban, amelyen f-re teljesülnek az előírt tulajdonságok. Belátható, hogy ha f kétszer folytonosan differenciálható a gyököt tartalmazó valamely I intervallumban, akkor f megfelelő alaki tulajdonsága (szigorú monotonitás és állandó konvexitás) esetén a sorozat monoton módon konvergál a gyökhöz, ha a kezdőpontot a gyök megfelelő oldalán vesszük fel az I intervallumban (lásd a kékkel jelölt pontot az alábbi ábrákon). Ilyenkor tehát az (x_k) sorozat minden eleme a kezdőpont és a gyök közé fog esni.*



3.5.1. Nemlineáris algebrai egyenletrendszerek megoldása

Eddig egyismeretlenes valós egyenletek megoldásával foglalkoztunk. Nézzük meg most a rendszereket! Ekkor az alapfeladat

$$\underline{f}(\underline{x}) = \underline{0},$$

ahol az $\underline{f}$ függvény $\mathbb{R}^m \to \mathbb{R}^m$ típusú, és feltesszük, hogy kellően sima. Ha ez az $\underline{f}$ függvény lineáris, akkor ez egy lineáris algebrai egyenletrendszer (az ilyenek megoldásával már részletesen foglalkoztunk), egyébként nemlineáris. Itt két módszert ismertetünk a feladat megoldására.

1. **Egyszerű iteráció.** Általában ezt használják. Lényege, hogy az $\underline{f}(\underline{x}) = \underline{0}$ egyenletet $\underline{F}(\underline{x}) = \underline{x}$ alakra hozzuk, és elindítjuk a fixponttétel szerinti

$$\underline{x}_{k+1} = \underline{F}(\underline{x}_k)$$

iterációt. Sajnos nem mindig könnyű megvizsgálni, hogy $\underline{F}$-re teljesülnek-e a fixponttétel feltételei, de ha a fenti iterációval előállított sorozat konvergál egy $\underline{x}^*$ vektorhoz, akkor $\underline{x}^* = \underline{F}(\underline{x}^*)$ miatt $\underline{f}(\underline{x}^*) = \underline{0}$ is teljesül (mivel $\underline{F}$ folytonos).

2. **Newton-módszer.** Idézzük fel a módszer képletét a skaláris esetben:

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

Ennek analógiájára felírható az

$$\underline{x}_{k+1} = \underline{x}_k - (J(\underline{x}_k))^{-1} \cdot \underline{f}(\underline{x}_k)$$

iteráció, ahol J az $\underline{f}$ függvény Jacobi-mátrixa (deriváltmátrixa). Ez a módszer megfelelő feltételek esetén másodrendben konvergens, ám viszonylag költséges. A számításigény csökkenthető, ha minden egyes iterációs lépésben a $J(\underline{x}_0)$ márixszal számolunk, bár ez csak elsőrendű konvergenciát biztosít.

4. fejezet

Interpolációs feladatok

A gyakorlatban sűrűn előfordul, hogy egy folytonos függvényt nem ismerünk a teljes értelmezési tartományán, csak néhány $x_0, x_1, \ldots, x_n$ alappontban felvett $f_0, f_1, \ldots, f_n$ értékét tudjuk. Ez a helyzet például akkor, ha az adatok diszkrét pontokban végzett mérésekből származnak. A matematikai analízis módszereit ezekre a diszkrét pontokban adott függvényekre közvetlenül nem tudjuk alkalmazni. Ezért a pontokra először is folytonos függvényt illesztünk. Az illesztésre leggyakrabban **polinom**okat használunk, ezek ugyanis a legkönnyebben kezelhető és a legkedvezőbb analitikus tulajdonságokat követő függvények. Előnyeik:

- A legfeljebb n-ed fokú polinomok halmaza (P_n) zárt az összeadásra és a skalárral való szorzásra (vektortér).
- A polinomokat könnyű deriválni ill. integrálni.
- Helyettesítési értékeik egyszerűen kiszámíthatók (Horner-szabály).
- Gyökeik jól meghatározhatók.
- Weierstrass tétele értelmében folytonos függvények jól közelíthetők polinomokkal.

4.1. Az interpolációs polinom

Tegyük fel, hogy az $f : \mathbb{R} \to \mathbb{R}$ függvényt csak az $x_0, x_1, \ldots, x_n$ pontokban ismerjük. Jelölje f_k az $f(x_k)$ függvényértéket, $k = 0, 1, \ldots, n$. Olyan p polinomot keresünk, amelynek fokszáma legfeljebb n, és teljesül, hogy $p(x_k) = f_k,\ k = 0, 1, \ldots, n$. Az x_k pontokat **interpolációs alappont**oknak, az f_k értékeket **interpolált érték**eknek, p-t pedig **interpolációs polinom**nak nevezzük. Az interpolációs polinom létezését és egyértelműségét biztosítja az alábbi tétel.

4.1.1 Tétel. *Pontosan egy olyan polinom létezik, amelynek fokszáma legfeljebb n, és az $n+1$ különböző pontban az f_k értékeket veszi fel.*

Biz.: a) Először a létezést bizonyítjuk a képlet bemutatásával.
Keressünk először olyan legfeljebb n-ed fokú polinomot, amely az x_m alappontban 1-et, a többi alappontban pedig nullát vesz fel. Jelölje ezt l_m.

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

Mivel az $x_0, x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_n$ pontokban el kell tűnnie, tartalmaznia kell az $x - x_k$ $(k \neq m)$ gyöktényezőket. Ez azt jelenti, hogy a keresett polinom

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

alakú, ahol c_m állandó. Ezt az állandót az $l_m(x_m) = 1$ feltételből határozhatjuk meg:

$$c_m = \frac{1}{\prod\limits_{\substack{k=0 \\ k \neq m}}^{n} (x_m - x_k)}.$$

Így az

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

alakhoz jutunk. Vegyük észre, hogy ha mindegyik alapponthoz felírjuk a megfelelő l_m függvényt $(m = 0, 1, \ldots, n)$, akkor ezekből lineáris kombinációval összerakhatunk egy olyan $p \in P_n$ polinomot, amely mindegyik pontban az előírt interpolált értéket veszi fel, nevezetesen a következő módon:

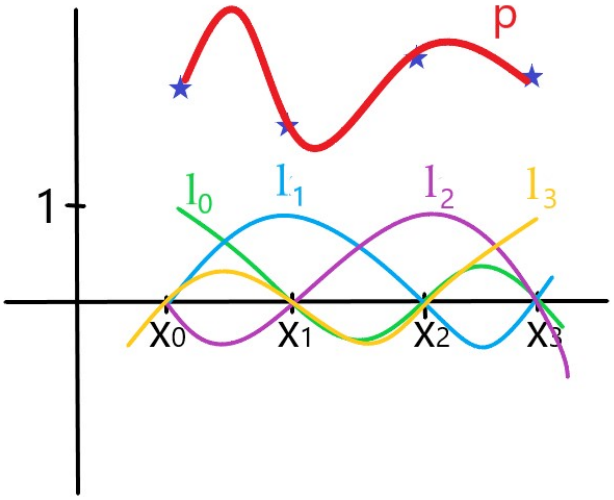
$$p(x) := \sum_{m=0}^{n} l_m(x) f_m.$$

Lássuk is be, hogy p az x_k pontokban az f_k értékeket veszi fel: az x_k-t behelyettesítve $p(x)$ képletébe

$$p(x_k) = \sum_{m=0}^{n} l_m(x_k) f_m.$$

Ebben az összegben csak az $m = k$ indexű tag nem nulla, és az éppen f_k-val egyenlő. Tehát

$$p(x_k) = f_k, \quad k = 0, 1, \ldots, n.$$



b) Az egyértelműséget indirekt módon bizonyítjuk. Tegyük fel, hogy két polinom is rendelkezik a kívánt tulajdonságokkal, azaz

$$p(x_k) = f_k \quad \text{és} \quad q(x_k) = f_k, \ k = 0, 1, \ldots, n.$$

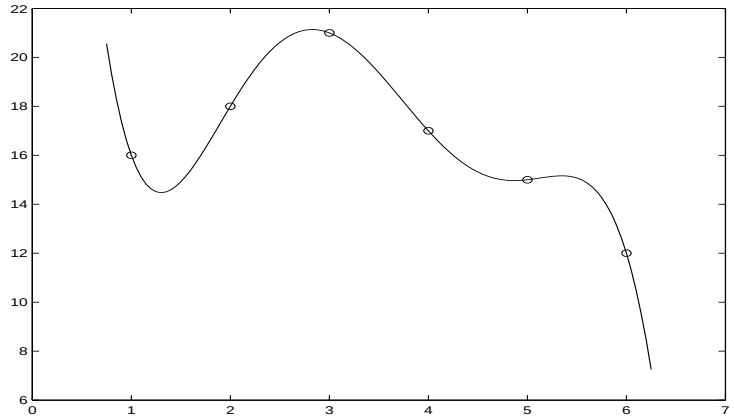
Jelölje d a $p - q$ különbséget. Ekkor

$$d(x_k) = p(x_k) - q(x_k) = f_k - f_k = 0, \ k = 0, 1, \ldots, n.$$

Így d egy n-nél nem magasabb fokszámú polinom, amelynek $n + 1$ különböző zérushelye van. Ez csak úgy lehetséges, ha d azonosan nulla, azaz $p = q$.

Az l_m polinomokat **Lagrange-féle interpolációs alappolinomok**nak, p-t pedig **interpolációs polinom**nak nevezzük. A p polinom fenti, az alappolinomokkal felírt alakját az **interpolációs polinom Lagrange-féle alakjá**nak hívjuk.

Illusztrációként az ábrán megtekinthetünk egy hat alappontra illeszkedő interpolációs polinomot.



4.1. ábra. Hat alappontra illeszkedő interpolációs polinom.

4.1.1. Newton-féle alak

A Lagrange-féle alak az l_m alappolinomok lineáris kombinációjával adja meg az interpolációs polinomot. Ezek eggyel alacsonyabb fokszámúak, mint az alappontok száma, és mindegyik függ az összes alapponttól. Mi történik, ha felveszünk egy újabb alappontot, és keressük az ilyen módon eggyel több pontra illeszkedő polinomot? Ekkor eggyel több eggyel magasabb fokszámú alappolinomot kell felírnunk, és ehhez a korábbi, a kevesebb alappontra kapott eredményeink nem használhatók, a számításokat előlről kell kezdenünk. Kérdés: hogyan kaphatunk olyan formulát, amelynek segítségével az $n+1$ pontra illesztett polinom közvetlenül számolható az első n pontra illesztett interpolációs polinomból?

Legyenek adva az (x_k, f_k), $k = 0, 1, \ldots, n$ pontok, és jelölje $N_j \in P_j$ az első j darab pontra illesztett interpolációs polinomot, azaz legyen

$$N_j(x_k) = f_k,\ k = 0, 1, \ldots, j.$$

Tekintsük az

$$N_j(x) - N_{j-1}(x) =: r_j(x)$$

különbségfüggvényt. Nyilvánvalóan, $r_j \in P_j$, és $r_j(x_k) = 0,\ k = 0, 1, \ldots, j-1$. Így r_j egy olyan legfeljebb j-ed fokú polinom, amelynek gyökei $x_0, x_1, \ldots, x_{j-1}$, azaz

$$r_j(x) = N_j(x) - N_{j-1}(x) = b_j(x - x_0)(x - x_1)\cdots(x - x_{j-1}) = b_j\omega_j(x),\ j = 1, \ldots, n$$

ahol b_j valamilyen állandó, és ω_j az x_0, x_1, x_{j-1} alappontokhoz tartozó ún. alappontpolinom. Ezzel felírhatjuk az

$$N_j(x) = N_{j-1}(x) + b_j\omega_j(x),\ j = 0, 1, \ldots, n$$

rekurziót az $\{N_j(x)\}$ polinomok kiszámítására. Adjuk meg a kezdő $N_0(x)$ függvényt! A definíció alapján $N_0(x) = f_0$ (konstans függvény). Az alappontpolinom $j = 0$-ra egyelőre nem értelmes, de ezt célszerű $\omega_0(x) := 1$ módon definiálni, mert ekkor $b_0 = f_0$ megválasztással N_0 felírható

$$N_0(x) = b_0\omega_0(x)$$

alakban, a további N_j függvények pedig:

$$\begin{aligned} N_1(x) &= N_0(x) + b_1\omega_1(x) = b_0\omega_0(x) + b_0\omega_1(x) \\ &\vdots \\ N_n(x) &= \sum_{j=0}^{n} b_j\omega_j(x) \end{aligned} \tag{4.1}$$

Kérdés: hogyan határozzuk meg a b_j együtthatókat?

Mivel $N_j(x)$ az első j pontra illesztett interpolációs polinom, azt fel tudjuk írni a Lagrange-féle alakban:

$$N_j(x) = \sum_{m=0}^{j} f_m \left(\prod_{\substack{k=0 \\ k\neq m}}^{j} \frac{x - x_k}{x_m - x_k} \right)$$

Az x^j együtthatóinak a két oldalon meg kell egyezniük. Ez az együttható a bal oldalon b_j, a jobb oldalon pedig

$$\sum_{m=0}^{j} f_m \left(\prod_{\substack{k=0 \\ k\neq m}}^{j} \frac{1}{x_m - x_k} \right)$$

Így tehát

$$b_0 = 1; \tag{4.2}$$

$$b_j = \sum_{m=0}^{j} f_m \left(\prod_{\substack{k=0 \\ k\neq m}}^{j} \frac{1}{x_m - x_k} \right), \ j = 1, \ldots, n. \tag{4.3}$$

4.1.2 Definíció. *A (4.2), (4.3) együtthatókkal felírt (4.1) alakú $N_n \in P_n$ polinomot Newton-féle interpolációs polinomnak nevezzük.*

4.1.3 Megjegyzés. *Mivel az interpolációs polinom egyértelmű, pontosabb az interpolációs polinom Newton-féle alakja elnevezés.*

Nézzük meg, hogy speciális n esetén mit jelent (4.1), és így $N_n(x)$!

1.) $n = 1$ eset:

$$b_1 = \sum_{m=0}^{1} f_m \left(\prod_{\substack{k=0 \\ k\neq m}}^{1} \frac{1}{x_m - x_k} \right) = f_0 \frac{1}{x_0 - x_1} + f_1 \frac{1}{x_1 - x_0} = \frac{f_1 - f_0}{x_1 - x_0}$$

$$\Rightarrow N_1(x) = f_0 + \frac{f_1 - f_0}{x_1 - x_0}(x - x_0).$$

2.) $n = 2$ eset:

$$b_2 = \sum_{m=0}^{2} f_m \left(\prod_{\substack{k=0 \\ k\neq m}}^{2} \frac{1}{x_m - x_k} \right) = f_0 \frac{1}{(x_0 - x_1)(x_0 - x_2)} + f_1 \frac{1}{(x_1 - x_0)(x_1 - x_2)} + f_2 \frac{1}{(x_2 - x_0)(x_2 - x_1)}$$

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

Az együtthatókra kapott kifejezések áttekinthetőbbé válnak a következő fogalom bevezetésével.

4.1.4 Definíció. *(Osztott differenciák)*
Nulladrendű osztott differenciáknak nevezzük az

$$f[x_i] := f(x_i)$$

függvényértékeket.
Elsőrendű osztott differenciáknak nevezzük az

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

hányadosokat.
Továbbmenve, ha $k \in \mathbb{N}^+$, akkor a k-adrendű osztott differenciákat az

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

rekurzív képlettel definiáljuk.

Azt sikerült tehát belátnunk, hogy

$$b_0 = f[x_0]; \tag{4.4}$$

$$b_1 = f[x_0, x_1]; \tag{4.5}$$

$$b_2 = f[x_0, x_1, x_2] \tag{4.6}$$

4.1.5 Állítás. *Általában is:* $b_j = f[x_0, x_1, \ldots, x_j]$

A bizonyítást itt nem részletezzük, teljes indukcióval elvégezhető.

A gyakorlatban a Newton-féle alak felírásához érdemes ún. differenciatáblázatot készíteni. Írjuk fel egymás alá az alappontokat, és mindegyik mellé a hozzá tartozó nulladrendű osztott differenciát, majd a további osztott differenciákat a 4.1 szerinti elrendezésben.

$$\begin{array}{llll}
x_0 & \underline{f[x_0]} & & \\
 & & \underline{f[x_0,x_1]} & \\
x_1 & f[x_1] & & \underline{f[x_0,x_1,x_2]} \\
 & & f[x_1,x_2] & & \underline{f[x_0,x_1,x_2,x_3]} \\
x_2 & f[x_2] & & f[x_1,x_2,x_3] & \\
 & & f[x_2,x_3] & & \\
x_3 & f[x_3] & & &
\end{array}$$

4.1. táblázat. A differenciatáblázat négy alappont esetén.

A következő tétel az interpolációs polinomot az osztott differenciákkal fejezi ki.

4.1.6 Tétel. *(Az interpolációs polinom Newton-féle alakja)*
A p interpolációs polinom előáll a

$$p(x) = f[x_0] + f[x_0, x_1] \cdot (x - x_0) + f[x_0, x_1, x_2] \cdot (x - x_0)(x - x_1) + \ldots$$
$$\ldots + f[x_0, \ldots, x_n] \cdot (x - x_0) \cdots (x - x_{n-1}) \qquad (4.7)$$

alakban.

Az interpolációs polinom felírásához tehát a differenciatáblázatban aláhúzott elemek kellenek. Látható, hogy egy újabb alappont hozzávételekor csak egy új tagot kell hozzáadni a korábbi alakhoz, és ez a tag a már elkészített differenciatáblázat kibővítésével könnyen kiszámítható.

Feladatok. 1. Írjuk fel az interpolációs polinom Newton-féle alakját, ha $x_0 = -1$, $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, és $f_0 = 1$, $f_1 = -1$, $f_2 = -1$, $f_3 = 1$.

Megoldás: Készítsük el a differenciatáblázatot:

Ebből $p(x) = 1 + (-2)(x + 1) + 1(x + 1)x$.

2. Vegyünk fel egy újabb alappontot: $x_4 = 3$, $f_4 = 2$, és írjuk fel az új $p(x)$-et!

Megoldás: A kibővített differenciatáblázat alapján $p(x) = 1 + (-2)(x+1) + 1(x+1)x + 0 + \frac{1}{6}(x+1)x(x-1)(x-2)$.

$$
\begin{array}{cccccc}
-1 & \underline{1} & & & & \\
 & & \underline{-2} & & & \\
0 & -1 & & \underline{1} & & \\
 & & 0 & & \underline{0} & \\
1 & -1 & & 1 & & \\
 & & 2 & & & \\
2 & 1 & & & &
\end{array}
$$

$$
\begin{array}{cccccc}
-1 & \underline{1} & & & & \\
 & & \underline{-2} & & & \\
0 & -1 & & \underline{1} & & \\
 & & 0 & & \underline{0} & \\
1 & -1 & & 1 & & \underline{\frac{1}{6}} \\
 & & 2 & & -\frac{1}{2} & \\
2 & 1 & & -\frac{1}{2} & & \\
 & & 1 & & & \\
3 & 2 & & & &
\end{array}
$$

4.1.2. Az interpolációs polinom hibája

Az interpolációs polinom alkalmazása folytonos függvény közelítésénél is szóba jöhet. Tegyük fel, hogy az $f : I \to \mathbb{R}$ folytonos függvényt szeretnénk polinommal közelíteni. Vegyünk fel alappontokat I-n, és az $(x_i, f(x_i))$ pontokra illesszünk interpolációs polinomot. Kérdés: mennyire jól közelíti a kapott p függvény az f-et? Ha f kellően sima, akkor tudunk képletet adni a hibára:

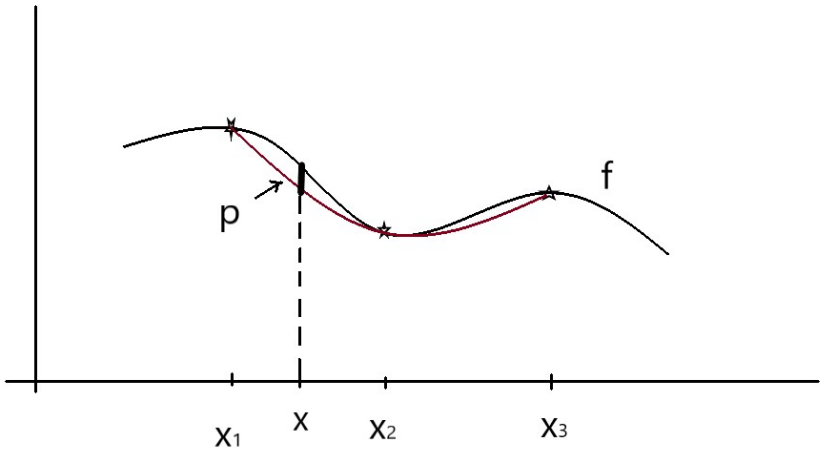
4.1.7 Tétel. *Legyen $f : I \to \mathbb{R}$ $n+1$-szer folytonosan differenciálható függvény, p az a legfeljebb n-edfokú polinom, amely az I intervallum $n+1$ különböző x_k $(k = 0, 1, \ldots, n)$ pontjában interpolálja f-et, és $x \in I$ tetszőleges pont. Ekkor az x-et és az összes x_k pontot tartalmazó legszűkebb intervallum tartalmaz a belsejében egy olyan ξ pontot, amelyre*

$$f(x) - p(x) = \frac{1}{(n+1)!}\omega_{n+1}(x) f^{(n+1)}(\xi), \tag{4.8}$$

ahol

$$\omega_{n+1}(x) := (x - x_0)(x - x_1) \cdots (x - x_n) \tag{4.9}$$

a korábbiakban már bevezetett alappontpolinom.



Biz.:

- $x = x_k$-ra nyilvánvaló az állítás, hiszen ekkor (4.8) mindkét oldala tetszőleges ξ-re nulla.

- Tegyük fel, hogy $x \neq x_k$ minden k-ra. Tekintsük a

$$g(t) := f(t) - p(t) - c\omega_{n+1}(t) \tag{4.10}$$

segédfüggvényt, ahol c egyelőre tetszőleges állandó. Nyilván

$$g(x_k) = f_k - f_k - 0 = 0, \quad k = 0, 1, \ldots, n.$$

Válasszuk meg c-t úgy, hogy a g függvény $t = x$ esetén is tűnjön el, azaz $f(x) - p(x) - c\omega_{n+1}(x) = 0$ legyen. Ebből

$$c = \frac{f(x) - p(x)}{\omega_{n+1}(x)}.$$

Ezen c-vel g-nek legalább $n + 2$ gyöke lesz: $x_0, x_1, \ldots, x_n$ és x, és valamennyi az I intervallumra esik. Rolle tétele szerint g'-nek legalább $n + 1$ gyöke van, g''-nek legalább n és így tovább, a $g^{(n+1)}$-nek legalább egy. Ezek abban a legszűkebb intervallumban vannak, amely tartalmazza az $x_0, x_1, \ldots, x_n$ és x pontokat. Legyen ξ a $g^{(n+1)}$ függvény egy ilyen gyöke. A (4.10) egyenlőséget $n + 1$-szer deriválva

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

A $t = \xi$ helyettesítéssel

$$f^{(n+1)}(\xi) = c(n+1)!$$

Figyelembe véve c megválasztását, éppen a (4.8) egyenlőséget kapjuk.

Alkalmazzuk a tételt $n = 0$-ra (egy darab alappont)! Ekkor p nulladfokú polinom, azaz konstans függvény, értéke minden pontban $f(x_0)$. Ekkor a (4.8) egyenlőség speciálisan

$$f(x) - f(x_0) = (x - x_0)f'(\xi), \text{ ahol } \xi \text{ } x \text{ és } x_0 \text{ közötti,}$$

és ez az analízisből ismert Lagrange-középértéktétel.
Tekintsük most az $n = 1$ esetet (két alappont). Ekkor lineáris interpolációról beszélünk. A (4.8) egyenlőség:

$$f(x) - p(x) = \frac{(x - x_0)(x - x_1)}{2} f''(\xi)$$

4.2. Hermite-interpoláció

Eddig az $x_0, x_1, \ldots, x_n$ alappontokban csak az $f_0, f_1, \ldots, f_n$ függvényértékek voltak előírva az interpoláció során. Most foglalkozzunk azzal az általánosabb esettel, amikor az alappontokban bizonyos rendig bezárólag a deriváltakat is előírjuk.

Tegyük fel, hogy az x_k pontban az m_k-adikig bezárólag írjuk elő a deriváltakat, vagyis adott $f_k^{(0)}, f_k^{(1)}, \ldots f_k^{(m_k)}$ ($m_k + 1$ darab szám) az összes $x_k, k = 0, 1, \ldots, n$ ponthoz. Ez összesen $N = n+1+m_0+m_1+\ldots+m_n$ darab feltételt jelent. Várható, hogy $N-1$-edfokú polinom egyértelműen illeszthető a pontokra. Valóban igaz a következő

4.2.1 Állítás. *Létezik egyetlen* $H \in P_{N-1} : H^{(i)}(x_k) = f_k^{(i)},\ k = 0, 1, \ldots, n, i = 0, 1 \ldots, m_k.$

A H polinomot **Hermite-féle interpolációs polinom**nak nevezzük. Ennek speciális esete, amikor mindegyik alappontban a függvényérték és az első derivált van előírva, ekkor Hermite–Fejér-interpolációról beszélünk, és a polinom fokszáma $N - 1 = 2n + 1$.

Az Hermite-féle interpolációs polinom felírása történhet

1. a határozatlan együtthatók módszerével;

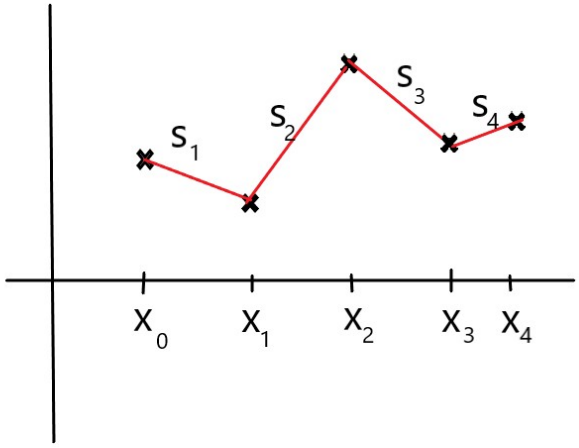
2. differenciatáblázattal.

Az utóbbit úgy készítjük el, hogy mindegyik x_k alappontot $m_k + 1$-szer veszünk fel az első oszlopban, mellé jük $f_k^{(0)}$-t írjuk, továbbá az i-ed rendű osztott differenciák oszlopában azonos alappontok esetén (azaz ha nullával osztanánk) az $\frac{f^{(i)}(x_k)}{i!}$ értéket írjuk be.

4.3. Spline-interpoláció

Az eddigiekben az interpolációs alappontokra egyetlen polinomot illesztettünk. A 4.1 ábrán látható ennek a megközelítésnek a fő hátránya. Az alappontok között – az ábrán bemutatott Lagrange-féle interpoláció polinom esetén ez különösen az első és az utolsó részintervallumon szembetűnő – a függvény túlságosan ingadozik. Minél nagyobb fokszámú az interpolációs polinom (azaz minél több alappont van), általában annál nagyobbak az illesztett görbe ingadozásai. Ennek oka az, hogy egy n-edfokú polinom deriváltja $(n-1)$-edfokú polinom, amelynek akár $n-1$ zérushelye is lehet. Ahol pedig a derivált nulla, ott az interpolációs polinomnak igen gyakran lokális szélsőértéke van, ami kacskaringós grafikont eredményez. Könnyen előfordul, hogy egy adott szakaszon monoton növő vagy csökkenő függvényértékek esetén az interpolációs polinom már nem monoton. Mindezen hiányosságokat kiküszöbölhetjük szakaszonként polinomiális interpolációval: az egyes alappontok között más-más (rendszerint alacsony fokszámú) polinomot alkalmazunk. Ezt nevezzük **spline-interpoláció**nak.

A spline-interpoláció legegyszerűbb esete a szakaszonként lineáris interpoláció, amikor a szomszédos pontokat egyenes szakaszokkal kötjük össze.



Ekkor az $[x_{k-1}, x_k]$ intervallumon az

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

függvényt írhatjuk fel, és a teljes interpolációs függvény:

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

Ez folytonos, hiszen $s_k(x_k) = s_{k+1}(x_k) = f_k,\ k = 1, 2, \ldots, n$. Hibája – az interpolációs polinom hibája alapján – $f \in C^2(I)$ függvényre

$$|f(x) - s(x)| \leq \frac{M_2}{8} h^2,$$

ahol $M_2 \geq |f''(x)|\ \forall x \in I$, és $h = \max_k |x_k - x_{k-1}|$.

Ekkor monoton f_k értékek esetén az $s(x)$ függvény is monoton, de hátrány, hogy s nem lesz differenciálható az $[x_0, x_n]$ intervalumon. Ez orvosolható, ha kvadratikus spline-interpolációt alkalmazunk, pl. a következő módon:

1) Válasszuk meg az $s'(x_0)$ értéket, jelölje ezt d_0.
2) Alkalmazzunk Hermite-interpolációt az $[x_0, x_1]$ szakaszon, vagyis keressük meg azt az $s_1 \in P_2$ polinomot, amelyre

$$s_1(x_0) = f_0,$$

$$s_1'(x_0) = d_0,$$

$$s_1(x_1) = f_1$$

3) Az $[x_1, x_2]$ szakaszon – ismét Hermite-interpolációt végezve – meghatározzuk azt az $s_2 \in P_2$ polinomot, amelyre

$$s_2(x_1) = f_1,$$

$$s_2'(x_1) = s_1'(x_1),$$

$$s_2(x_2) = f_2$$

És így tovább, a teljes s függvény szakaszonként P_2-beli, és a deriváltja folytonos az egész $[x_0, x_n]$ intervallumon.

Még magasabb fokszámú polinomokat is illeszthetünk, pl. legfeljebb harmadfokú polinomok esetén köbös spline-t kapunk...

4.4. Trigonometrikus interpoláció

Ha a diszkrét pontokban ismert értékek egy periodikus függvényhez tartoznak, akkor a pontokra érdemes lehet trigonometrikus polinomot illeszteni. Ezt nevezzük trigonometrikus interpolációnak.

Tegyük fel, hogy f egy 2π szerint periodikus függvény, amelynek értékeit az

$$x_k = \frac{k}{n+1} 2\pi,\ k = 0, 1, \ldots, n$$

pontokban ismerjük. Keressük azt a

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

alakú trigonomtrikus polinomot, amelyre

$$t_m(x_k) = f_k, \ k = 0, 1, \ldots, n$$

Az $a_0, a_j, b_j, \ j = 1, 2, \ldots, m$ számokat az f függvény diszkrét Fourier-együtthatóinak nevezzük. Ekkor összesen $n+1$ darab egyenletünk van $2m+1$ darab diszkrét Fourier-együtthatóra. Ezért ha n páros, akkor azt várjuk, hogy $m = n/2$, ha pedig n páratlan, akkor $m = \frac{n+1}{2}$ fokú trigonometrikus polinomra lesz szükségünk az illesztéshez. Az előbbi esetben ugyanannyi egyenlet lesz, ahány ismeretlen, az utóbbi esetben pedig $2m+1 = n+2$ darab együtthatóra lesz $n+1$ darab egyenlet, így a rendszer alulhatározott. Ilyenkor viszont

$$b_m \sin(mx_k) = b_m \sin\left(\frac{n+1}{2}\frac{k}{n+1}2\pi\right) = b_m \sin(\pi k) = 0,$$

így b_m értéke nem befolyásolja az interpolációt, b_m választható nullának. Ekkor kiegyensúlyozott trigonometrikus polinomról beszélünk.

Igazolható, hogy mindkét esetben valóban megvalósítható az interpoláció, az illesztett trigonometrikus polinom egyértelmű, és a diszkrét Fourier-együtthatók a következők lesznek:

1. Ha n páros:

$$a_0 = \frac{1}{n+1} \sum_{k=0}^{n} f_k,$$

$$a_j = \frac{2}{n+1} \sum_{k=0}^{n} f_k \cos(jx_k), \ j = 1, 2, \ldots, m$$

$$b_j = \frac{2}{n+1} \sum_{k=0}^{n} f_k \sin(jx_k), \ j = 1, 2, \ldots, m$$

2. Ha n páratlan:

$$a_0 = \frac{1}{n+1} \sum_{k=0}^{n} f_k$$

$$a_j = \frac{2}{n+1} \sum_{k=0}^{n} f_k \cos(jx_k), \ j = 1, 2, \ldots, m-1$$

$$a_m = \frac{1}{n+1} \sum_{k=0}^{n} f_k \cos(mx_k)$$

$$b_j = \frac{2}{n+1} \sum_{k=0}^{n} f_k \sin(jx_k), \;\; j = 1, 2, \ldots, m-1$$

Megjegyezzük, hogy ezek az összegek az f függvény Fourier-sorában szereplő együtthatókat megadó integrálok közelítései.

4.5. A legkisebb négyzetek módszere

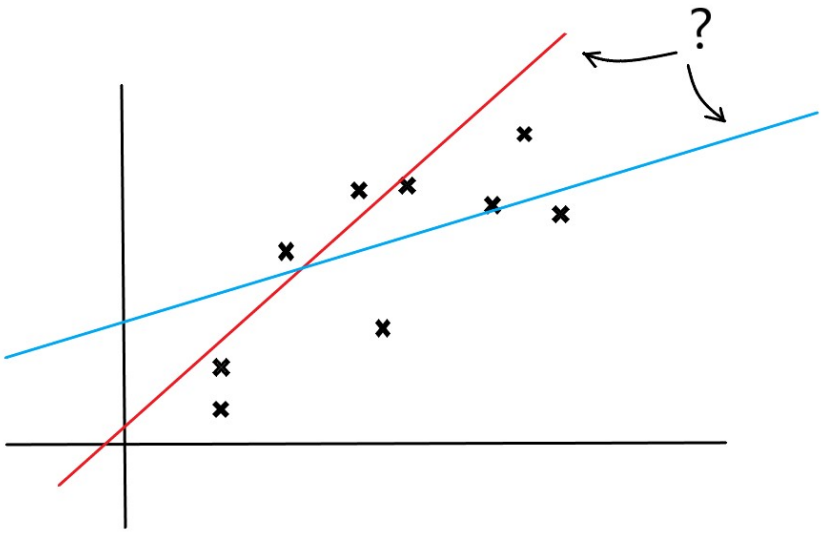
Jelölje továbbra is $x_0, x_1, \ldots, x_n$ az alappontokat, és $f_0, f_1, \ldots, f_n$ a hozzájuk tartozó függvényértékeket. Az összes fajta interpolációs polinom éppen átmegy mindegyik ponton. Sokszor azonban az értékek mérésekből származnak, tehát nem egészen pontosak, így enyhíthetünk azon az előírásunkon, hogy az illesztett függvény mindegyiken pontosan átmenjen, elég nekünk, ha valamilyen értelemben közel halad a pontokhoz. (Ezt ugyanakkor már nem nevezzük interpolációnak!) Emellett a pontokra ránézve esetleg látjuk, hogy azok durván egy egyenes vagy parabola mentén helyezkednek el. Olyan p polinomot szeretnénk találni, amely adott fokszámú, és valamilyen értelemben a legközelebb halad az összes ponthoz. Kérdés: hogyan értsük a közelséget? Lehetne a

$$\sum_{k=0}^{n} |f_k - p(x_k)|$$

összeget használni, de ennek hátránya, hogy az abszolút érték függvény nem differenciálható. Ezért helyette minden tagban négyzetre emeljük a különbséget. A legkisebb négyzetek módszere azt az adott fokszámú p polinomot határozza meg, amelyre a

$$\sum_{k=0}^{n} [f_k - p(x_k)]^2$$

kifejezés értéke minimális.



Nézzük először azt az esetet, amikor a pontokra $p(x) = c_0 + c_1 x$ alakú elsőfokú polinomot, azaz egyenest illesztünk. Mely c_0 és c_1 esetén lesz minimális a fenti összeg, vagyis a

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

kifejezés? Itt x_k és f_k, $k = 0, 1, \ldots, n$ adottak, és a (c_0, c_1) számpártól függő

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

$\mathbb{R}^2 \to \mathbb{R}$ függvény minimumhelyét keressük.Ez ott lehet, ahol F deriváltja (gradiense) nulla. Deriváljuk tehát c_0 és c_1 szerint az F függvényt:

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

Ezeket nullával egyenlővé téve – egyszerűsítések után – a

$$\begin{aligned} \sum_{k=0}^{n} [f_k - c_0 - c_1 x_k] &= 0 \\ \sum_{k=0}^{n} [f_k x_k - c_0 x_k - c_1 x_k^2] &= 0. \end{aligned}$$

lineáris algebrai egyenletrendszert kapjuk. Mátrixos alakban:

$$\begin{bmatrix} n+1 & \sum_{k=0}^{n} x_k \\ \sum_{k=0}^{n} x_k & \sum_{k=0}^{n} x_k^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} \sum_{k=0}^{n} f_k \\ \sum_{k=0}^{n} f_k x_k \end{bmatrix}$$

Ezt kell megoldanunk a c_0, c_1 ismeretlenekre. Általános, esetben, ha N-edfokú polinomot akarunk a pontokra illeszteni, akkor $N+1$-ismeretlenes lineáris algebrai egyenletrendszert kapunk.

5. fejezet

Numerikus integrálás

Legyen $f : [a, b] \to \mathbb{R}$ egy Riemann-integrálható függvény. Kíváncsiak vagyunk a függvény integráljára az $[a, b]$ intervallumon. Az analízisből ismeretes, hogy ha f Riemann-integrálható az $[a, b]$ intervallumon és létezik F primitív függvénye, akkor a Newton–Leibniz-szabály értelmében

$$\int_a^b f(x)\mathrm{d}x = F(b) - F(a).$$

A primitív függvényt azonban nem mindig tudjuk meghatározni. Ezért van szükség olyan módszerekre, amelyekkel az $\int_a^b f(x)\mathrm{d}x$ integrált legalább közelítőleg ki tudjuk számítani.

A legegyszerűbb ötlet erre azon alapszik, hogy a Riemann-integrál (nemnegatív függvény esetén) szemléletesen a grafikon alatti területet adja meg, közelítsük tehát ezen alakzat területét valamilyen egyszerű alakzat területével, amely a függvény néhány pontban felvett értékéből kiszámítható. Ennek különféle lehetőségeit tekintjük át az 5.1. alfejezetben. A közelítő integrálás formuláit szokás kvadratúraformuláknak is nevezni. (A „kvadratúra” szó arra az elemi területszámítási módszerre utal, amikor a függvény grafikonját négyzethálós papírra visszük, és összeadjuk a görbe alatti kis négyzetek területét, de általánosan is így nevezzük a közelítő integrálási módszereket.)

5.1. Az alapvető kvadratúraformulák

Jelölje $h := b - a$ az intervallum hosszát. Ismerkedjünk meg először két alapvető kvadratúraformulával.

1. **Középponti szabály:** annak a téglalapnak a területével közelítjük az integrált, amelynek alapja h, magassága pedig az intervallum $c = \frac{a+b}{2}$ felezőpontjában felvett

függvényérték, azaz

$$k(f) := h \cdot f(c).$$

2. **Trapézszabály:** annak a trapéznak a területével közelítjük az integrált, amelynek alapja h, oldalainak hosszai pedig az intervallum két végpontjában felvett függvényértékek, azaz

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

Egy kvadratúraformulát jól jellemez az, hogy hogyan viselkedik különböző fokszámú polinomokra. Egy kvadratúraformula rendje annak a legalacsonyabb fokú polinomnak a fokszáma, amelyre a formula már nem adja meg pontosan az integrált. Pl. a középponti és a trapézszabály minden f elsőfokú polinomra pontos, de a másodfokú polinomokra már nem, ezért másodrendű kvadratúraformuláknak nevezzük őket. Általában ha egy p-edrendű kvadratúraformulát megfelelően sima függvényre alkalmazunk egyre csökkenő hosszúságú intervallumokon, akkor a területszámítás hibája h p-edik hatványával arányos.

Alkalmazzuk a középponti és a trapézszabályt az $f(x) = x^2$ függvényre a $[0, 1]$ intervallumon!
Az integrál pontos értéke

$$\int_0^1 x^2 \mathrm{d}x = \left.\frac{x^3}{3}\right|_0^1 = \frac{1}{3}.$$

A középponti szabály eredménye

$$k(f) = 1 \cdot \left(\frac{1}{2}\right)^2 = \frac{1}{4},$$

a trapézszabályé pedig

$$t(f) = 1 \cdot \frac{0+1}{2} = \frac{1}{2}.$$

Így $k(f)$ hibája $\frac{1}{3} - \frac{1}{4} = \frac{1}{12}$, $t(f)$ hibája pedig $\frac{1}{3} - \frac{1}{2} = -\frac{1}{6}$. Látjuk, hogy a hibák ellentétes előjelűek, és a középponti szabály kétszer olyan pontos, mint a trapézszabály. Ezt az észrevételt felhasználhatjuk arra, hogy még pontosabb formulát gyártsunk.
Nyilván, hogyha $t(f)$ hibája pontosan -2-szerese $k(f)$ hibájának a vizsgált $f(x) = x^2$ függvény esetén, akkor a pontos integrált I-vel jelölve fennáll az

$$I - t(f) = -2(I - k(f))$$

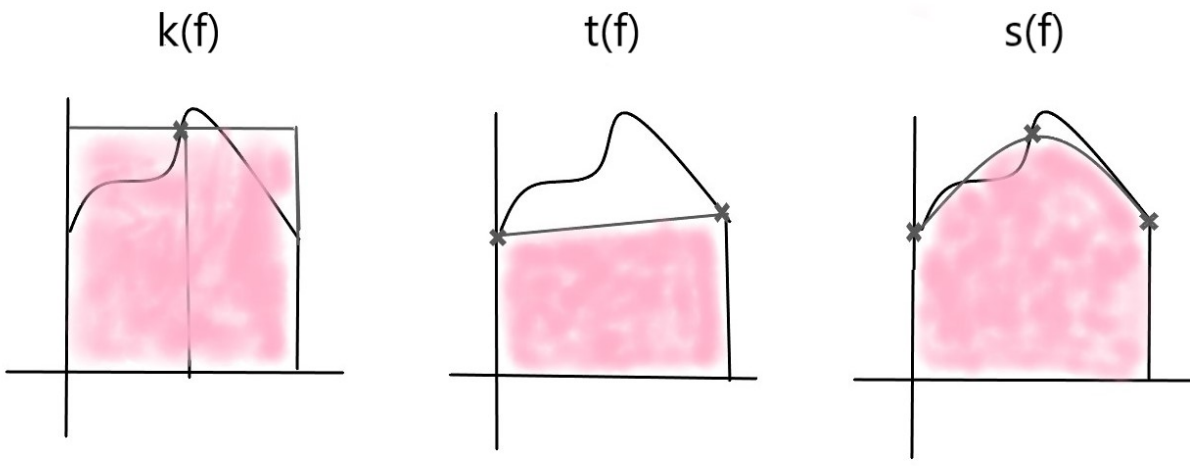
egyenlőség. Ebből I-t kifejezve az

$$I = \frac{2}{3}k(f) + \frac{1}{3}t(f)$$

képletet kapjuk, amelyet kezelhetünk újabb területszámítási formulaként. Az így nyert

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

formulát **Simpson-formulá**nak vagy **parabolaformulá**nak nevezzük. Erről egyelőre annyit tudunk, hogy az $f(x) = x^2$ függvényre pontos a $[0, 1]$ intervallumon.
A különböző kvadratúraformulák szemléltetésére szolgál a 5.1. ábra. (A Simpson-formulánál feltüntetett alakzatot később magyarázzuk meg!)

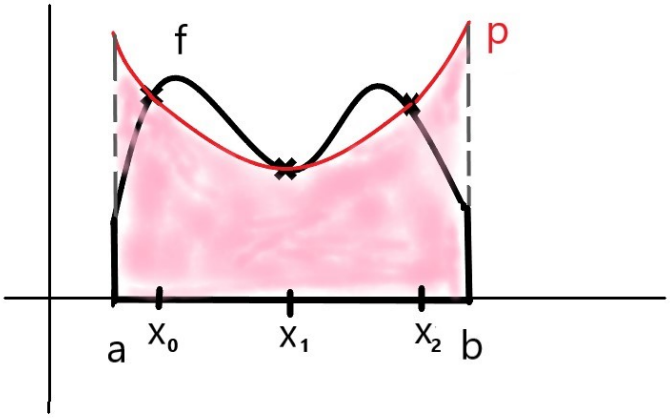


5.1. ábra. A középponti szabály, a trapézszabály és a Simpson-formula szemléltetése. A közelítő integrál a rózsaszínnel színezett alakzatok területe.

Kvadratúraformulákat másképpen is származtathatunk, pl. úgy, hogy f-et helyettesítjük valamely interpolációs polinomjával, és ezen interpolációs polinom grafikon alatti területét számítjuk ki. Ezzel foglalkozik a következő alfejezet, amelyben már részletesebben megvizsgáljuk a kvadratúraformulák hibáját. Megtudjuk azt is, hogy a Simpson-formulát miért nevezik másképpen parabolaformulának.

5.2. Interpolációs típusú kvadratúraformula

Legyen $[a, b] \subset \mathbb{R}$ zárt intervallum, és $x_0, x_1, \ldots, x_n \in [a, b]$ tetszőleges alappontok. Az $f : [a, b] \to \mathbb{R}$ függvény integrálját közelítsük az őt interpoláló polinom integráljával!



Az interpolációs polinom hibájáról tanultak értelmében, ha $f \in C^{n+1}[a,b]$, akkor egy $x \in [a,b]$ pontban a hiba

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x), \tag{5.1}$$

ahol $p(x) = \sum_{k=0}^{n} l_k(x) f(x_k)$ az interpolációs polinom, és ξ az x-et és az alappontokat tartalmazó legszűkebb intervallum valamely pontja. Integráljuk a (5.1) egyenletet az $[a,b]$ intervallumon! Vigyázzunk, hogy mivel a fenti képletben ξ függ az x ponttól, és itt x szerint integrálunk, a jobb oldalon ξ x-függését figyelembe kell venni.

$$\int_a^b f(x)\mathrm{d}x - \int_a^b \left(\sum_{m=0}^{n} l_m(x) f(x_m) \right) \mathrm{d}x = \int_a^b \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\omega_{n+1}(x)\mathrm{d}x. \tag{5.2}$$

A bal oldali második tag az integrál közelítése, amely másképpen a

$$\sum_{m=0}^{n} \left(\int_a^b l_m(x)\mathrm{d}x \right) f(x_m) \tag{5.3}$$

alakba írható, a jobb oldali tag pedig a közelítés képlethibája.

5.2.1 Definíció. *A 5.3 kifejezést* **interpolációs típusú kvadratúraformulá***nak hívjuk.*

5.2.2 Definíció. *Általánosan a* $\sum_{m=0}^{n} A_m f(x_m)$ *alakú kifejezést* **lineáris kvadratúraformulá***nak, az* A_m *számokat pedig a kvadratúraformula együtthatóinak nevezzük.*

Az interpolációs típusú kvadratúraformula tehát olyan lineáris kvadratúraformula, amelynek együtthatóira speciálisan

$$A_m := \int_a^b l_m(x)\mathrm{d}x,\ m = 0, 1, \ldots, n.$$

Sokféleképpen lehet származtatni lineáris kvadratúraformulákat. Az interpolációs típusú azonban rendelkezik a következő előnyös tulajdonsággal.

5.2.3 Tétel. *Egy lineáris kvadratúraformula akkor és csak akkor pontos minden legfeljebb n-edfokú polinomra, amikor interpolációs típusú.*

Biz.: ($\Leftarrow$) Visszafelé az állítás nyilvánvalóan következik abból, hogy egy legfeljebb n-edfokú polinom interpolációs polinomja önmaga.
($\Rightarrow$) Mivel az $l_j, j = 0, 1, \ldots, n$ Lagrange-féle alappolinomok n-edfokú polinomok, ezért mindegyikükre pontos a kvadratúraformula. Azaz

$$\int_a^b l_j(x)\mathrm{d}x = \sum_{m=0}^{n} A_m l_j(x_m) = A_j.$$

Az utolsó lépésben felhasználtuk, hogy $l_j(x_m) = 1$, ha $m = j$, és 0, ha $m \neq j$.

5.2.1. Newton–Cotes-formulák

5.2.4 Definíció. *A (5.3) interpolációs típusú kvadratúraformulát Newton–Cotes-formulának nevezzük, ha $[a, b]$ felosztása egyenlőközű. Továbbá, zárt Newton–Cotes-formuláról beszélünk, ha a és b is alappont.*

Speciális esetek

- $n = 1$, azaz két alappont van: $x_0 = a$ és $x_1 = b$. Ekkor az interpolációs polinom lineáris függvény. Számítsuk ki a formula együtthatóit!

$$A_0 = \int_a^b l_0(x)\mathrm{d}x = \int_a^b \frac{x - x_1}{x_0 - x_1}\mathrm{d}x = \int_a^b \frac{x - b}{a - b}\mathrm{d}x = \frac{1}{a - b}\left[\frac{x^2}{2} - bx\right]_a^b =$$
$$= \frac{1}{a - b}\left(\frac{b^2}{2} - b^2 - \frac{a^2}{2} + ab\right) = \frac{1}{a - b}\frac{-b^2 - a^2 + 2ab}{2} = -\frac{(a - b)^2}{2(a - b)} = \frac{b - a}{2} = \frac{h}{2}.$$

Hasonlóan,

$$A_1 = \int_a^b l_1(x)\mathrm{d}x = \int_a^b \frac{x - x_0}{x_1 - x_0}\mathrm{d}x = \frac{b - a}{2} = \frac{h}{2}.$$

Így a jól ismert

$$t(f) := \frac{h}{2}(f(a) + f(b)) \tag{5.4}$$

trapézformulát kapjuk. Nem jutottunk tehát új módszerhez, de észrevételünket felhasználhatjuk arra, hogy meghatározzuk a trapézformula képlethibáját. A képlethiba $f \in C^2[a,b]$ függvényekre, (5.2)-ből, integrál-középértéktételt alkalmazva ($x \mapsto (x-a)(x-b)$ integrálható, előjeltartó függvény, és $f''(\xi(x))$ folytonos függvény):

$$\int_a^b f - t(f) = \int_a^b \frac{f''(\xi(x))}{2!}(x-a)(x-b)\mathrm{d}x$$
$$= f''(\kappa)\int_a^b \frac{1}{2}(x-a)(x-b)\mathrm{d}x = -\frac{h^3}{12}f''(\kappa)$$

ahol $\kappa \in [a,b]$ valamely pont.

- $n=2$, azaz három alappont van: $x_0 = a, x_1 = \frac{a+b}{2}$ és $x_2 = b$. Ekkor az interpolációs polinom legfeljebb másodfokú, az együtthatók:

$$A_0 = \int_a^b \frac{(x-\frac{a+b}{2})(x-b)}{(a-\frac{a+b}{2})(a-b)}\mathrm{d}x = \frac{h}{6} = A_2, \quad A_1 = \frac{4h}{6}.$$

Ezekből kapjuk a már szintén ismert

$$s(f) = \frac{h}{6}\left(f(a) + 4f(c) + f(b)\right), \quad c = \frac{a+b}{2}$$

Simpson- vagy **parabolaformulát**, amely tehát nem más, mint a függvényt közelítő, $f(a)$, $f(c)$ és $f(b)$ értékekre illeszkedő másodfokú interpolációs polinom grafikon alatti területe, lásd a 5.1. ábra bal alsó grafikonját. Képlethibája $f \in C^4[a,b]$ függvényekre (bizonyítás nélkül)

$$\int_a^b f - s(f) = \int_a^b \frac{f'''(\xi(x))}{3!}(x-a)\left(x-\frac{a+b}{2}\right)(x-b)\mathrm{d}x = -\frac{h^5}{2880}f^{IV}(\kappa),$$

ahol $\kappa \in [a,b]$ valamely pont.

5.2.5 Következmény. *A Simpson-formula a harmadfokú polinomokra is pontos, mert azok negyedik deriváltja nulla. Negyedfokúakra már nem pontos, így negyedrendű kvadratúra-formuláról van szó.*

5.2.2. Összetett kvadratúraformulák

Kérdés: Mit tegyünk, ha ezek az egyszerű formulák nem elég pontosak? Olyan formula kellene, amellyel az integrált tetszőleges pontossággal meg tudjuk kapni.

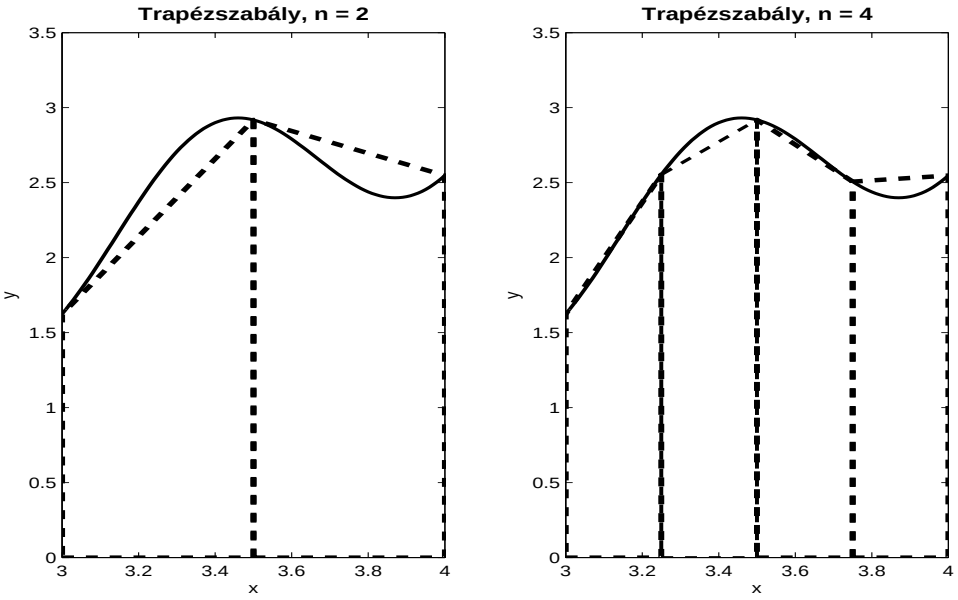
Az összes fenti formula képlethibájából jól látható, hogy minél kisebb h, azaz minél rövidebb az $[a,b]$ intervallum, annál kisebb a hiba. Ez azt az ötletet adja, hogy osszuk fel

kis részekre az $[a,b]$-t az $x_0 = a, x_1, \ldots, x_m = b$ osztópontokkal (így m db részintervallumot kapunk) és szakaszonként alkalmazzuk ezeket a formulákat: minden részintervallumon közelítjük az integrált, és ezeket a közelítő integrálokat összegezzük (hiszen az integrál additív). Az egyszerűség kedvéért dolgozzunk ekvidisztáns felosztással, ekkor minden egyes részintervallum hossza $\Delta h = \frac{b-a}{m} = \frac{h}{m}$.

Összetett trapézformula

Ha pl. a trapézformulát alkalmazzuk, és $m = 2$ ill. $m = 4$ egyenlő részre osztjuk az $[a,b]$-t, akkor a 5.2 ábrán szaggatott vonallal berajzolt kis trapézok összterületével közelítjük az integrált. Ekkor

$$\int_a^b f = \sum_{i=1}^{m} \int_{x_{i-1}}^{x_i} f \approx \sum_{i=1}^{m} \frac{\Delta h}{2}(f(x_{i-1})+f(x_i)) = \frac{\Delta h}{2}(f(x_0)+2\sum_{i=1}^{m-1} f(x_i)+f(x_m)) =: t_m(f).$$



5.2. ábra. Az $f(x) = \frac{1}{2}\sin 6x + x - 1$ függvény grafikon alatti területének kiszámítása a $[3,4]$ intervallumon az összetett trapézszabállyal $n = 2$ és $n = 4$ részintervallum esetén. Az integrált a szaggatott vonallal határolt tartományok területének összege annál jobban megközelíti, minél több részre daraboljuk az intervallumot.

Az összetett trapézformula képlethibája $f \in C^2[a,b]$ függvényekre

$$\int_a^b f - t_m(f) = \sum_{i=1}^{m} \frac{-\Delta h^3}{12} \cdot f''(\kappa_i) = -\frac{h^3}{12m^2} \cdot \frac{1}{m} \sum_{i=1}^{m} f''(\kappa_i), \tag{5.5}$$

ahol $\kappa_i \in [x_{i-1}, x_i]$. Az $\frac{1}{m}\sum_{i=1}^{m} f''(\kappa_i)$ kifejezés m darab függvényérték számtani közepe. Tegyük fel, hogy $f \in C^2[a,b]$. Ekkor f'' valamely $\eta \in [a,b]$ helyen felveszi ezt a számtani közepet. Ebből a képlethiba a

$$-\frac{h^3}{12m^2} f''(\eta)$$

alakba írható. Mivel f'' folytonos az $[a,b]$-n, így korlátos is, azaz $|f''| \leq K$. Így a hiba abszolút értékben felülről becsülhető a következőképpen:

$$\left| -\frac{h^3}{12m^2} f''(\eta) \right| \leq \frac{h^3}{12m^2} K = \tilde{K} \Delta h^2,$$

ahol $\tilde{K} = \frac{h}{12} K$ szintén konstans. A $\tilde{K}\Delta h^2$ felső korlát $\Delta h \to 0$ esetén nullához tart, így a lépésköz finomításával tetszőleges pontosság elérhető.
Annak jellemzésére, hogy egy közelítő módszer hibája hogyan viselkedik egyre kisebb lépésközökre, bevezetjük a következő fogalmat.

5.2.6 Definíció. *Tegyük fel, hogy $g : K(0) \to \mathbb{R}$ olyan függvény, amelyhez van olyan $\tilde{p} \in \mathbb{N}^+$ szám és $K \in \mathbb{R}$ korlát, hogy a 0-hoz kellően közeli t pontokban*

$$|g(t)| \leq K|t|^{\tilde{p}}.$$

Jelölje p a legnagyobb ilyen tulajdonságú $\tilde{p}$ számot! Ekkor azt mondjuk, hogy a g függvény p-ed rendben tart 0-hoz a 0 pontban. Ezt jelölje $g(t) = \mathcal{O}(t^p)$ (ejtsd: "ordó t^p ".)

Például a $g(t) = t^2$ függvényre $|t^2| \leq |t|$ a $[-1,1]$-en, $|t^2| \leq |t|^2$ (mindenhol), de $|t^2| \leq K|t|^3$ már semmilyen K-ra nem teljesül a 0 egyetlen környezetén sem, ugyanis $\frac{|t^2|}{|t|^3} = \frac{1}{|t|}$ $t \to 0$-ra kinő a végtelenbe. Tehát a $g(t) = t^2$ függvény a 0-ban másodrendben tart 0-hoz.
Az összetett trapézszabály hibakorlátja szintén másodrendben tart nullához. Ez azt jelenti, hogy ha Δh-t q-ad részére csökkentjük, akkor a hibakorlát q^2-ed részére csökken. (Hiszen valamely Δh_1 lépésközhöz $\tilde{K}\Delta h_1^2$ hibakorlát tartozik, míg a $\Delta h_1/q$ lépésközhöz $\tilde{K}(\frac{\Delta h_1}{q})^2 = \frac{\tilde{K}\Delta h_1^2}{q^2}$.)

Összetett Simpson-formula

Legyen ismét $x_i = a + i\Delta h$ $(i = 0, 1 \ldots, m), \Delta h = \frac{b-a}{m}, y_i := \frac{x_{i-1}+x_i}{2}$ $(i = 1, \ldots, m)$. Az összetett Simpson-formula a következőképpen írható fel:

$$\int_a^b f \approx \sum_{i=1}^{m} \frac{\Delta h}{6}(f(x_{i-1}) + 4f(y_i) + f(x_i)) =$$

$$\frac{\Delta h}{6}(f(x_0) + f(x_m) + 2\sum_{i=1}^{m-1} f(x_i) + 4\sum_{i=1}^{m} f(y_i)) =: s_m(f).$$

Képlethibája $f \in C^4[a,b]$ függvényekre

$$\int_a^b f - s_m(f) = -\frac{h^5}{2880m^4} \cdot f^{IV}(\eta), \quad \eta \in [a,b] \text{ valamely pont.}$$

Mivel $|f^{IV}| \leq K$, ez a hiba abszolút értékben felülről becsülhető a

$$\frac{h^5}{2880m^4} K \leq \tilde{K} \Delta h^4$$

kifejezéssel. Ez a hibakorlát $\Delta h \to 0$ esetén negyedrendben tart nullához, így az összetett Simpson-formula magasabb rendben pontos közelítést nyújt, mint az összetett trapézformula.

5.2.3. Gauss-kvadratúrák

Eddig az interpolációs kvadratúra-formulák alkalmazásánál adottnak vettük az alappontokat. Láttuk, hogy a $\sum_{m=0}^{n} A_m f_m$ interpolációs formula pontos minden legfeljebb n-ed fokú polinomra. Kérdés: ha az f függvény értékeit tetszőleges alappontokban meg tudjuk határozni, akkor a kvadratúra-formula rendje növelhető-e az alappontok megfelelő megválasztásával? Vegyük észre, hogy abban a legegyszerűbb esetben, amikor csak egyetlen x_0 alappont van, és az illesztett interpolációs polinom konstans függvény, akkor a formula minden konstans függvényre pontos, de minden elsőfokú függvényre is pontos lesz, amennyiben x_0-nak a $c = \frac{a+b}{2}$ felezőpontot választjuk. (Ez eredményezi a középponti formulát, amelyről tudjuk, hogy másodrendű). Hasonlóképpen, a Simpson-formula három alappontra támaszkodik ($n = 2$), ám nem csak másodfokú polinomokra pontos, ahogyan azt három alappont esetén várnánk, hanem harmadfokúakra is. Érdekes kérdés az, hogy lehetne-e még magasabb rendet elérni a három pont másfajta megválasztásával, valamint hogy általában mennyivel növelhető a rend.

Vizsgáljuk meg részletesebben azt az esetet, amikor két alappont van: x_0 és x_1, és az egyszerűség kedvéért legyen $[a,b] = [-1,1]$. Az interpolációs kvadratúraformula ekkor

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

ahol

$$A_0 = \int_{-1}^{1} l_0(x) \mathrm{d}x = \int_{-1}^{1} \frac{x - x_1}{x_0 - x_1} \mathrm{d}x = \frac{1}{x_0 - x_1} \left[\frac{x^2}{2} - x_1 x \right]_{-1}^{1} = \frac{2x_1}{x_1 - x_0},$$

illetve x_0 és x_1 felcserélésével

$$A_1 = \int_{-1}^{1} l_1(x)\mathrm{d}x = \frac{2x_0}{x_0 - x_1} = -\frac{2x_0}{x_1 - x_0}.$$

Mint tudjuk, a formula tetszőleges x_0 és $x_1 \in [-1, 1]$ esetén pontos minden $f \in P_1$ polinomra, de ha x_0-ra és x_1-re szabadon megválasztható paraméterként tekintünk, akkor lehetséges, hogy elérhető kettővel magasabb rend is. Nyilvánvalóan egy kvadratúraformula akkor és csak akkor pontos minden q-ad fokú polinomra, amikor pontos az $f(x) = x^l$ függvényekre $\forall l = 0, 1, \ldots, q$ esetén. Így a formula akkor lesz egy renddel jobb, ha pontos az $f(x) = x^2$ függvényre, és két renddel pedig akkor, ha még az $f(x) = x^3$ függvényre is pontos. Így két egyenletet tudunk felírni az ismeretlen x_0 és x_1 pontokra:
1.) A formula akkor lesz pontos az $f(x) = x^2$ függvényre, ha

$$\sum_{m=0}^{1} A_m (x_m)^2 = A_0 \cdot x_0^2 + A_1 \cdot x_1^2 = \int_{-1}^{1} x^2 \mathrm{d}x = \frac{2}{3},$$

azaz

$$\frac{2x_1}{x_1 - x_0} x_0^2 - \frac{2x_0}{x_1 - x_0} x_1^2 = \frac{2}{3},$$

amely egyenlőség az

$$(1) \quad x_0 x_1 = -\frac{1}{3}$$

egyenletre egyszerűsödik.
2.) Hasonlóan, a formula akkor lesz pontos az $f(x) = x^3$ függvényre, ha

$$\frac{2x_1}{x_1 - x_0} x_0^3 - \frac{2x_0}{x_1 - x_0} x_1^3 = \int_{-1}^{1} x^3 \mathrm{d}x = 0.$$

Ebből egyszerűsítések után az $x_0 + x_1 = 0$, vagyis az

$$(2) \quad x_0 = -x_1$$

összefüggést kapjuk. Az (1)(2) egyenletrendszernek egyetlen megoldása van:

$$x_0 = -\frac{1}{\sqrt{3}}, \quad x_1 = \frac{1}{\sqrt{3}}.$$

Ezekhez az alappontokhoz számítsuk ki az A_0 és A_1 súlyokat:

$$A_0 = \frac{2x_1}{x_1 - x_0} = 1, \quad A_1 = -\frac{2x_0}{x_1 - x_0} = 1.$$

Tehát a kapott kvadratúraformula, amely a $[-1,1]$ intervallumon pontos minden $f \in P_3$ függvényre, a következő:

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

Általában, mivel minden további fokszámra akkor lesz pontos a formula, ha az alappontok eleget tesznek egy újabb egyenlőségnek, és összesen $n+1$ darab alappont van, az interpolációs kvadratúraformula rendje legfeljebb annyival növelhető meg, ahány alappont van. Kiszámolható, hogy három alappont esetén az alábbi megválasztással nyerünk olyan formulát, amely nemcsak minden másodfokú polinomra, hanem minden ötödfokúra is pontos (tehát harmadrendű helyett hatodrendű):

$$x_0 = -\sqrt{\frac{3}{5}}, \quad x_0 = 0, \quad x_2 = \sqrt{\frac{3}{5}}.$$

A megfelelő súlyok: $A_0 = A_2 = \frac{5}{9}, A_1 = \frac{8}{9}$. (Nem meglepő, hogy nem a Simpson-formula alappontjait kaptuk.)

5.2.7 Megjegyzés. *Ha pedig az integrálandó függvény nem a $[-1,1]$, hanem egy tetszőleges $[a,b]$ intervallumon van értelmezve, akkor a függvényt a $[-1,1]$ intervallumra áttranszformálva már közvetlenül használhatjuk a fent levezetett Gauss-féle formulákat.*

6. fejezet

Numerikus deriválás

Függvények deriváltjának közelítő kiszámítására szükségünk lehet például akkor, ha a függvényt csak diszkrét pontokban ismerjük, esetleg ott is csak közelítőleg. A deriváltakat közelítő formulák a differenciálegyenletek numerikus megoldásánál is hasznosak.

Tegyük fel, hogy az $f : \mathbb{R} \to \mathbb{R}$, $f \in D(I)$ (I egy valós intervallum) függvényt az $x_0, x_1, \ldots, x_n \in I$ pontokban ismerjük, és a szomszédos pontok az egyszerűség kedvéért legyenek egymástól azonos h távolságra. Célünk az $f'(x_i)$ ill. $f''(x_i)$ érték közelítése az f ezen diszkrét pontokban felvett függvényértékei segítségével.

6.1. Az első derivált közelítése

A derivált legegyszerűbb közelítő formuláját közvetlenül a derivált definíciójából nyerhetjük. Ismeretes, hogy az $f : \mathbb{R} \to \mathbb{R}$ függvény $x_i \in \mathrm{int}D(f)$-beli deriváltján az

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

határértéket értjük. Ha tehát h kellően kicsi, akkor $f'(x_i)$ jól közelíthető a következő hányadossal:

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

Elnevezése: **jobb oldali (vagy haladó) differenciahányados**.

A x_i-től balra elhelyezkedő szomszédos rácspontot is használhatjuk, így jutunk a **bal oldali (vagy retrográd) differenciahányados**hoz:

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

A kettő számtani közepe újabb közelítő deriválási formulát eredményez, ez a **centrális (vagy központi) differenciahányados**:

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

Kérdés, mennyire jók ezek a közelítések. Ennek jellemzésére bevezetjük a közelítés rendjének a fogalmát:

6.1.1 Definíció. *Jelölje az f függvény x_i-beli valamelyik deriváltját Df, ennek közelítését pedig $\Delta f(h)$. Azt mondjuk, hogy az x_i pontban a közelítés rendje $p \in \mathbb{N}^+$, ha $|Df - \Delta f(h)| = \mathcal{O}(h^p)$. (Ekkor $\exists K > 0 : |Df - \Delta f(h)| \leq Kh^p$.)*

6.1.2 Állítás. *A haladó és a retrográd differencia is elsőrendű közelítése egy $f \in C^2(I)$ függvény első deriváltjának.*

Biz.: Az f függvényt x_i körül Taylor-sorba fejtve

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

ahol $\eta \in [x_i, x_i + h]$ valamely pont. Ebből

$$\Delta f_+ = f'(x_i) + \frac{h}{2} f''(\eta),$$

és így

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

ahol $K = \frac{|f''(\eta)|}{2}$. A retrográd sémára hasonlóan...

6.1.3 Állítás. *A központi differenciahányados másodrendű közelítése egy $f \in C^3(I)$ függvény első deriváltjának.*

Biz.: Szintén f x_i körüli sorfejtéséből adódik:

$$f(x_i + h) = f(x_i) + hf'(x_i) + \frac{h^2}{2!} f''(x_i) + \frac{h^3}{3!} f'''(\eta_1),$$

ahol $\eta_1 \in [x_i, x_i + h]$ valamely pont, és

$$f(x_i - h) = f(x_i) - hf'(x_i) + \frac{h^2}{2!} f''(x_i) - \frac{h^3}{3!} f'''(\eta_2),$$

ahol $\eta_2 \in [x_i - h, x_i]$ valamely pont,

$$\Rightarrow \Delta f_c = \frac{f(x_i + h) - f(x_i - h)}{2h} = f'(x_i) + \frac{h^2}{6} f'''(\eta),$$

ahol kihasználtuk, hogy mivel $f \in C^3(I)$, az f''' függvény az η_1 és η_2 helyen felvett értékek számtani közepét is felveszi valamely $\eta \in I$ pontban.

6.2. A második derivált közelítése

Ha a függvény első deriváltját tudjuk közelíteni, akkor ennek felhasználásával megbecsülhető a második derivált, hiszen definíció szerint

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

és így kis h esetén felírhatjuk pl. a jobb oldali különbségi hányadost az első derivált függvényre:

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

Ez a formula még közvetlenül ritkán használható, mert az alappontokban jellemzően csak magának az f függvénynek az értékeit ismerjük, az első deriváltakét nem. Ezért az itt szereplő első deriváltakat közelítsük pl. retrográd differenciákkal. Ebből az

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

közelítést nyerjük, amely a második derivált leggyakoribb véges különbséges közelítése. Elnevezése: centrális séma a második derivált közelítésére.

6.2.1 Állítás. *A $\Delta^2 f_c$ centrális séma másodrendű közelítése egy $f \in C^4(I)$ függvény második deriváltjának.*

Ha még több rácsponti értéket használunk, akkor az első, második vagy magasabb rendű deriváltakat még magasabb rendben közelíthetjük, erre gyakorlaton láthatunk példákat.

6.3. A lépéstávolság dilemmája

A gyakorlatban az $f(x_i)$ értékek általában hibával terheltek, számítógépen dolgozva például elkerülhetetlenül fellép a számábrázolási hiba. Ebből érdekes dolog adódik a h lépéstávolság megválasztásával kapcsolatban. A képletek azt sugallják, hogy minél kisebb a h lépéstávolság, annál kisebb a hiba. Ez azonban nem így van!

Tegyük fel, hogy $f'(x_i)$-t szeretnénk közelíteni a központi sémával, de az $f(x_{i-1})$ és $f(x_{i+1})$ pontos értéke helyett az $\tilde{f}_{i-1}$ ill. $\tilde{f}_{i+1}$ hibás értékekkel számolunk. Tegyük fel, hogy $f(x_{i+1}) = \tilde{f}_{i+1} + \varepsilon_{i+1}$ ill. $f(x_{i-1}) = \tilde{f}_{i-1} + \varepsilon_{i-1}$, ahol $|\varepsilon_{i-1}|, |\varepsilon_{i+1}| \leq \varepsilon > 0$. Ekkor

$$\Delta f_c = \frac{f_{i+1} - f_{i-1}}{2h} = \Delta \tilde{f}_c + \frac{\varepsilon_{i+1} - \varepsilon_{i-1}}{2h}.$$

A bal oldalon $\Delta f_c = f'(x_i) + \frac{h^2}{6} f'''(\eta)$, így

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

ahol $M_3 := \sup |f'''|$. A felső becslést adó függvény tehát

$$g(h) = \frac{h^2}{6} M_3 + \frac{\varepsilon}{h},$$

és itt kis h lépéstávolságra a második tag kinő. Következésképpen akkor lesz a legkisebb a hiba, ha h se nem túl nagy, se nem túl kicsi. Közelítőleg optimális h értéket úgy kaphatunk, hogy a $g(h)$ függvényt minimalizáljuk h-ban. Hol nulla a $g'(h)$ derivált?

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

amiből az optimális lépéstávolság $h_{\text{opt}} = \sqrt[3]{\frac{3\varepsilon}{M_3}}$.

7. fejezet

Közönséges differenciálegyenletek megoldása

Számos fizikai, kémiai, biológiai stb. folyamat leírásához elengedhetetlenül fontosak a közönséges ill. parciális differenciálegyenletek. Ebben a fejezetben a közönséges differenciálegyenletekre vonatkozó kezdetiérték-feladatok numerikus megoldásával foglalkozunk.

Tekintsük az

$$\begin{cases} y'(t) = f(t, y(t)) \quad t \in (t_0, T] \\ y(t_0) = y_0 \end{cases} \tag{7.1}$$

skaláris kezdetiérték-feladatot. A megoldás létezését és egyértelműségét feltételezzük. (Ezek biztosítottak például, ha f azonkívül, hogy folytonos, kielégíti a Lipschitz-feltételt.) Az ilyen feladatok pontos megoldását zárt alakban rendszerint nem tudjuk meghatározni, csak speciális jobb oldali f függvények esetén. Ezért a legtöbb esetben közelítő módszereket kell alkalmaznunk.

7.1. Véges különbséges módszerek

A (7.1) közelítő megoldására az egyik legelterjedtebb módszercsalád a véges különbséges módszerek, ezért mi is ezek ismertetésére szorítkozunk. (A második félévben részletesen lesz szó sokféle egyéb módszerről is.)

A véges különbséges módszerek alkalmazásakor a feladatot először **diszkretizáljuk**: a $[t_0, T]$ intervallumon definiáljuk az

$$\omega_\tau := \{t_n = t_0 + n\tau, n = 0, 1, \ldots, N, \tau = (T - t_0)/N\}$$

egyenlőközű rácsot. A pontos megoldásra (amely egy folytonos függvény) a rácspontokban akarunk közelítést nyerni. A közelítő megoldás tehát nem egy folytonos függvény, hanem

egy ún. **rácspontfüggvény** lesz. Jelölje y_n a megoldás t_n pontbeli közelítését, ahol $n = 0, 1, \ldots, N$. A véges különbséges módszerek úgy működnek, hogy a rácson időben előrefelé lépegetve határozzuk meg a közelítő megoldást a rácspontokban. Ismerkedjünk meg a legegyszerűbb módszerekkel!

7.1.1. Az explicit Euler-módszer

A kezdeti t_0 időpillanatban a megoldás a kezdeti feltételből ismert y_0 érték. Hogyan közelíthetjük a megoldás t_1-beli értékét? Fejtsük sorba az y pontos megoldást a t_0 körül: ha y kétszer folytonosan differenciálható, akkor

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

Az utolsó tagot elhagyva az

$$y(t_0 + \tau) \approx y_0 + f(t_0, y_0)\tau$$

közelítő egyenlőség adódik. A jobb oldali kifejezés lesz a t_1-beli numerikus megoldás, ezt jelölje y_1. Tehát

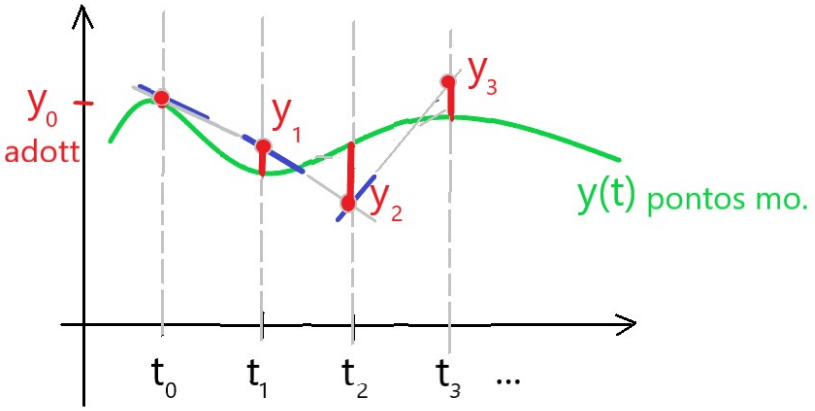
$$y_1 := y_0 + f(t_0, y_0)\tau$$

Hasonlóan léphetünk a $t_2, t_3, \ldots$ pontokra. Így az

$$y_{n+1} := y_n + f(t_n, y_n)\tau, \quad n = 0, 1, \ldots, N - 1$$

módszert nyerjük, amelynek neve **explicit Euler-módszer**.

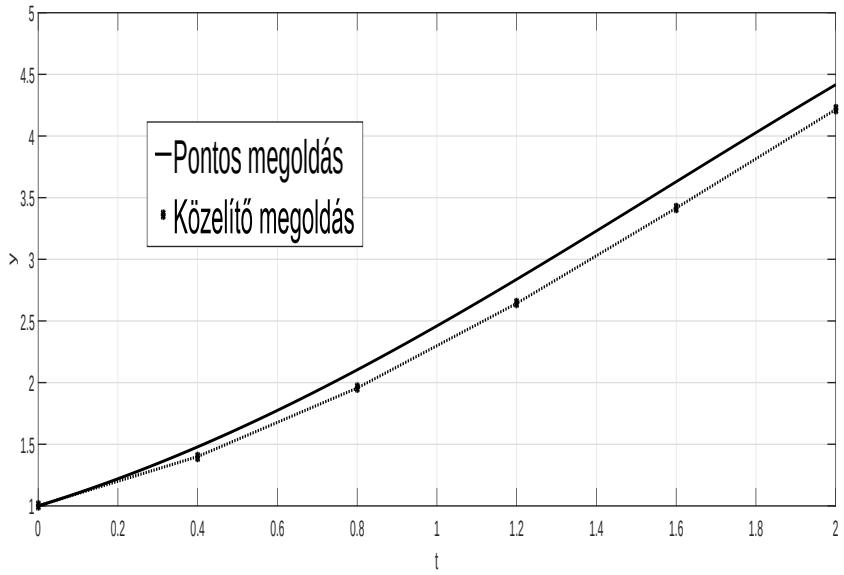
A 7.1.1 ábrán zöld folytonos vonal szemlélteti az (ismeretlen) pontos megoldást, ehhez szeretnénk közel kerülni. Az explicit Euler-módszerrel kapott numerikus megoldást a piros pöttyök mutatják. A kék pálcikák az adott ponton átmenő megoldásgörbe meredekségét jelzik, ezzel a meredekséggel lépünk a következő rácspontra. A piros szakaszok hossza az egymást követő lépésekben elkövetett hibát adja meg.



Az explicit Euler-módszerrel kapott közelítő megoldást szemlélteti a 7.1. ábra, ahol a módszert az

$$\begin{cases} y'(t) = \sin t + 1 \quad t \in (0, 2] \\ y(0) = 1 \end{cases} \tag{7.2}$$

feladatra alkalmaztuk.



7.1. ábra. A (7.2) feladat pontos megoldása (folytonos vonal) és az explicit Euler-módszerrel, $ht = 0.4$-es lépésközzel kapott közelítő megoldása (a szürke vonallal összekötött pontok) a $[0, 2]$ intervallumon.

7.1.2. Az implicit Euler-módszer

Ha az explicit Euler-módszer képletében az f függvényt nem a t_n, hanem a t_{n+1} pontban értékeljük ki, akkor

$$y_{n+1} = y_n + f(t_{n+1}, y_{n+1})\tau,\ i = 0, \ldots, N-1.$$

formulával definiált **implicit Euler-módszer**hez jutunk. Az elnevezést az indokolja, hogy ez egy implicit egyenlet a keresett y_{n+1} értékre, azaz ebből az egyenlőségből általános esetben nem fejezhető ki y_{n+1}. Jelölje $x := y_{n+1}$, ekkor a megoldandó egyenlet (nullára rendezve)

$$x - y_n - f(t_{n+1}, x)\tau = 0.$$

Erre alkalmazhatjuk pl. a már látott Newton-iterációt! Legyen

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

ennek keressük a zérushelyét. A Newton-módszer képletéből

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

ahol

$$F'(x) = 1 - \tau \cdot \partial_2 f(t_{n+1}, x).$$

A Newton-iteráció (különösen rendszer esetén) költségessé teszi a módszer alkalmazását.

7.1.3. Runge–Kutta-módszerek

Az explicit Euler-módszernél az n-edik lépésben a t_n pontból az $f(t_n, y_n)$ meredekséggel léptünk τ távolságon. Az $f(t_n, y_n)$ nem más, mint annak a megoldásgörbének a meredeksége, amely a (t_n, y_n) ponton átmegy. Hogyan nyerhetünk további módszereket? Például úgy, hogy több helyen is kiértékeljük az f függvényt, és ezeket a meredekségeket súlyozzuk. A legnépszerűbb Runge–Kutta-módszernél például négy meredekséget számolunk ki a következők szerint:

$$\begin{aligned}
k_1 &= f(t_n, y_n) \\
k_2 &= f(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}k_1) \\
k_3 &= f(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}k_2) \\
k_4 &= f(t_n + \tau, y_n + \tau k_3)
\end{aligned}$$

Ezután az y_{n+1} numerikus megoldást az alábbi módon nyerjük:

$$y_{n+1} = y_n + \tau \left(\frac{1}{6} k_1 + \frac{1}{3} k_2 + \frac{1}{3} k_3 + \frac{1}{6} k_4 \right)$$

A közönséges differenciálegyenletek numerikus megoldásáról bővebben a II. félévben lesz szó.

8. fejezet

Sajátérték-feladatok közelítő megoldása

Ismeretes, hogy az $A \in \mathbb{C}^{n\times n}$ mátrixnak a $\lambda \in \mathbb{C}$ szám sajátértéke, ha létezik olyan $v \in \mathbb{C}^n,\ v \neq 0$ vektor, amelyre $Av = \lambda v$. Ekkor a v vektort az A mátrix λ sajátértékhez tartozó sajátvektorának nevezzük. Az A sajátértékeinek a halmazát $\sigma(A)$ jelölje. A továbbiakban az egyszerűség kedvéért feltesszük, hogy A valós mátrix, továbbá a sajátértékek és sajátvektorok is valósak. Ez a helyzet például, ha az A mátrix szimmetrikus.

Tudjuk, hogy a λ szám pontosan akkor sajátértéke A-nak, ha megoldása A karakterisztikus egyenletének, vagyis ha $\det(A - \lambda I) = 0$. Ha tehát A sajátértékeit kell meghatározni, akkor erre elvileg direkt módszert jelentene a karakterisztikus egyenlet megoldása, de ez egy n-ed fokú egyenlet, amelyre $n > 4$ esetén nincs megoldóképlet. Ezért a gyakorlatban valamilyen iterációs módszert alkalmazunk.

Először nézzük meg, hogy ha ismerjük az egyik v sajátvektor $\tilde{v}$ közelítését, akkor hogyan közelíthetjük a megfelelő λ sajátértéket. Induljunk ki az $A\tilde{v} \approx \lambda\tilde{v}$ közelítő egyenlőségből. Koordinátánként kiírva az alábbi n darab egyenletet kapjuk λ-ra:

$$a_{i1}\tilde{v}_1 + a_{i2}\tilde{v}_2 + \ldots + a_{in}\tilde{v}_n = \lambda\tilde{v}_i, \quad i = 1, 2, \ldots, n.$$

Ezek közül bármelyik olyan egyenletből kifejezhető λ, amelynek jobb oldalán $\tilde{v}_i \neq 0$. (Mivel a sajátvektor nem nullvektor, így legalább egy ilyen biztosan létezik). Akár azt is megtehetjük, hogy külön-külön mindegyikből kifejezzük a λ számot, és az átlagukat vesszük.

Van azonban más lehetőség is. Szorozzuk meg skalárisan az $A\tilde{v} \approx \lambda\tilde{v}$ közelítő egyenlőség mindkét oldalát a $\tilde{v}$ vektorral:

$$(\tilde{v}, A\tilde{v}) = \lambda(\tilde{v}, \tilde{v}),$$

majd ebből fejezzük ki λ-t:

$$\lambda \approx \frac{(\tilde{v}, A\tilde{v})}{(\tilde{v}, \tilde{v})}$$

Vajon melyik a jobb módszer? Az alábbi tétel megmutatja, hogy az utóbbi a lehető legjobb, amivel próbálkozhatunk.

8.0.1 Állítás. *Legyen $x \in \mathbb{R}^n,\ x \neq 0$, és $A \in \mathbb{R}^{n\times n}$. Ekkor*

$$\min_{\alpha\in\mathbb{R}} \|Ax - \alpha x\|_2^2 = \|Ax - R(x)x\|_2^2,$$

ahol

$$R(x) = \frac{(x, Ax)}{(x, x)}.$$

(Tehát egy x vektor A mátrixszorosához az x vektor $R(x)$-szerese van legközelebb (a 2-es normában.)

Biz.:

$$\|Ax - \alpha x\|_2^2 = (Ax - \alpha x, Ax - \alpha x) = \alpha^2(x, x) - 2\alpha(x, Ax) + (Ax, Ax).$$

Ez α-ban másodfokú polinom, amelynek (x, x) főegyütthatója pozitív (mivel $x \neq 0$), ezért minimuma ott van, ahol az α szerinti deriváltja 0:

$$2\alpha(x, x) - 2(x, Ax) = 0 \Rightarrow \alpha = \frac{(x, Ax)}{(x, x)}.$$

Érdemes külön nevet adni az $R(x)$ szorzónak:

8.0.2 Definíció. *Legyen $x \in \mathbb{R}^n, x \neq 0, A \in \mathbb{R}^{n\times n}$. Ekkor az*

$$R(x) = \frac{(x, Ax)}{(x, x)}$$

számot az x vektorhoz tartozó **Rayleigh-hányados***nak nevezzük.*

(Nyilván, ha $\|x\|_2 = 1$, akkor $R(x) = (x, Ax)$.) Ha x eleve sajátvektor, akkor $R(x)x = Ax$, valamint ha van egy közelítésünk a sajátvektorra, akkor a Rayleigh-hányados segítségével kaphatunk jó becslést a sajátvektorhoz tartozó sajátértékre.

8.1. A hatványmódszer

Ezzel a módszerrel egy mátrix egyszeresen domináns sajátértékét (ha az létezik) és a hozzá tartozó sajátvektort tudjuk megtalálni. Legyen $A \in \mathbb{R}^{n\times n}$, és sajátértékei $\lambda_i,\ i =$

$1, 2, \ldots, n$, amelyekről feltesszük, hogy

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|$$

Tegyük fel, hogy A normális (azaz kommutál a transzponáltjával; ez teljesül például, ha A szimmetrikus). Ekkor A-nak van ortonormált sajátvektorrendszere: $v_1, v_2, \ldots, v_n$. Legyen $x \in \mathbb{R}^n$ olyan kezdővektor, amelynek van a v_1 vektor irányába eső komponense, tehát az

$$x = \alpha_1 v_1 + \alpha_2 v_2 + \ldots \alpha_n v_n$$

előállításban $\alpha_1 \neq 0$. Ekkor

$$A^k x = A^k(\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n) = \alpha_1 A^k v_1 + \ldots + \alpha_n A^k v_n =$$

$$\alpha_1 \lambda_1^k v_1 + \ldots + \alpha_n \lambda_n^k v_n = \lambda_1^k \left(\alpha_1 v_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k v_2 + \ldots + \alpha_n \left(\frac{\lambda_n}{\lambda_1} \right)^k v_n \right).$$

A zárójelben a másodiktól kezdve az összes tag nullához tart, ha $k \to \infty$. Ezért k növelésével $A^k x$ egyre inkább a v_1 sajátvektor irányába fog mutatni, azaz egyre jobb közelítése lesz egy λ_1-hez tartozó sajátvektornak.

Itt gondot jelent, hogy amennyiben λ_1 nem $+1$ vagy -1, akkor a kapott vektorok hossza vagy nullához, vagy ∞-hez tart, így numerikusan kezelhetetlenné válik a feladat (alul- vagy túlcsordulás lép fel a számítógépen). Ezért az egyes lépések után kapott vektort célszerű normálni. Ha a k. lépésben kapott normált vektort $y^{(k)}$ jelöli, akkor a Rayleigh-hányados alapján

$$\nu^{(k)} := (y^{(k)}, Ay^{(k)})$$

közelíti a hozzá tartozó sajátértéket.

8.1.1 Megjegyzés. *A hatványmódszernél feltettük, hogy a kezdővektornak van az első sajátvektor irányába eső komponense. De ezt hogyan biztosíthatjuk, ha nem ismerjük a sajátvektorokat? Szerencsére a gyakorlatban ez nem olyan fontos, mert a számábrázolási hibák miatt az iteráció során az iterációs vektornak lesz v_1 irányú komponense.*

8.2. Inverz iteráció

Eddig az egyszeresen domináns sajátérték és a hozzá tartozó sajátvektor meghatározásával foglalkoztunk. Kérdés: hogyan határozható meg a többi sajátérték?

Ismeretesek a következők:

1) Ha $\det(A) \neq 0$, és $\lambda \in \sigma(A)$, akkor $\frac{1}{\lambda} \in \sigma(A^{-1})$. Ez a következőképpen igazolható:

$$\lambda \in \sigma(A) \Leftrightarrow \exists x \neq 0 : Ax = \lambda x.$$

Szorozzuk meg ezt az egyenlőséget balról az A^{-1} mátrixszal:

$$A^{-1}Ax = \lambda A^{-1}x \Rightarrow \frac{1}{\lambda}x = A^{-1}x \Rightarrow \frac{1}{\lambda} \in \sigma(A^{-1}).$$

2) Ha μ nem egyezik meg A egyik λ_i sajátértékével sem, akkor $\det(A-\mu I) \neq 0$, így létezik $(A-\mu I)^{-1}$. Ha A-nak x sajátvektora a λ_i sajátértékkel, akkor

$$Ax = \lambda_i x \Rightarrow (A-\mu I)x = Ax - \mu Ix = \lambda_i x - \mu x = (\lambda_i - \mu)x,$$

ezért és az 1) pont alapján $(A-\mu I)^{-1}$ sajátértékei az $\frac{1}{\lambda_i - \mu}$, $i = 1, 2, \ldots, n$ számok.

8.2.1 Következmény. *Legyen $A \in \mathbb{R}^{n\times n}$ szimmetrikus mátrix, amelyre $\det A \neq 0$. Ha μ elég közel van valamelyik λ_j sajátértékhez, akkor $(A-\mu I)^{-1}$ domináns sajátértéke $\frac{1}{\lambda_j - \mu}$ lesz. Egy adott μ számhoz legközelebbi sajátértéket meg tudjuk tehát határozni, ha az $(A-\mu I)^{-1}$ mátrixszal hajtjuk végre a hatványmódszert. (Pl. ha $\mu = 0$, akkor a 0-hoz legközelebbi, tehát a legkisebb abszolút értékű sajátértéket tudjuk megtalálni.) A k-adik lépésben*

$$x^{(k)} = (A-\mu I)^{-1}y^{(k-1)} \Leftrightarrow (A-\mu I)x^{(k)} = y^{(k-1)},$$

azaz egy lineáris algebrai egyenletrendszert kell megoldani, ami költséges.

Az inverz iteráció lehetséges módosítása a Rayleigh-hányados iteráció: az inverz iteráció egy lépésében kapunk becslést a sajátvektorra, ebből a Rayleigh-hányados segítségével kaphatunk egy újabb sajátértékbecslést, az inverz iteráció következő lépésében μ helyett ezt az $R(y^{(k-1)})$ értéket használjuk.

Felhasznált irodalom:

Faragó István, Horváth Róbert: Numerikus módszerek, Typotex, 2016.
Fekete Imre: Alkalmazott analízis I. (diasor)