Matematika III

Fejezetek a numerikus analízisből

Írta: Dr. Gáspár Csaba

Bevezetés

Ez a jegyzet a Széchenyi István Egyetem mérnöki BSc-hallgatói számára készült. A jegyzet bevezetést ad a korszerű numerikus módszerek elméletébe és gyakorlatába. Hangsúlyozottan bevezető jellegű, tehát a teljességre megközelítőleg sem törekszik egyik fejezetében sem.

A jegyzet feltételezi a Matematika I. és II. tárgyak, azaz az elemi analízis, a lineáris algebra és a többváltozós függvények témaköréből korábban tanultak készségszerű ismeretét.

A jegyzetben érintett témakörök a következők:

- Lineáris egyenletrendszerek és mátrixfelbontások
- A legkisebb négyzetek módszere
- Interpoláció
- Közelítő integrálás
- Deriváltak közelítése

Mindegyik fejezet végén néhány gyakorló feladat található, részletes megoldásokkal együtt. Ez azonban nem helyettesíti a lényegesen bővebb feladatgyűjteményeket.

A jegyzetben alkalmazott jelölések:

- $\mathbf{N}$: a pozitív egész számok halmaza
- $\mathbf{R}$: a valós számok halmaza
- $\mathbf{R}^n$: a rendezett valós szám-n-esek vektortere
- $\mathbf{C}$: a komplex számok halmaza
- $\mathbf{M}_{n\times m}$: az n sorból és m oszlopból álló mátrixok vektortere

Fogjuk még alkalmazni az $\mathcal{O}(x)$ jelölést, mely egy olyan, x-től függő kifejezést jelent, melynek abszolút értéke felülről becsülhető $C \cdot x$-szel valamely alkalmas, x-től független $C \geq 0$ konstans mellett.

Tipikus példák: egy $N \times N$-es reguláris mátrixú lineáris egyenletrendszer megoldásának műveletigénye (ha a mátrix semmilyen speciális tulajdonsággal nem rendelkezik) $\mathcal{O}(N^3)$, azaz a megoldáshoz szükséges műveletek száma $\leq C \cdot N^3$, ahol a C szám N-től független. Ez jól mutatja a műveletigény viselkedését: ha az ismeretlenek száma megduplázódik, a megoldás műveletigénye az eredeti műveletigény mintegy nyolcszorosára nő.

Másik példa: az $[a, b]$ intervallumon értelmezett, elegendően sima f függvény integrálját egy ekvidisztáns, h lépésközű alappontrendszeren felépített összetett trapézformula $\mathcal{O}(h^2)$ hibával közelíti, azaz a pontos integrál és a trapézformula által adott közelítő integrál eltérése $\leq C \cdot h^2$, ahol a C szám h-től nem függ (de az f integrandusztól valamint az intervallumtól igen). Ez jól mutatja a közelítés pontosságát: ha a lépésközt felére, harmadára csökkentjük (azaz a részintervallumok számát kétszeresére, háromszorosára növeljük), akkor a közelítés hibája az eredetinek kb. negyedére, kilencedére csökken.

Javasoljuk, hogy a jegyzetben található módszereket, mintapéldákat és feladatokat az Olvasó valósítsa is meg – célszerűen MATLAB-ban. Numerikus módszereket igazán csak "működés közben" lehet megérteni, a programozás, a hibajavítás, az egyszerűbb, majd bonyolultabb példákon át való kipróbálás során.

Kérjük továbbá az Olvasókat, hogy a jegyzettel kapcsolatos észrevételeket, esetleges sajtóhibák felfedezését stb. tudassák a szerzővel a

gasparcs@sze.hu

email címen.

Eredményes felhasználást kíván a szerző:

Dr. Gáspár Csaba

1 Lineáris egyenletrendszerek és mátrixfelbontások

1.1 Motiváció

A gyakorlatban rengeteg probléma vezet lineáris (elsőfokú) egyenletrendszerekre. Ezek numerikus megoldása alapvető fontosságú. Hogy érzékeltessük a lényeges különbséget a "tiszta", "elméleti" és a numerikus matematikai problémák között, tekintsük a következő két példát:

Példa: Oldjuk meg az alábbi egyenletrendszert:

$$1000x + 999y = 1$$

$$999x + 998y = 1$$

A rendszer determinánsa $998000 - 999^2$, ami nem 0, így pontosan egy megoldás létezik. Könnyen ellenőrizhető, hogy ez: $x = 1$, $y = -1$.

Most tekintsük az alig-alig módosított egyenletrendszert:

$$1000x + 999y = 1$$

$$999x + 998y = 0.999$$

Intuitíve azt várjuk, hogy mivel az adatok csak "kicsit" változtak (1 ezreléknyit), azért a megoldás is csak "kicsit" fog különbözni az előzőtől. Ezzel szemben most a megoldás: $x = 0.001$, $y = 0$. Mivel pedig az egyenletek adatai (mátrixelemek és/vagy a jobb oldal elemei) gyakran mérésekből származnak, fontos, hogy ne csak a megoldást tudjuk kiszámítani, hanem annak hibáját is meg tudjuk becsülni. A fenti példa a legegyszerűbb példa a *rosszul kondícionált* egyenletrendszerekre, ahol a megoldás igen érzékeny az adatok megváltozására.

Példa: Mennyi időbe telhet (a leggyorsabb számítógépeken) egy 200×200-as mátrix determinánsának definíció szerinti kiszámítása (pl. az első sor szerinti kifejtéssel)?

Megoldás: Jelölje c_N egy $N \times N$-es mátrix determinánsának kiszámításakor csak a szükséges szorzások számát. A sor szerinti kifejtés alapján nyilván $c_N = N \cdot c_{N-1}$, innen:

$$c_N = N \cdot c_{N-1} = N \cdot (N-1) \cdot c_{N-2} = ... = N!$$

azaz $c_{200} = 200!$. Ez felfoghatatlanul nagy szám. Ha tekintünk egy hipotetikus párhuzamos működésű számítógépet, melynek mérete a Földével

egyezik, az egyes processzorok atomnyi méretűek, az információterjedés sebessége pedig a fénysebesség, a szükséges számítási idő meghaladná a ma feltételezett ősrobbanás óta eltelt időt.

Tehát nem elég tudni egy módszerről, hogy az elvileg helyes, az is szükséges, hogy a műveletigényét becsülni tudjuk, és az elfogadható felső korlát alá essék.

1.2 Lineáris egyenletrendszerek megoldhatósága

Most felidézzük a lineáris egyenletrendszerekről az alapképzésben már megismert legfontosabb megoldhatósági tételeket.

Legyen $A \in \mathbf{M}_{N\times N}$ adott négyzetes mátrix (elemei legyenek $a_{k,j}$ $(k, j = 1, 2, ..., N)$), $b \in \mathbf{R}^N$ adott vektor (elemei legyenek b_1, b_2,...,b_N), és tekintsük az

$$Ax = b \tag{1}$$

lineáris egyenletrendszert, ahol $x \in \mathbf{R}^N$ a keresett megoldásvektor. A fenti tömör jelölésmód egyenértékű az alábbi, hagyományos felírású lineáris egyenletrendszerrel:

$$a_{11}x_1 + a_{12}x_2 + ... + a_{1N}x_N = b_1$$

$$a_{21}x_1 + a_{22}x_2 + ... + a_{2N}x_N = b_2$$

$$.....$$

$$a_{N1}x_1 + a_{N2}x_2 + ... + a_{NN}x_N = b_N$$

Az egyenletrendszert *homogénnek* nevezzük, ha $b = \mathbf{0}$, azaz $b_1 = b_2 = ... = b_N = 0$. Nyilvánvaló, hogy ekkor $x_1 = x_2 = ... = x_N = 0$ mindig megoldás: ezt *triviális megoldásnak* nevezzük, míg a homogén egyenlet minden olyan megoldását, ahol legalább egy x_j zérustól különbözik, *nemtriviális megoldásnak* nevezzük.

A homogén egyenletek esetében a jellemző probléma az, hogy létezik-e nemtriviális megoldás, míg az inhomogén egyenlet esetén az a kérdés, hogy van-e egyáltalán megoldása, és ha igen, akkor hány.

Idézzük fel az (1) egyenletrendszer megoldhatóságát biztosító tételeket:

Tétel: Az $A \in \mathbf{M}_{N\times N}$ mátrix pontosan akkor reguláris, ha az $Ax = b$ egyenletnek minden jobb oldal mellett létezik megoldása. Ekkor a megoldás egyértelmű is (éspedig $A^{-1}b$-vel egyenlő).

Tétel: Az $A \in \mathbf{M}_{N\times N}$ mátrix pontosan akkor reguláris, ha az $Ax = \mathbf{0}$ homogén egyenletnek csak a triviális megoldása létezik. Másszóval, A pontosan akkor szinguláris, ha a homogén egyenletnek létezik nemtriviális megoldása. Ekkor pedig végtelen sok nemtriviális megoldás is létezik (bármely x megoldás esetén annak tetszőleges konstansszorosa is megoldás).

1.3 A Gauss-elimináció

Most áttekintjük – az alapképzésben megismert – egyik legfontosabb egyenletmegoldó algoritmus, a Gauss-elimináció működését.

Tekintsük az alábbi lineáris egyenletrendszert:

$$\begin{gathered}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ... + a_{1N}x_N = b_1 \\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + ... + a_{2N}x_N = b_2 \\
a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + ... + a_{3N}x_N = b_3 \\
..... \\
a_{N1}x_1 + a_{N2}x_2 + a_{N3}x_3 + ... + a_{NN}x_N = b_N
\end{gathered}$$

ahol most tegyük fel, hogy az A mátrix reguláris. A Gauss-elimináció (vagy kiküszöböléses módszer) lépései a következők:

1. Osszuk le az 1. egyenletet az a_{11} együtthatóval:

$$\begin{gathered}
x_1 + a'_{12}x_2 + a'_{13}x_3 + ... + a'_{1N}x_N = b'_1 \\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + ... + a_{2N}x_N = b_2 \\
a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + ... + a_{3N}x_N = b_3 \\
..... \\
a_{N1}x_1 + a_{N2}x_2 + a_{N3}x_3 + ... + a_{NN}x_N = b_N
\end{gathered}$$

2. Az 1. sor a_{k1}-szeresét vonjuk ki a k-adik sorból ($k = 2, 3, ..., N$), ezáltal kiküszöböljük x_1-et a 2.,3.,...,N. egyenletből. Eredményül az alábbi szerkezetű egyenletrendszerhez jutunk:

$$\begin{gathered}
x_1 + a'_{12}x_2 + a'_{13}x_3 + ... + a'_{1N}x_N = b'_1 \\
a'_{22}x_2 + a'_{23}x_3 + ... + a'_{2N}x_N = b'_2 \\
a'_{32}x_2 + a'_{33}x_3 + ... + a'_{3N}x_N = b'_3 \\
..... \\
a'_{N2}x_2 + a'_{N3}x_3 + ... + a'_{NN}x_N = b'_N
\end{gathered}$$

3. A 2.,3.,...,N. egyenlet már csak $(N-1)$ ismeretlent tartalmaz, így az 1., 2.

pont lépéseit megismételhetjük: x_2-t kiküszöböljük a 3.,4.,...,N. egyenletből, majd hasonlóan, x_3-at a 4.,5.,...,N. egyenletből, és így tovább. Végül a következő alakú egyenletrendszert kapjuk:

$$\begin{aligned} x_1 + \tilde{a}_{12}x_2 + \tilde{a}_{13}x_3 + ... + \tilde{a}_{1N}x_N &= \tilde{b}_1 \\ x_2 + \tilde{a}_{23}x_3 + ... + \tilde{a}_{2N}x_N &= \tilde{b}_2 \\ x_3 + ... + \tilde{a}_{3N}x_N &= \tilde{b}_3 \\ ..... & \\ x_N &= \tilde{b}_N \end{aligned}$$

4. Az utolsó egyenletből x_N máris ismert: visszahelyettesítve az $(N-1)$-edik egyenletbe, x_{N-1} számítható; x_N-et és x_{N-1}-et visszahelyettesítve az $(N-2)$-edik egyenletbe, x_{N-2} számítható, és így tovább, így az egyes ismeretleneket index szerint csökkenő sorrendben határozzuk meg. Ezt a műveletsort a Gauss-elimináció *visszahelyettesítési* részének, míg az előző lépéseket *eliminációs résznek* nevezzük.

Az algoritmust az alábbi egyszerű példán szemléltetjük:

Példa: Oldjuk meg az alábbi egyenletrendszert:

$$\begin{array}{rcrcrcr} 2x_1 & - & 6x_2 & + & 10x_3 & = & -12 \\ 2x_1 & - & 5x_2 & + & 3x_3 & = & -4 \\ 3x_1 & - & 2x_2 & + & x_3 & = & 3 \end{array}$$

Megoldás: Osszuk le az 1. egyenletet 2-vel:

$$\begin{array}{rcrcrcr} x_1 & - & 3x_2 & + & 5x_3 & = & -6 \\ 2x_1 & - & 5x_2 & + & 3x_3 & = & -4 \\ 3x_1 & - & 2x_2 & + & x_3 & = & 3 \end{array}$$

Az 1. egyenlet 2-szeresét vonjuk ki a 2. egyenletből, majd az 1. egyenlet 3-szorosát vonjuk ki a 3. egyenletből, ezzel a 2. és 3. egyenletből kiküszöböljük x_1-et:

$$\begin{array}{rcrcrcr} x_1 & - & 3x_2 & + & 5x_3 & = & -6 \\ & & x_2 & - & 7x_3 & = & 8 \\ & & 7x_2 & - & 14x_3 & = & 21 \end{array}$$

A 2. egyenletet már nem kell x_2 együtthatójával leosztani, lévén az 1-gyel egyenlő. Vonjuk ki a 2. egyenlet 7-szeresét a 3. egyenletből, ezzel a 3.

egyenletből x_2-t is kiküszöböltük:

$$\begin{array}{rcrcrcr} x_1 & - & 3x_2 & + & 5x_3 & = & -6 \\ & & x_2 & - & 7x_3 & = & 8 \\ & & & & 35x_3 & = & -35 \end{array}$$

Elosztva a 3. egyenletet 35-tel, az eliminációs részt befejeztük:

$$\begin{array}{rcrcrcr} x_1 & - & 3x_2 & + & 5x_3 & = & -6 \\ & & x_2 & - & 7x_3 & = & 8 \\ & & & & x_3 & = & -1 \end{array}$$

Az utolsó egyenletből x_3 már ki van számítva. Visszahelyettesítve a 2. egyenletbe, innen x_2 is számítható (ugyanide jutunk, ha a 3. egyenlet 7-szeresét hozzáadjuk a 2. egyenlethez):

$$\begin{array}{rcrcrcr} x_1 & - & 3x_2 & + & 5x_3 & = & -6 \\ & & x_2 & & & = & 1 \\ & & & & x_3 & = & -1 \end{array}$$

Végül, x_2-t és x_3-t az 1. egyenletbe helyettesítve vissza, x_1 is számítható:

$$\begin{array}{rcrcr} x_1 & & & = & 2 \\ & x_2 & & = & 1 \\ & & x_3 & = & -1 \end{array}$$

Ezzel az egyenletrendszer megoldását előállítottuk. Visszahelyettesítéssel meggyőződhetünk róla, hogy az így nyert megoldás valóban kielégíti az eredeti egyenletrendszert.

Vegyük észre, hogy a számítás végrehajtásához az x_1, x_2, x_3 szimbólumokat és az egyenlőségjeleket újra meg újra leírni felesleges: a számításokat voltaképpen csak az együtthatókon, azok mátrixán hajtjuk végre. Így a fenti számítási lépések az alábbi tömör formába írhatók (a mátrix utolsó oszlopa előtti függőleges vonal csak a jobb áttekinthetőséget szolgálja):

$$\left(\begin{array}{ccc|c} 2 & -6 & 10 & -12 \\ 2 & -5 & 3 & -4 \\ 3 & -2 & 1 & 3 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & -3 & 5 & -6 \\ 2 & -5 & 3 & -4 \\ 3 & -2 & 1 & 3 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|c} 1 & -3 & 5 & -6 \\ 0 & 1 & -7 & 8 \\ 0 & 7 & -14 & 21 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & -3 & 5 & -6 \\ 0 & 1 & -7 & 8 \\ 0 & 0 & 35 & -35 \end{array}\right) \to$$

$$\rightarrow \left(\begin{array}{ccc|c} 1 & -3 & 5 & -6 \\ 0 & 1 & -7 & 8 \\ 0 & 0 & 1 & -1 \end{array}\right) \rightarrow \left(\begin{array}{ccc|c} 1 & -3 & 5 & -6 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{array}\right) \rightarrow$$

$$\rightarrow \left(\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{array}\right)$$

Megmutatható, hogy a Gauss-elimináció műveletigénye $\mathcal{O}(N^3)$, ami azt mutatja, hogy numerikus szempontból a Gauss-elimináció nem "olcsó": ha az ismeretlenek száma duplájára nő, akkor a szükséges műveletszám kb. *nyolcszorosára* emelkedik.

Az algoritmust ebben a formában nem mindig lehet végrehajtani, ui. lehetséges, hogy valamelyik együttható, mellyel osztanunk kéne, 0-val egyenlő. Legyen pl. $a_{11} = 0$. Ekkor az első és valamelyik későbbi egyenlet cseréjével elérhető, hogy az új egyenletrendszer első egyenletében x_1 együtthatója ne legyen 0, ellenkező esetben a mátrix első oszlopa csupa 0-ból állna, de ekkor a mátrix szinguláris volna, kiinduló feltételünkkel ellentétben. Ugyanez áll az elimináció további lépéseiben fellépő (egyre kisebb méretű) egyenletrendszerekre. A numerikus számítás pontosságának szempontjából az a célszerű, hogy a együtthatók, melyekkel leosztjuk az egyenleteket (az ún *főegyütthatók*), abszolút értékben minél nagyobbak legyenek (hogy az osztás számítási hibája minél kisebb legyen). Ezért az egyenletek cseréjekor célszerű az aktuális, mondjuk k-adik egyenletet azzal a későbbi pl. r-edik egyenlettel felcserélni, melyre $|a_{rk}|$ a lehető legnagyobb ($r = k, k+1, ..., N$) még akkor is, ha $a_{kk} \neq 0$. Ezt a megoldási stratégiát *részleges főelemkiválasztásnak* nevezzük, és ez már minden reguláris A mátrix esetén működik. Valamivel több számítási munkával jár, de még nagyobb pontosságot biztosít a *teljes főelemkiválasztás*, amikor a k-adik egyenlettel való eliminációs lépéskor az összes hátralévő $|a_{pq}|$ érték maximumát keressük ($p, q = k, k+1, ..., N$), és ekkor nemcsak az egyenleteket cseréljük meg, hanem az ismeretlenek sorrendjét is megváltoztatjuk, hogy a főegyüttható az imént meghatározott maximális abszolút értékű elem legyen.

A Gauss-elimináció egy válfaja a *Gauss–Jordan-elimináció*, amikor az aktuális pl. k-adik egyenlet segítségével nemcsak a későbbi egyenletekből küszöböljük ki a k-adik ismeretlent, hanem a *megelőzőekből* is. Így a visszahelyettesítési lépések elmaradnak, és az elimináció befejeztével azonnal nyerjük az ismeretlenek értékeit. (Egyszerűsége ellenére a Gauss–Jordan-elimináció műveletigénye nagyobb a Gauss-elimináció műveletigényénél).

Példa: Tekintsük az előző példa egyenletrendszerét. Az algoritmus első két

lépése egyezik a Gauss-elimináció első két lépésével, eltérés csak a 3. lépéstől van:

$$\left(\begin{array}{ccc|c} 2 & -6 & 10 & -12 \\ 2 & -5 & 3 & -4 \\ 3 & -2 & 1 & 3 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & -3 & 5 & -6 \\ 2 & -5 & 3 & -4 \\ 3 & -2 & 1 & 3 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|c} 1 & -3 & 5 & -6 \\ 0 & 1 & -7 & 8 \\ 0 & 7 & -14 & 21 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & 0 & -16 & 18 \\ 0 & 1 & -7 & 8 \\ 0 & 0 & 35 & -35 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|c} 1 & 0 & -16 & 18 \\ 0 & 1 & -7 & 8 \\ 0 & 0 & 1 & -1 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{array}\right)$$

A Gauss-elimináció szinguláris mátrixú egyenletrendszerek megoldására is alkalmas. Ekkor az a jellemző, hogy az elimináció valamelyik lépésében az egyik egyenlet *összes* együtthatója zérussá válik. Amennyiben az illető egyenlet jobb oldala nem zérus, akkor az egyenletrendszernek nincs megoldása; ha a jobb oldal is zérus, akkor van megoldás, sőt, ekkor mindig *végtelen sok* megoldás van. Ekkor ui. valamelyik ismeretlen (esetleg több is) szabadon megválasztható, a többi pedig ezek függvényében fejezhető ki.

Az elmondottakat egy példán szemléltetjük:

Példa: Oldjuk meg a következő egyenletrendszert:

$$\begin{array}{rcrcrcl} x_1 & - & 2x_2 & + & x_3 & = & 1 \\ -2x_1 & + & x_2 & + & x_3 & = & 4 \\ x_1 & + & x_2 & - & 2x_3 & = & 1 \end{array}$$

Megoldás: A Gauss-elimináció lépéseit az előző példákban megismert tömör jelölésmóddal írjuk le:

$$\left(\begin{array}{ccc|c} 1 & -2 & 1 & 1 \\ -2 & 1 & 1 & 4 \\ 1 & 1 & -2 & 1 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & -2 & 1 & 1 \\ 0 & -3 & 3 & 6 \\ 0 & 3 & -3 & 0 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|c} 1 & -2 & 1 & 1 \\ 0 & 1 & -1 & -2 \\ 0 & 3 & -3 & 0 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & -2 & 1 & 1 \\ 0 & 1 & -1 & -2 \\ 0 & 0 & 0 & 6 \end{array}\right)$$

Az utolsó egyenlet együtthatói mind 0-val lettek egyenlők, de a jobb oldal nem zérus. Ez ellentmondás, így az egyenletrendszernek nincs megoldása.

Ha viszont a megfelelő homogén egyenletet tekintjük:

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

akkor már tudjuk, hogy van nemtriviális megoldás, hiszen a mátrix szinguláris (ezt akár a determináns zérus voltából láthatjuk, akár onnan, hogy ha a mátrix reguláris volna, az előző egyenletrendszernek is lenne, éspedig egyetlen megoldása). Lássuk, hogyan működik a Gauss-elimináció ebben az esetben:

$$\left(\begin{array}{ccc|c} 1 & -2 & 1 & 0 \\ -2 & 1 & 1 & 0 \\ 1 & 1 & -2 & 0 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & -2 & 1 & 0 \\ 0 & -3 & 3 & 0 \\ 0 & 3 & -3 & 0 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|c} 1 & -2 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 3 & -3 & 0 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & -2 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Az utolsó egyenlet a semmitmondó $0 = 0$ egyenlőséggé egyszerűsödött. Valamelyik, célszerűen az utolsó ismeretlent így tetszőlegesen megválaszthatjuk: $x_3 := t$, ahol $t \in \mathbf{R}$ tetszőleges szám. Ezt beírva a 3. egyenlet helyére, a visszahelyettesítések már nehézség nélkül elvégezhetők:

$$\left(\begin{array}{ccc|c} 1 & -2 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & t \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & -2 & 0 & -t \\ 0 & 1 & 0 & t \\ 0 & 0 & 1 & t \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|c} 1 & 0 & 0 & t \\ 0 & 1 & 0 & t \\ 0 & 0 & 1 & t \end{array}\right)$$

Tehát végtelen sok nemtriviális megoldás van, és ezek általános alakja: $x_1 = t$, $x_2 = t$, $x_3 = t$.

Végül megmutatjuk, hogyan használható a Gauss-elimináció mátrixinvertálásra. Legyen $A \in \mathbf{M}_{N\times N}$ egy reguláris mátrix. Ekkor érvényes az

$$AA^{-1} = I$$

mátrixegyenlőség. Jelölje az egyelőre ismeretlen A^{-1} inverz mátrix oszlopait a_1, a_2, ... , a_N, az I egységmátrix oszlopait pedig e_1, e_2, ... ,e_N (ezek épp a

standard bázis elemei $\mathbf{R}^N$-ben):

$$A \cdot \left(\begin{array}{c|c|c|c} a_1 & a_2 & ... & a_N \end{array} \right) = \left(\begin{array}{c|c|c|c} e_1 & e_2 & ... & e_N \end{array} \right)$$

A mátrixszorzás definíciója értelmében ez a mátrixegyenlőség ekvivalens N db vektoregyenlőséggel, éspedig:

$$Aa_k = e_k \qquad (k = 1, 2, ..., N)$$

Azt kaptuk tehát, hogy a fenti N db egyenletrendszert megoldva, a megoldásvektorokból mint oszlopokból összeállított mátrix épp az eredeti A mátrix inverzével egyezik.

Tehát egy mátrixinverzióhoz N db speciális jobb oldalú, de azonos mátrixú egyenletrendszert kell megoldani. Ez történhet egyidejűleg is, mivel az eliminációs lépésekkel egyidőben a jobb oldalakon végrehajtandó műveleteket egyszerre *mindegyik* jobb oldallal megtehetjük. Az eljárás a fentiekben használt tömör jelölésmóddal igen szemléletes: az elimináció elején a bal oldali részmátrix az eredeti mátrix, a jobb oldali pedig az egységmátrix, míg az algoritmus befejeztével a bal oldali részmátrixból egységmátrix lesz, ekkor a jobb oldali részmátrix az inverz mátrixszal lesz egyenlő.

Példa: Számítsuk ki az alábbi mátrix inverzét:

$$A := \left(\begin{array}{ccc} -3 & -2 & 0 \\ 0 & 3 & 2 \\ -2 & 0 & 1 \end{array} \right)$$

Megoldás: Az algoritmus lépései az eddigiekben alkalmazott tömör jelölésmóddal:

$$\left(\begin{array}{ccc|ccc} -3 & -2 & 0 & 1 & 0 & 0 \\ 0 & 3 & 2 & 0 & 1 & 0 \\ -2 & 0 & 1 & 0 & 0 & 1 \end{array} \right) \rightarrow$$

$$\rightarrow \left(\begin{array}{ccc|ccc} 1 & 2/3 & 0 & -1/3 & 0 & 0 \\ 0 & 3 & 2 & 0 & 1 & 0 \\ -2 & 0 & 1 & 0 & 0 & 1 \end{array} \right) \rightarrow$$

$$\rightarrow \left(\begin{array}{ccc|ccc} 1 & 2/3 & 0 & -1/3 & 0 & 0 \\ 0 & 3 & 2 & 0 & 1 & 0 \\ 0 & 4/3 & 1 & -2/3 & 0 & 1 \end{array} \right) \rightarrow$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 2/3 & 0 & -1/3 & 0 & 0 \\ 0 & 1 & 2/3 & 0 & 1/3 & 0 \\ 0 & 4/3 & 1 & -2/3 & 0 & 1 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 2/3 & 0 & -1/3 & 0 & 0 \\ 0 & 1 & 2/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1/9 & -2/3 & -4/9 & 1 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 2/3 & 0 & -1/3 & 0 & 0 \\ 0 & 1 & 2/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1 & -6 & -4 & 9 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 2/3 & 0 & -1/3 & 0 & 0 \\ 0 & 1 & 0 & 4 & 3 & -6 \\ 0 & 0 & 1 & -6 & -4 & 9 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -3 & -2 & 4 \\ 0 & 1 & 0 & 4 & 3 & -6 \\ 0 & 0 & 1 & -6 & -4 & 9 \end{array}\right)$$

Az inverz mátrix tehát:

$$A^{-1} = \begin{pmatrix} -3 & -2 & 4 \\ 4 & 3 & -6 \\ -6 & -4 & 9 \end{pmatrix}$$

amit az $A^{-1}A$ mátrixszorzás elvégzésével könnyen ellenőrizhetünk.

1.4 Mátrixok LU-felbontása

Legyen $A = [a_{kj}] \in \mathbf{M}_{N\times N}$ reguláris mátrix, mellyel a Gauss-elimináció elvégezhető (azaz nincs szükség sorcserékre).

Tétel: Az A mátrix egyértelműen előáll

$$A = LU$$

alakban, ahol L *normált alsó háromszögmátrix* (azaz $L_{kk} = 1$, és $L_{kj} = 0$, ha $j > k$) U pedig *felső háromszögmátrix* (azaz $U_{kj} = 0$, ha $j < k$).

Az LU-felbontás gyakorlati haszna a következő. Ha megoldandó egy $Ax = b$ egyenletrendszer esetleg *sok különböző* b *jobb oldallal*, akkor ezek helyett elég *egyszer* végrehajtani az LU-felbontást; ezekután most már elegendő

megoldani az $L(Ux) = b$ egyenleteket, azaz az $Ly = b$, $Ux = y$ egyenletpárokat (minden b jobb oldali vektorral), ami numerikusan sokkal olcsóbb, tekintve, hogy itt csak előre- és visszahelyettesítéseket kell végrehajtani (kiküszöbölési lépéseket már nem).

Igazolható, hogy az LU-felbontás műveletigénye $\mathcal{O}(N^3)$, míg egy-egy fenti alakú, $Ly = b$, $Ux = y$ egyenletpár megoldásának műveletigénye csak legfeljebb $\mathcal{O}(N^2)$ (speciális esetekben ennél kevesebb is lehet, pl. ha L és U ritka mátrixok).

A tétel és a következő algoritmus helyességének bizonyítása nem nehéz, de hosszadalmas; több apróbb állítás eredménye, így ettől eltekintünk.

A felbontás végrehajtása Gauss-eliminációval: A k-adik sorral való eliminációkor vonjuk ki a k-adik sor $l_{m,k} := a_{m,k}/a_{k,k}$-szorosát az m-edik sorból $(m = k+1, ..., N)$. Ezekből az $l_{m,k}$ számokból az L mátrix összeállítható:

$$L = \begin{pmatrix} 1 & 0 & 0 & ... & 0 \\ l_{2,1} & 1 & 0 & ... & 0 \\ l_{3,1} & l_{3,2} & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ l_{N,1} & l_{N,2} & l_{N,3} & ... & 1 \end{pmatrix}$$

Az eliminációs lépések után pedig az eredeti A mátrixból épp U-t kapjuk.

Példa: Határozzuk meg az

$$A = \begin{pmatrix} 2 & -6 & 10 \\ 2 & -5 & 3 \\ 3 & -2 & 1 \end{pmatrix}$$

mátrix LU-felbontását!

Megoldás:

$$\begin{pmatrix} 2 & -6 & 10 \\ 2 & -5 & 3 \\ 3 & -2 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0 & 0 \\ . & 1 & 0 \\ . & . & 1 \end{pmatrix}$$

$$\begin{pmatrix} 2 & -6 & 10 \\ 0 & 1 & -7 \\ 3 & -2 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ . & . & 1 \end{pmatrix}$$

$$\begin{pmatrix} 2 & -6 & 10 \\ 0 & 1 & -7 \\ 0 & 7 & -14 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \frac{3}{2} & . & 1 \end{pmatrix}$$

$$U = \begin{pmatrix} 2 & -6 & 10 \\ 0 & 1 & -7 \\ 0 & 0 & 35 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \frac{3}{2} & 7 & 1 \end{pmatrix} = L$$

Egy másik példa, amelyre két megoldást is mutatunk; az első lényegében egyezik az előző példa eljárásával, míg a második talán még egyszerűbb. Határozzuk meg az alábbi mátrix LU-felbontását:

$$A := \begin{pmatrix} 2 & -2 & 4 \\ -2 & -1 & -1 \\ 4 & 1 & 3 \end{pmatrix}$$

Megoldás: Mindenekelőtt L-et a már ismert elemeivel kitöltjük: a főátlóban 1-esek állnak, a főátló felett zérusok:

$$A = \begin{pmatrix} 2 & -2 & 4 \\ -2 & -1 & -1 \\ 4 & 1 & 3 \end{pmatrix} \qquad L = \begin{pmatrix} 1 & 0 & 0 \\ . & 1 & 0 \\ . & . & 1 \end{pmatrix}$$

Majd elkezdjük az eliminációt: A első sorának $\frac{-2}{2}$-szeresét kivonjuk a második, $\frac{4}{2}$-szeresét a harmadik sorból, e szorzókat ($\frac{-2}{2}$ és $\frac{4}{2}$) pedig beírjuk L első oszlopának második ill. harmadik helyére:

$$\begin{pmatrix} 2 & -2 & 4 \\ 0 & -3 & 3 \\ 0 & 3 & -5 \end{pmatrix} \qquad L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & . & 1 \end{pmatrix}$$

Folytatjuk az eliminációt: a mátrix második sorának $\frac{3}{-3}$-szorosát kivonjuk a harmadik sorból, a $\frac{3}{-3}$ szorzót pedig beírjuk L második oszlopának harmadik helyére. Ezzel az eredeti mátrixból felső háromszög mátrixot kaptunk (ez lesz a felbontás U mátrixa), és egyidejűleg az L-et is megkaptuk.

$$U = \begin{pmatrix} 2 & -2 & 4 \\ 0 & -3 & 3 \\ 0 & 0 & 2 \end{pmatrix} \qquad L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix}$$

(Ellenőrizzük az eredményt az LU szorzat közvetlen kiszámításával!)

Más megoldás: Töltsük ki az L és U mátrixok előre ismert elemeit:

- L főátlójában csupa 1-esek állnak;

- L főátlója felett csupa 0-k állnak;
- L első oszlopának további elemei: a_{21}/a_{11}, a_{31}/a_{11}, a_{41}/a_{11}, ...
- U főátlója alatt csupa 0-k állnak;
- U első sora mindig egyezik A első sorával.

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & \ell_{32} & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & -2 & 4 \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}$$

Ezekután L k-adik sorát U j-edik oszlopával szorozva az eredeti A mátrix a_{kj}-edik elemét kell kapjuk. A sor-oszlop-szorzások sorrendjének ügyes megválasztásával elérhető, hogy minden így nyert egyenletben csak egy ismeretlen szerepeljen, amit aztán mindjárt be is írhatunk L ill. U megfelelő helyére.

Például, L második sorát U második oszlopával szorozva: $(-1)\cdot(-2)+1\cdot u_{22}+0=-1$, ahonnan u_{22} kifejezhető: $u_{22}=-3$:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & \ell_{32} & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & -2 & 4 \\ 0 & -3 & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}$$

Most L harmadik sorát szorozzuk U második oszlopával: $2\cdot(-2)+\ell_{32}\cdot(-3)+0=-1$, ahonnan ℓ_{32} kifejezhető: $\ell_{32}=-1$:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & -2 & 4 \\ 0 & -3 & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}$$

Ezzel L-et már ki is számítottuk. L második sorát szorozva U harmadik oszlopával: $(-1)\cdot 4+1\cdot u_{23}+0=-1$, ahonnan u_{23} kifejezhető: $u_{23}=3$:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & -2 & 4 \\ 0 & -3 & 3 \\ 0 & 0 & u_{33} \end{pmatrix}$$

Végül L harmadik sorát szorozva U harmadik oszlopával: $2\cdot 4+(-1)\cdot 3+1\cdot u_{33}=3$, ahonnan u_{33} kifejezhető: $u_{33}=-2$:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & -2 & 4 \\ 0 & -3 & 3 \\ 0 & 0 & -2 \end{pmatrix}$$

Ezzel a kívánt LU-felbontást megkaptuk.

Megjegyzés: Az LU-felbontás a determináns kiszámítására is ad egy eljárást, ami numerikusan sokkal olcsóbb, mint a definíció alkalmazása (sor vagy oszlop szerinti kifejtés). Ha ui. $A = LU$, akkor $\det(A) = \det(L) \cdot \det(U)$. A jobb oldali determinánsok pedig rendkívül egyszerűen számíthatók, épp a diagonálelemek szorzatával egyenlők (miért?). Ezért $\det(L) = 1$, tehát:

$$\det(A) = \det(U) = U_{11} \cdot U_{22} \cdot ... \cdot U_{NN}.$$

1.5 Feladatok

1. Oldjuk meg Gauss-eliminációval az alábbi egyenletrendszert:

$$\begin{array}{rcrcrcr} x & + & 4y & + & 2z & = & 5 \\ -3x & + & 2y & + & z & = & -1 \\ 4x & - & y & - & z & = & 2 \end{array}$$

Megoldás: A számítás sémája a következő:

$$\left(\begin{array}{rrr|r} 1 & 4 & 2 & 5 \\ -3 & 2 & 1 & -1 \\ 4 & -1 & -1 & 2 \end{array}\right) \to \left(\begin{array}{rrr|r} 1 & 4 & 2 & 5 \\ 0 & 14 & 7 & 14 \\ 0 & -17 & -9 & -18 \end{array}\right) \to$$

$$\to \left(\begin{array}{rrr|r} 1 & 4 & 2 & 5 \\ 0 & 1 & \frac{1}{2} & 1 \\ 0 & -17 & -9 & -18 \end{array}\right) \to \left(\begin{array}{rrr|r} 1 & 4 & 2 & 5 \\ 0 & 1 & \frac{1}{2} & 1 \\ 0 & 0 & -\frac{1}{2} & -1 \end{array}\right) \to$$

$$\to \left(\begin{array}{rrr|r} 1 & 4 & 2 & 5 \\ 0 & 1 & \frac{1}{2} & 1 \\ 0 & 0 & 1 & 2 \end{array}\right) \to \left(\begin{array}{rrr|r} 1 & 4 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{array}\right) \to$$

$$\to \left(\begin{array}{rrr|r} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{array}\right)$$

A megoldás tehát: $x = 1$, $y = 0$, $z = 2$.

2. Oldjuk meg a következő lineáris egyenletrendszert Gauss-Jordan-eliminációval:

$$\begin{array}{rcrcrcr} 2x_1 & - & 3x_2 & + & x_3 & = & -1 \\ x_1 & - & 2x_2 & - & 3x_3 & = & 6 \\ 2x_1 & + & x_2 & + & x_3 & = & 3 \end{array}$$

Megoldás: Felcserélve az 1. és 2. sorokat, majd az (új) 1. sor (-2)-szeresét hozzáadva a 2., majd a 3. sorhoz:

$$\left(\begin{array}{rrr|r} 2 & -3 & 1 & -1 \\ 1 & -2 & -3 & 6 \\ 2 & 1 & 1 & 3 \end{array}\right) \to \left(\begin{array}{rrr|r} 1 & -2 & -3 & 6 \\ 2 & -3 & 1 & -1 \\ 2 & 1 & 1 & 3 \end{array}\right) \to$$

$$\left(\begin{array}{ccc|c} 1 & -2 & -3 & 6 \\ 0 & 1 & 7 & -13 \\ 0 & 5 & 7 & -9 \end{array}\right)$$

A második sor segítségével elimin——áljunk lefelé:

$$\left(\begin{array}{ccc|c} 1 & -2 & -3 & 6 \\ 0 & 1 & 7 & -13 \\ 0 & 0 & -28 & 56 \end{array}\right)$$

Folytassuk az eliminálást felfelé is. (Ebben különbözik a Gauss-Jordan-algoritmus a Gauss-eliminációtól.) A második sor kétszeresét az első sorhoz adva:

$$\left(\begin{array}{ccc|c} 1 & 0 & 11 & -20 \\ 0 & 1 & 7 & -13 \\ 0 & 0 & -28 & 56 \end{array}\right) \quad \rightarrow \quad \left(\begin{array}{ccc|c} 1 & 0 & 11 & -20 \\ 0 & 1 & 7 & -13 \\ 0 & 0 & 1 & -2 \end{array}\right)$$

A harmadik oszlopban felfelé eliminálva:

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -2 \end{array}\right)$$

Innen végül:

$$x_1 = 2, \qquad x_2 = 1, \qquad x_3 = -2$$

3. Oldjuk meg a következő lineáris egyenletrendszert Gauss-eliminációval:

$$\begin{array}{rcrcrcl} 2x_1 & - & x_2 & + & x_3 & = & 3 \\ 2x_1 & + & 2x_2 & - & 4x_3 & = & 4 \\ x_1 & - & 2x_2 & + & 3x_3 & = & 1 \end{array}$$

Megoldás: Mindenekelőtt: a rendszer mátrixa most szinguláris (mert determinánsa 0; ellenőrizzük!). Ez azt jelenti, hogy a megoldhatóság a jobb oldaltól függ: vagy van megoldás (és akkor mindjárt végtelen sok is), vagy nincs megoldás. A számítás sémája a következő:

$$\left(\begin{array}{ccc|c} 2 & -1 & 1 & 3 \\ 2 & 2 & -4 & 4 \\ 1 & -2 & 3 & 1 \end{array}\right) \rightarrow \left(\begin{array}{ccc|c} 1 & -2 & 3 & 1 \\ 2 & 2 & -4 & 4 \\ 2 & -1 & 1 & 3 \end{array}\right) \rightarrow$$

$$\left(\begin{array}{ccc|c} 1 & -2 & 3 & 1 \\ 0 & 6 & -10 & 2 \\ 0 & 3 & -5 & 1 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & -2 & 3 & 1 \\ 0 & 1 & -\frac{5}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Az utolsó eliminálásnál a 3. sor csupa 0 lett. Ez azt jelenti, hogy az egyik ismeretlen, pl. x_3 szabadon megválasztható: $x_3 := t$ (ahol $t \in \mathbf{R}$ tetszőleges). Így a 3. sor így alakul:

$$\left(\begin{array}{ccc|c} 1 & -2 & 3 & 1 \\ 0 & 1 & -\frac{5}{3} & \frac{1}{3} \\ 0 & 0 & 1 & t \end{array}\right)$$

Az eliminálást folytatva:

$$\left(\begin{array}{ccc|c} 1 & -2 & 0 & 1-3t \\ 0 & 1 & 0 & \frac{1}{3}+\frac{5}{3}t \\ 0 & 0 & 1 & t \end{array}\right) \quad \to \quad \left(\begin{array}{ccc|c} 1 & 0 & 0 & \frac{5}{3}+\frac{1}{3}t \\ 0 & 1 & 0 & \frac{1}{3}+\frac{5}{3}t \\ 0 & 0 & 1 & t \end{array}\right)$$

Az egyenletrendszernek tehát végtelen sok megoldása van, mégpedig tetszőleges $t \in \mathbf{R}$ mellett:

$$x_1 = \frac{5}{3} + \frac{1}{3}t, \quad x_2 = \frac{1}{3} + \frac{5}{3}t, \quad x_3 = t$$

4. Számítsuk ki Gauss-eliminációval az alábbi mátrix inverzét:

$$A := \begin{pmatrix} -2 & 3 & 1 \\ -1 & 1 & 1 \\ 2 & -2 & -1 \end{pmatrix}$$

Megoldás: A számítás sémája a következő:

$$\left(\begin{array}{ccc|ccc} -2 & 3 & 1 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 & 1 & 0 \\ 2 & -2 & -1 & 0 & 0 & 1 \end{array}\right)$$

Első lépésben cseréljük fel az 1. és a 2. sort, majd végezzük el az első oszlop eliminálását.

$$\left(\begin{array}{ccc|ccc} -1 & 1 & 1 & 0 & 1 & 0 \\ -2 & 3 & 1 & 1 & 0 & 0 \\ 2 & -2 & -1 & 0 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & -1 & -1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 1 & -2 & 0 \\ 0 & 0 & 1 & 0 & 2 & 1 \end{array}\right) \to$$

$$\left(\begin{array}{ccc|ccc} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 1 & 2 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 \end{array}\right)$$

Kaptuk tehát, hogy:

$$A^{-1} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 0 & 1 \\ 0 & 2 & 1 \end{pmatrix}$$

5. Számítsuk ki Gauss-eliminációval az alábbi mátrix inverzét:

$$A := \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 2 \\ -1 & 3 & 2 \end{pmatrix}$$

Megoldás: A számítás sémája a következő:

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 0 \\ -1 & 3 & 2 & 0 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 0 \\ 0 & 3 & 3 & 1 & 0 & 1 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 0 \\ 0 & 1 & 1 & \frac{1}{3} & 0 & \frac{1}{3} \end{array}\right)$$

Cseréljük meg a 2. és 3. egyenletet (a hozzájuk tartozó jobb oldalakkal együtt, hogy az eliminációt folytatni tudjuk:

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 2 & 0 & 1 & 0 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 & \frac{1}{2} & 0 \end{array}\right) \to$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 1 & 0 & \frac{1}{3} & -\frac{1}{2} & \frac{1}{3} \\ 0 & 0 & 1 & 0 & \frac{1}{2} & 0 \end{array}\right)$$

Az inverz mátrix tehát:

$$A^{-1} = \begin{pmatrix} 1 & -\frac{1}{2} & 0 \\ \frac{1}{3} & -\frac{1}{2} & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 \end{pmatrix}$$

6. Határozzuk meg az alábbi mátrix LU-felbontását, és ennek segítségével számítsuk ki a mátrix determinánsát:

$$A := \begin{pmatrix} 4 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 4 \end{pmatrix}$$

Megoldás: Mindenekelőtt L-et a már ismert elemeivel kitöltjük: a főátlóban 1-esek állnak, a főátló felett zérusok:

$$A = \begin{pmatrix} 4 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 4 \end{pmatrix} \qquad L = \begin{pmatrix} 1 & 0 & 0 \\ . & 1 & 0 \\ . & . & 1 \end{pmatrix}$$

Majd elkezdjük az eliminációt: A első sorának $\frac{2}{4}$-szeresét kivonjuk a második, $\frac{1}{4}$-szeresét a harmadik sorból, e szorzókat ($\frac{2}{4}$ és $\frac{1}{4}$) pedig beírjuk L első oszlopának második ill. harmadik helyére:

$$\begin{pmatrix} 4 & 2 & 1 \\ 0 & 3 & \frac{3}{2} \\ 0 & \frac{3}{2} & \frac{15}{4} \end{pmatrix} \qquad L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{4} & . & 1 \end{pmatrix}$$

Folytatjuk az eliminációt: a mátrix második sorának $\frac{1}{2}$-szeresét kivonjuk a harmadik sorból, az $\frac{1}{2}$ szorzót pedig beírjuk L második oszlopának harmadik helyére. Ezzel az eredeti mátrixból felső háromszög mátrixot kaptunk (ez lesz a felbontás U mátrixa), és egyidejűleg az L-et is megkaptuk.

$$U = \begin{pmatrix} 4 & 2 & 1 \\ 0 & 3 & \frac{3}{2} \\ 0 & 0 & 3 \end{pmatrix} \qquad L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{4} & \frac{1}{2} & 1 \end{pmatrix}$$

(Ellenőrizzük az eredményt az LU szorzat közvetlen kiszámításával! Próbáljuk ki a mintafeladatban mutatott mátrixszorzásos módszert is!)

A determináns az U mátrix diagonálelemeinek szorzata:

$$\det(A) = \det(U) = 4 \cdot 3 \cdot 3 = 36.$$

7. Határozzuk meg az alábbi mátrix LU-felbontását, és ennek segítségével számítsuk ki a mátrix determinánsát:

$$A := \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

Megoldás: Mindenekelőtt L-et a már ismert elemeivel kitöltjük: a főátlóban 1-esek állnak, a főátló felett zérusok:

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \qquad L = \begin{pmatrix} 1 & 0 & 0 \\ . & 1 & 0 \\ . & . & 1 \end{pmatrix}$$

Majd elkezdjük az eliminációt: A első sorának $\frac{1}{2}$-szeresét kivonjuk a második sorból (a harmadik sorból nem kell kivonni semmit, mert az első eleme 0), a szorzókat ($\frac{1}{2}$ és 0) pedig beírjuk L első oszlopának második ill. harmadik helyére:

$$\begin{pmatrix} 2 & 1 & 0 \\ 0 & \frac{3}{2} & 1 \\ 0 & 1 & 2 \end{pmatrix} \qquad L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & . & 1 \end{pmatrix}$$

Folytatjuk az eliminációt: a mátrix második sorának $\frac{2}{3}$-szorosát kivonjuk a harmadik sorból, az $\frac{2}{3}$ szorzót pedig beírjuk L második oszlopának harmadik helyére. Ezzel az eredeti mátrixból felső háromszög mátrixot kaptunk (ez lesz a felbontás U mátrixa), és egyidejűleg az L-et is megkaptuk.

$$U = \begin{pmatrix} 2 & 1 & 0 \\ 0 & \frac{3}{2} & 1 \\ 0 & 0 & \frac{4}{3} \end{pmatrix} \qquad L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & \frac{2}{3} & 1 \end{pmatrix}$$

(Ellenőrizzük az eredményt az LU szorzat közvetlen kiszámításával! Próbáljuk ki a mintafeladatban mutatott mátrixszorzásos módszert is!)

A determináns az U mátrix diagonálelemeinek szorzata:

$$\det(A) = \det(U) = 2 \cdot \frac{3}{2} \cdot \frac{4}{3} = 4.$$

2 A legkisebb négyzetek módszere

2.1 Motiváció

Gyakorta fellépő probléma, számoknál bonyolultabb objektumok pl. vektorok, függvények nagyságát valamint ezek távolságát alkalmas módon definiálni.

Tipikus példák:

- Ha egy függvényt egy másikkal (egyszerűbbel) közelítünk, akkor kettőjük távolsága mérheti a közelítés jóságát abban az értelemben, hogy minél kisebb ez a távolság, annál jobb a közelítés.

- Ha egy sokismeretlenes egyenletrendszert oldunk meg, óhatatlanul számítási hibákat követünk el, tekintve, hogy a számítógép is csak véges sok tizedesjeggyel dolgozik. Sőt, olykor a kiindulási adatok is hibával terheltek, mert pl. mérés eredményei. Ilyenkor jó tudni (vagy legalább megbecsülni), hogy a közelítő megoldás milyen távol van a pontos megoldástól: itt is, minél kisebb ez a távolság, annál jobb a közelítés.

Ebben a fejezetben a sík- és térgeometriából jól ismert nagyság- és távolságfogalmakat általánosítjuk $\mathbf{R}^N$-beli vektorokra, és valamilyen közös intervallumon értelmezett függvényekre. Ezen kívül igen jól használható alkalmazásokat mutatunk függvényközelítésekre és lineáris egyenletrendszerek közelítő megoldására.

2.2 Skaláris szorzat, norma és távolság $\mathbf{R}^N$-ben

Adott N dimenziószám mellett tekintsük az $\mathbf{R}^N$ vektorteret, melynek elemei a rendezett, valós $(x_1, x_2, ..., x_N)$ alakú szám-N-esek (vektorok). Két rendezett valós szám-N-est, $x := (x_1, ..., x_N)$-t és $y := (y_1, ..., y_N)$-t egyenlőnek tekintünk, ha a komponenseik rendre megegyeznek, azaz $x_1 = y_1, ..., x_N = y_N$. $N = 2$ vagy 3 esetén $\mathbf{R}^N$ azonosítható a geometriai síkkal ill. térrel egy rögzített derékszögű koordinátarendszer megadásával; ekkor a sík (tér) pontjai és a koordinátáik alkotta (x_1, x_2) ill. (x_1, x_2, x_3) rendezett valós számpárok (számhármasok) kölcsönösen egyértelműen megfeleltethetők egymásnak.

Az $x := (x_1, x_2)$, $y = (y_1, y_2)$ síkbeli pontok *távolsága* a Pitagorásztételből adódik: $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$. Térbeli pontokra hasonlóan: ha $x := (x_1, x_2, x_3)$, $y = (y_1, y_2, y_3)$, akkor a távolság:

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}.$$

Az x pont távolságát az origótól az x vektor *hosszának* is nevezzük.

Ezen egyszerű geometriai fogalmakat kézenfekvő módon általánosíthatjuk az $\mathbf{R}^N$ vektortér esetére is:

Definíció: Az $x := (x_1, x_2, ..., x_N) \in \mathbf{R}^N$ vektor hosszának vagy *normájának* az

$$||x|| := \sqrt{x_1^2 + x_2^2 + ... + x_N^2} = \sqrt{\sum_{j=1}^{N} x_j^2}$$

számot nevezzük. Az $x := (x_1, x_2, ..., x_N)$, $y := (y_1, y_2, ..., y_N) \in \mathbf{R}^N$ vektorok *távolsága* alatt pedig az

$$||x - y|| = \sqrt{\sum_{j=1}^{N} (x_j - y_j)^2}$$

számot értjük.

Sokszor célszerű bevezetni két vektor skaláris szorzatát is:

Definíció: Az $x := (x_1, x_2, ..., x_N)$, $y := (y_1, y_2, ..., y_N) \in \mathbf{R}^N$ vektorok *skaláris szorzatának* az alábbi számot nevezzük:

$$\langle x, y \rangle := \sum_{j=1}^{N} x_j y_j = x_1 y_1 + ... + x_N y_N$$

számot nevezzük.

Elemi számolásokkal adódnak a norma alapvető tulajdonságai:

Állítás:

- Tetszőleges $x \in \mathbf{R}^N$-re $||x|| \geq 0$ és $||x|| = 0$ pontosan akkor teljesül, ha x maga a zérusvektor, azaz $x = (0, 0, ..., 0)$.
- Tetszőleges $x \in \mathbf{R}^N$ vektor és $\alpha \in \mathbf{R}$ szám esetén: $||\alpha \cdot x|| = |\alpha| \cdot ||x||$.
- Tetszőleges $x, y \in \mathbf{R}^N$ vektor esetén: $||x+y|| \leq ||x|| + ||y||$ (háromszög-egyenlőtlenség).

Ezek – a háromszög-egyenlőtlenség kivételével – egyszerű, elemi meggondolásokból adódnak (ellenőrizzük!); a háromszög-egyenlőtlenséget kicsit később igazoljuk.

Ugyancsak elemi számolásokkal kaphatók meg a skaláris szorzatra vonatkozó alapvető egyenlőségek, melyek megkönnyítik a kétkezi számításokat:

Állítás: Tetszőleges $x, y, z \in \mathbf{R}^N$ vektorra és $\alpha \in \mathbf{R}$ számra:

- $||x|| = \sqrt{\langle x, x \rangle}$
- $\langle x, y \rangle = \langle y, x \rangle$
- $\langle \alpha x, y \rangle = \alpha \cdot \langle x, y \rangle$
- $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$

A fenti állításokból azonnal következik, hogy $\langle x, \alpha y \rangle = \alpha \cdot \langle x, y \rangle$ és $\langle x+y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ is igaz.

Különleges jelentősége van az alábbi tételnek:

Tétel (Cauchy-egyenlőtlenség): Tetszőleges $x, y \in \mathbf{R}^N$ esetén:

$$|\langle x, y \rangle| \leq ||x|| \cdot ||y||,$$

és egyenlőség csak abban az esetben áll fenn, ha x és y lineárisan összefüggők, azaz egyik a másik konstansszorosa.

Bizonyítás: Tetszőleges $\alpha \in \mathbf{R}$ esetén nyilván igaz, hogy $||x - \alpha y||^2 \geq 0$, azaz:

$$\sum_{j=1}^{N} (x_j - \alpha y_j)^2 = \sum_{j=1}^{N} \left(x_j^2 + 2\alpha x_j y_j + \alpha^2 j_j^2 \right) =$$

$$= ||x||^2 - 2\alpha \langle x, y \rangle + \alpha^2 ||y||^2 \geq 0$$

Legyen most már $\alpha := \frac{||x||}{||y||}$ (ha $y \neq \mathbf{0}$; $y = \mathbf{0}$ esetén az állítás a $0 = 0$ egyenlőségre egyszerűsödik). Ekkor:

$$||x||^2 + 2\frac{||x||}{||y||}\langle x, y \rangle + \frac{||x||^2}{||y||^2} \cdot ||y||^2 \geq 0,$$

ahonnan a lehetséges összevonások és egyszerűsítések után az

$$\langle x, y \rangle \leq ||x|| \cdot ||y||$$

egyenlőtlenséget kapjuk. Mivel ez minden $x, y \in \mathbf{R}^N$ esetén igaz, azért y helyébe $(-y)$-t írva is igaz marad:

$$\langle x, -y \rangle \leq ||x|| \cdot || - y|| = ||x|| \cdot ||y||,$$

ahonnan $\langle x, y\rangle \geq -||x|| \cdot ||y||$. Kaptuk, hogy:

$$-||x|| \cdot ||y|| \leq \langle x, y\rangle \leq ||x|| \cdot ||y||,$$

azaz valóban, $|\langle x, y\rangle| \leq ||x|| \cdot ||y||$. Egyenlőség pedig nyilván csak akkor van, ha $||x - \alpha y||^2 = 0$, azaz $x = \alpha y$.

A bizonyításból érdemes egy önmagában is érdekes és fontos egyenlőséget kiemelni:

$$||x - y||^2 = ||x||^2 - 2\langle x, y\rangle + ||y||^2,$$

és hasonlóan látható az is, hogy:

$$||x + y||^2 = ||x||^2 + 2\langle x, y\rangle + ||y||^2,$$

A Cauchy-egyenlőtlenségből a háromszög-egyenlőtlenség már egyszerűen adódik:

$$||x+y||^2 = ||x||^2+2\langle x, y\rangle+||y||^2 \leq ||x||^2+2||x|| \cdot ||y||+||y||^2 = (||x||+||y||)^2.$$

Mindkét oldal négyzetgyökét véve, épp a háromszög-egyenlőtlenséget kapjuk.

2.3 Skaláris szorzat, norma és távolság az $L_2(a, b)$ térben

Legyen $(a, b) \subset \mathbf{R}$ egy (véges vagy végtelen) intervallum. Jelölje $L_2(a, b)$ mindazon $f : (a, b) \to \mathbf{R}$ függvények összességét, melyekre az $\int_a^b |f(x)|^2\, dx$ integrál létezik és véges (a négyzetesen integrálható függvények tere).

Könnyen ellenőrizhető, hogy $L_2(a, b)$ vektortér a szokásos függvényműveletekre nézve.

Ilyen függvényekre szintén bevezethető a norma, a távolság és a skaláris szorzat; vegyük észre a nagyfokú analógiát az $\mathbf{R}^N$-ben bevezetett definíciókkal.

Definíció: Az $f \in L_2(a, b)$ függvény *normájának* az

$$||f|| := \sqrt{\int_a^b |f(x)|^2\, dx}$$

számot nevezzük. Az $f, g \in L_2(a, b)$ függvények *távolsága* alatt az

$$||f - g|| = \sqrt{\int_a^b |f(x) - g(x)|^2\, dx}$$

számot értjük. Az $f, g \in L_2(a,b)$ függvények skaláris szorzatának pedig az alábbi számot nevezzük:

$$\langle f, g \rangle := \int_a^b f(x)g(x)\,dx.$$

Az alábbi állítások elemi számolásokkal adódnak:

Állítás:

- Tetszőleges $f \in L_2(a,b)$-re $||f|| \geq 0$, és $||f|| = 0$ pontosan akkor teljesül, ha f azonosan 0 az (a,b) intervallumon.
- Tetszőleges $f \in L_2(a,b)$ függvény és $\alpha \in \mathbf{R}$ szám esetén: $||\alpha \cdot f|| = |\alpha| \cdot ||f||$.
- Tetszőleges $f, g \in L_2(a,b)$ vektor esetén: $||f + g|| \leq ||f|| + ||g||$ (háromszög-egyenlőtlenség).

Állítás: Tetszőleges $f, g, h \in L_2(a,b)$ függvényekre és $\alpha \in \mathbf{R}$ számra:

- $||f|| = \sqrt{\langle f, f \rangle}$
- $\langle f, g \rangle = \langle g, f \rangle$
- $\langle \alpha f, g \rangle = \alpha \cdot \langle f, g \rangle$
- $\langle f, g + h \rangle = \langle f, g \rangle + \langle f, h \rangle$

A fenti állításokból azonnal következik, hogy $\langle f, \alpha g \rangle = \alpha \cdot \langle f, g \rangle$ és $\langle f+g, h \rangle = \langle f, h \rangle + \langle g, h \rangle$ is igaz.

A Cauchy-egyenlőtlenség most is igaz, és lényegében ugyanúgy igazolható, mint $\mathbf{R}^N$-ben, így ettől eltekintünk. Ebben a formájában néha *Schwarz-egyenlőtlenségnek* vagy *Cauchy-Schwarz-egyenlőtlenségnek* is nevezik:

Tétel (Cauchy-Schwarz-egyenlőtlenség): Tetszőleges $f, g \in L_2(a,b)$ függvények esetén:

$$|\langle f, g \rangle| \leq ||f|| \cdot ||g||,$$

azaz

$$\left| \int_a^b f(x)g(x)\,dx \right| \leq \sqrt{\int_a^b |f(x)|^2)\,dx} \cdot \sqrt{\int_a^b |g(x)|^2)\,dx}$$

és egyenlőség csak abban az esetben áll fenn, ha f és g lineárisan összefüggők, azaz egyik a másik konstansszorosa.

A Cauchy-egyenlőtlenség segítségével a háromszög-egyenlőtlenség ugyanúgy igazolható, mint az $\mathbf{R}^N$ tér esetében. Érdemes megjegyezni továbbá az alábbi hasznos egyenlőségeket: tetszőleges $f, g \in L_2(a,b)$ függvények esetén:

$$||f+g||^2 = ||f||^2 + 2\langle f,g\rangle + ||g||^2,$$

és

$$||f-g||^2 = ||f||^2 - 2\langle f,g\rangle + ||g||^2.$$

2.4 A legkisebb négyzetek módszere

2.4.1 Lineáris regresszió

Legyenek adottak az x_1, x_2, ..., x_M és az y_1, y_2, ..., y_M számok (ahol $M \geq 2$). Ezek a gyakorlatban sokszor mérési adatok. Feltételezzük (egyéb, nem feltétlenül matematikai meggondolások alapján), hogy az x_j adatoktól az y_j adatok közel egy elsőfokú polinom szerint függenek, azaz az összefüggés formulája:

$$y = a_0 + a_1 x,$$

ahol a_0, a_1 egyelőre ismeretlen paraméterek, és épp ezekre vagyunk kíváncsiak.

Pontosabban: keressünk olyan $y = a_0 + a_1 x$ alakú polinomot, mely a "lehető legjobban" illeszkedik az x_1, x_2, ..., x_M, y_1, y_2, ..., y_M adatokra.

A tökéletes illeszkedés ez lenne:

$$a_0 + a_1 x_1 = y_1$$

$$a_0 + a_1 x_2 = y_2$$

$$...$$

$$a_0 + a_1 x_M = y_M$$

Tömör jelölésekkel:

$$A\underline{a} = \underline{y},$$

ahol $A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ ... & ... \\ 1 & x_M \end{pmatrix} \in \mathbf{M}_{M\times 2}$, $\underline{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \in \mathbf{R}^2$, $\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_M \end{pmatrix} \in \mathbf{R}^M$.

Az egyenletek száma (M) a gyakorlatban jóval meghaladhatja az ismeretlenek számát (2): ezt néha *túlhatározott egyenletrendszernek* is nevezik. Ennek általában nincs megoldása. Azonban felvethető, hogy milyen $\underline{a} \in \mathbf{R}^2$ mellett minimális az

$$F(\underline{a}) := ||A\underline{a} - \underline{y}||^2$$

eltérésnégyzet. Minél kisebb ez a – hibaként is felfogható – szám, annál jobbnak tekinthető az illeszkedés. Tökéletes illeszkedés esetén ez a hiba nyilván 0.

Ez a *legkisebb négyzetek módszerének* alapgondolata.

A minimumot realizáló $\underline{a} \in \mathbf{R}^2$ vektor tekinthető az $A\underline{a} = \underline{y}$ egyfajta "általánosított megoldásának", pontosabban: "legkisebb négyzetes megoldásának".

A fenti minimumfeladat pedig könnyen megoldható. Ismeretes, hogy az

$$F(a_0, a_1) := ||A\underline{a} - \underline{y}||^2 = \sum_{k=1}^{M} (a_0 + a_1 x_k - y_k)^2$$

kétváltozós függvénynek csak ott lehet minimuma, ahol mindkét változója szerinti parciális deriváltja zérus, azaz:

$$\frac{\partial F}{\partial a_0} = \sum_{k=1}^{M} 2 \cdot (a_0 + a_1 x_k - y_k) \cdot 1 = 0$$

$$\frac{\partial F}{\partial a_1} = \sum_{k=1}^{M} 2 \cdot (a_0 + a_1 x_k - y_k) \cdot x_k = 0$$

Az ismeretlen a_0, a_1 paraméterekre tehát az alábbi egyenletrendszert kaptuk:

$$a_0 \cdot \left(\sum_{k=1}^{M} 1 \right) + a_1 \cdot \left(\sum_{k=1}^{M} x_k \right) = \sum_{k=1}^{M} y_k$$

$$a_0 \cdot \left(\sum_{k=1}^{M} x_k \right) + a_1 \cdot \left(\sum_{k=1}^{M} x_k^2 \right) = \sum_{k=1}^{M} x_k y_k$$

Az egyenletrendszer determinánsa:

$$\left(\sum_{k=1}^{M} 1 \right) \cdot \left(\sum_{k=1}^{M} x_k^2 \right) - \left(\sum_{k=1}^{M} x_k \right)^2,$$

és ez a Cauchy-egyenlőtlenség miatt csak akkor lehet 0, ha az $(1, 1, ..., 1)$ és az $(x_1, x_2, ..., x_M)$ vektorok egymás konstansszorosai, azaz mindegyik x_k egyenlő. Ettől, a gyakorlat számára érdektelen esettől eltekintve, a determináns mindig pozitív, így az egyenletrendszer egyértelműen megoldható.

2.4.2 Kvadratikus és polinomiális regresszió

Legyenek adottak az x_1, x_2, ..., x_M és az y_1, y_2, ..., y_M számok (ahol $M \geq 3$). Finomítva az előző szakasz problémakörének vizsgálatát, tegyük fel, hogy az x_j adatoktól az y_j adatok közel egy másodfokú polinom szerint függenek, azaz a közelítő összefüggés formulája:

$$y = a_0 + a_1 x + a_2 x^2,$$

ahol a_0, a_1, a_2 egyelőre ismeretlen paraméterek.

Másképp megfogalmazva, keressünk olyan $y = a_0 + a_1 x + a_2 x^2$ alakú polinomot, mely a "lehető legjobban" illeszkedik az x_1, x_2, ..., x_M, y_1, y_2, ..., y_M adatokra.

A tökéletes illeszkedés ez lenne:

$$a_0 + a_1 x_1 + a_2 x_1^2 = y_1$$

$$a_0 + a_1 x_2 + a_2 x_2^2 = y_2$$

$$...$$

$$a_0 + a_1 x_M + a_2 x_M^2 = y_M$$

Röviden:

$$A\underline{a} = \underline{y},$$

ahol $A = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ ... & ... & ... \\ 1 & x_M & x_M^2 \end{pmatrix} \in \mathbf{M}_{M\times 3}$,

$$\underline{a} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \in \mathbf{R}^3, \quad \underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_M \end{pmatrix} \in \mathbf{R}^M.$$

Ha $M \geq 3$ (és ez a gyakorlatban legtöbbször így is van), akkor ez is egy túlhatározott egyenletrendszer, és általában nincs megoldása. A legkisebb négyzetes (vagy általánosított) megoldás az az $\underline{a} \in \mathbf{R}^3$ vektor, mely minimalizálja az

$$F(\underline{a}) := ||A\underline{a} - \underline{y}||^2$$

függvényt. Minimum csak ott lehet, ahol mindhárom változó szerinti parciális derivált zérus, azaz:

$$\frac{\partial F}{\partial a_0} = \sum_{k=1}^{M} 2 \cdot (a_0 + a_1 x_k + a_2 x_k^2 - y_k) \cdot 1 = 0$$

$$\frac{\partial F}{\partial a_1} = \sum_{k=1}^{M} 2 \cdot (a_0 + a_1 x_k + a_2 x_k^2 - y_k) \cdot x_k = 0$$

$$\frac{\partial F}{\partial a_2} = \sum_{k=1}^{M} 2 \cdot (a_0 + a_1 x_k + a_2 x_k^2 - y_k) \cdot x_k^2 = 0$$

Kaptuk, hogy az ismeretlen a_0, a_1, a_2 paraméterek kielégítik a következő egyenletrendszert:

$$a_0 \cdot \left(\sum_{k=1}^{M} 1 \right) + a_1 \cdot \left(\sum_{k=1}^{M} x_k \right) + a_2 \cdot \left(\sum_{k=1}^{M} x_k^2 \right) = \sum_{k=1}^{M} y_k$$

$$a_0 \cdot \left(\sum_{k=1}^{M} x_k \right) + a_1 \cdot \left(\sum_{k=1}^{M} x_k^2 \right) + a_2 \cdot \left(\sum_{k=1}^{M} x_k^3 \right) = \sum_{k=1}^{M} x_k y_k$$

$$a_0 \cdot \left(\sum_{k=1}^{M} x_k^2 \right) + a_1 \cdot \left(\sum_{k=1}^{M} x_k^3 \right) + a_2 \cdot \left(\sum_{k=1}^{M} x_k^4 \right) = \sum_{k=1}^{M} x_k^2 y_k$$

Nem kell ragaszkodni a legfeljebb első- vagy másodfokú közelítéshez. Pontosan ugyanezzel a technikával könnyen látható, hogy ha egy

$$y = a_0 + a_1 x + ... + a_p x^p$$

legfeljebb p-edfokú polinomot keresünk, mely a lehető legjobban illeszkedik az x_1, x_2, ..., x_M, y_1, y_2, ..., y_M adatokra (ahol most $M \geq p+1$), akkor az ismeretlen a_0, a_1, ..., a_p együtthatók minimalizálják az

$$F(\underline{a}) := ||A\underline{a} - \underline{y}||^2 = \sum_{k=1}^{M} \left(\sum_{j=0}^{p} a_j x_k^j - y_k \right)^2$$

függvényt, ahol

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 & ... & x_1^p \\ 1 & x_2 & x_2^2 & ... & x_2^p \\ ... & ... & ... & ... & ... \\ 1 & x_M & x_M^2 & ... & x_M^p \end{pmatrix} \in \mathbf{M}_{M \times (p+1)},$$

$$\underline{a} = \begin{pmatrix} a_0 \\ ... \\ a_p \end{pmatrix} \in \mathbf{R}^{p+1}, \quad \underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_M \end{pmatrix} \in \mathbf{R}^M.$$

A minimumhelyen az összes változó szerinti parciális derivált eltűnik, ami az alábbi egyenletrendszerre vezet:

$$\frac{\partial F}{\partial a_m} = \sum_{k=1}^{M} 2 \cdot \left(\sum_{j=0}^{p} a_j x_k^j - y_k \right) \cdot x_k^m = 0,$$

azaz

$$\sum_{j=0}^{p} a_j \cdot \left(\sum_{k=1}^{M} x_k^{j+m} \right) = \sum_{k=1}^{M} x_k^m y_k \qquad (m = 0, 1, ..., p)$$

2.4.3 Egyenletmegoldás a legkisebb négyzetek módszerével

Az előző regressziós problémát általánosítva, tekintsük az

$$Ax = b$$

egyenletrendszert, ahol $A \in \mathbf{M}_{M \times N}$, $b \in \mathbf{R}^M$, és az egyenletrendszer x megoldását $\mathbf{R}^N$-ben keressük. Tegyük fel, hogy $M \geq N$ (azaz akár több egyenletünk van mint ahány ismeretlenünk, tehát az egyenletrendszer túlhatározott). Ekkor általában nincs megoldás. Általánosított vagy legkisebb négyzetes megoldásnak nevezzük azt az $x \in \mathbf{R}^N$ vektort, mely minimalizálja a maradékvektor normanégyzetét, azaz az

$$F(x) := ||Ax - b||^2$$

függvényt. Ilyen általánosított megoldás már mindig létezik. Nyilvánvaló, hogy ha történetesen x pontos megoldás, tehát $Ax = b$, akkor $F(x) = 0$, és mivel F mindenütt nemnegatív, azért ez az x egyúttal minimumhely is. Ilyen értelemben tehát az F-et minimalizáló vektorok valóban általánosított megoldásnak tekinthetők.

A minimumhely megkeresése a szokásos differenciálszámítási eszközökkel történik. Nyilván

$$F(x_1, x_2, ..., x_N) = ||Ax - b||^2 = \sum_{k=1}^{M} \left(\sum_{j=1}^{N} A_{kj} x_j - b_k \right)^2,$$

ahonnan minden $m = 1, 2, ..., N$-re:

$$\frac{\partial F}{\partial x_m} = \sum_{k=1}^{M} 2 \cdot \left(\sum_{j=1}^{N} A_{kj} x_j - b_k \right) \cdot A_{km} = 0$$

A bal oldalon az összegezések sorrendjét felcserélve:

$$\frac{\partial F}{\partial x_m} = \sum_{j=1}^{N} \left(\sum_{k=1}^{M} A_{kj} A_{km} \right) \cdot x_j = \sum_{k=1}^{M} A_{km} b_k$$

Jelölje szokásos módon $A^* \in \mathbf{M}_{N \times M}$ az A mátrix adjungáltját, akkor definíció szerint a bal oldalon $A_{kj} = A^*_{jk}$, a jobb oldalon pedig $A_{km} = A^*_{mk}$. Azt kaptuk, hogy:

$$\frac{\partial F}{\partial x_m} = \sum_{j=1}^{N} \left(\sum_{k=1}^{M} A^*_{jk} A_{km} \right) \cdot x_j = \sum_{k=1}^{M} A^*_{mk} b_k$$

$m = 1, 2, ..., N$-re. Ez pedig a mátrixszorzás definíciója miatt az alábbi tömör vektoregyenlőséggel egyenértékű:

$$A^* A x = A^* b$$

Ezt az egyenletet az eredeti $Ax = b$ egyenlet *Gauss-féle normálegyenletének* nevezzük. Mivel nyilván $A^* A \in \mathbf{M}_{N \times N}$, azért a Gauss-féle normálegyenlet mátrixa $M \gg N$ esetén sokkal kisebb méretű, mint az eredeti A mátrix.

Ha $A^* A$ reguláris, akkor tehát a legkisebb négyzetes (általánosított) megoldás egyértelműen létezik. Sok esetben viszont a Gauss-féle normálegyenlet rosszul kondícionált, ami az egyenletmegoldás során numerikus nehézségeket okozhat.

Megjegyzések:

- A Gauss-féle normálegyenlethez egyszerűbb úton is eljuthatunk. Tekintsük az $F(x) := ||Ax - b||^2$ függvényt: ennek csak ott lehet minimuma, ahol a gradiensvektor zérusvektorral egyenlő. A gradiensvektor kiszámítása pedig standard módon történhet. Legyen $h \in \mathbf{R}^N$ tetszőleges, akkor:

  $$F(x+h) = ||Ax + Ah - b||^2 = ||Ax - b||^2 + 2\langle Ax - b, Ah \rangle + ||Ah||^2 =$$
  $$= F(x) + 2\langle A^*(Ax - b), h \rangle + \mathcal{O}(||h||^2),$$

  azaz $\operatorname{grad} F(x) = 2A^*(Ax - b)$. Ez pedig nyilván akkor zérusvektor, ha x megoldása a Gauss-féle normálegyenleteknek.

- A legkisebb négyzetek módszere akkor is használható, ha $M < N$ (alulhatározott egyenletrendszer). Ekkor az a jellemző, hogy a Gauss-féle normálegyenletnek több megoldása is van. Ezekből sokszor célszerű kiválasztani a legkisebb normájút. Ennek részleteivel itt nem foglalkozunk.

- Fontos látni, hogy a közelítés jóságát nem kötelező a maradékvektor normanégyzetével mérni. Legalább ennyire jogos a maradékvektor legnagyobb abszolút értékű komponensét (annak abszolút értékét) minimalizálni: ez az *egyenletesen legjobb közelítés*. Azonban ez a minimumfeladat általában technikailag sokkal nehezebb probléma, mint a legkisebb négyzetek módszere, így inkább az utóbbit használják – ha úgy tetszik, kényelmi okokból.

2.4.4 Függvényillesztés a legkisebb négyzetek módszerével

Legyen $[a.b] \subset \mathbf{R}$ egy véges intervallum, és $f \in L_2(a,b)$ adott függvény. A gyakorlatban sokszor f egy bonyolult és/vagy képlettel nem is adott függvény, melyet szeretnénk egyszerűbb függvényekkel – például polinomokkal – közelíteni.

Legyen tehát a p közelítő függvény egy legfeljebb N-edfokú polinom:

$$p(x) := a_0 + a_1 x + a_2 x^2 + ... + a_N x^N$$

A közelítés jóságát többféleképp is értelmezhetjük.

1) Felveszünk egy $a \leq x_1 < x_2 < ... < x_M \leq b$ alappontrendszert, és az a_0, a_1, ..., a_N együtthatókat úgy választjuk meg, hogy az *eltérések négyzetösszege*:

$$F(a_0, a_1, ..., a_N) := \sum_{k=1}^{M} (p(x_k) - f(x_k))^2 = \sum_{k=1}^{M} \left(\sum_{j=0}^{N} a_j x_k^j - f(x_k) \right)^2$$

minimális legyen. Ez egy polinomiális regressziós probléma az x_1, x_2, ...,x_M alappontrendszeren, a hozzárendelt $f(x_1)$, $f(x_2)$, ..., $f(x_M)$ függvényértékekre nézve. Az előzőekből ennek már tudjuk a megoldását. Az ismeretlen együtthatók az alábbi egyenletrendszer megoldásából számíthatók:

$$\sum_{j=0}^{N} a_j \cdot \left(\sum_{k=1}^{M} x_k^{j+m} \right) = \sum_{k=1}^{M} x_k^m f(x_k) \qquad (m = 0, 1, ..., N)$$

2) Az a_0, a_1, ..., a_N együtthatókat úgy választjuk meg, hogy az f és p függvényeknek az $L_2(a,b)$-norma szerinti *távolságnégyzete*, azaz a

$$G(a_0, a_1, ..., a_N) := \int_a^b (p(x) - f(x))^2 \, dx = \int_a^b \left(\sum_{j=0}^{N} a_j x^j - f(x) \right)^2 dx$$

szám minimális legyen.

A már ismert technikával: minimum csak ott lehet, ahol mindegyik változó szerinti parciális derivált zérus, azaz mindegyik $m = 0, 1, ..., N$-re:

$$\frac{\partial G}{\partial a_m} = \int_a^b 2 \cdot \left(\sum_{j=0}^{N} a_j x^j - f(x) \right) \cdot x^m \, dx = 0$$

Innen az ismeretlen együtthatókra az alábbi egyenletrendszert nyerjük:

$$\sum_{j=0}^{N} a_j \left(\int_a^b x^{j+m} \, dx \right) = \int_a^b f(x) \cdot x^m \, dx \qquad (m = 0, 1, ..., N)$$

Az egyenletrendszer mátrixának elemei igen egyszerűen számíthatók:

$$A_{mj} = \int_a^b x^{j+m} \, dx = \frac{b^{j+m+1} - a^{j+m+1}}{j + m + 1}$$

Ha speciálisan $[a, b] = [0, 1]$, akkor a mátrix különösen egyszerű alakú lesz:

$$A_{mj} = \frac{1}{j + m + 1} \qquad (j, m = 0, 1, ..., N)$$

Ezt a mátrixot $(N + 1)$-edrendű *Hilbert-mátrixnak* nevezzük.

A jobb oldali $\int_a^b f(x) \cdot x^m \, dx$ integrálok *elvben* f ismeretében kiszámíthatók: a gyakorlatban ezek kiszámítása sokszor csak közelítő módszerekkel történhet. Ilyen közelítő integrálási módszerekkel egy másik fejezetben fogunk foglalkozni.

Megjegyzés: A Hilbert-mátrixok szimmetrikusak és pozitív definitek, de rendkívül rosszul kondícionáltak. Emiatt előszeretettel használják egyenletmegoldó algoritmusok tesztelésére.

A kétféle függvényközelítés hasonlóságát és különbözőségét az alábbi példán keresztül szemléltetjük.

Példa: Tekintsük az

$$f(x) := \cos \frac{\pi x}{2} \qquad (x \in [-1, \, 1])$$

formulával értelmezett függvényt. Közelítsük f-et egy legfeljebb másodfokú, $p(x) := a_0 + a_1 x + a_2 x^2$ alakú polinommal.

1) *Kvadratikus regresszió*: Ehhez egy (legalább 3 pontból álló) alappontrendszerre van szükség. Legyen pl. $x_1 := -1$, $x_2 := 0$, $x_3 := 1$. Akkor a függvényértékek az alappontokban: $f_1 = 0$, $f_2 = 1$, $f_3 = 0$.

Az adatok alapján azonnal ellenőrizhető, hogy a $p(x) := 1-x^2$ formulával definiált másodfokú polinom épp illeszkedik a fenti adatokra. Természetesen a formális kvadratikus regresszió is ugyanezt az eredményt adja. Most $M = 3$, a regressziós egyenletrendszer pedig a következő (ellenőrizzük!):

$$\begin{pmatrix} 3 & 0 & 2 \\ 0 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

melynek egyetlen megoldása: $\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$, azaz a regressziós polinom valóban $p(x) = a_0 + a_1 x + a_2 x^2 = 1 - x^2$.

Ha 3 helyett pl. 5 alappontot alkalmazunk, a regressziós polinom egy kicsit változik. Legyen most $M := 5$ és $x_1 := -1$, $x_2 := -0.5$, $x_3 := 0$, $x_4 := 0.5$, $x_5 := 1$. Ekkor f alapponti értékei: $f_1 = 0$, $f_2 = \frac{\sqrt{2}}{2}$, $f_3 = 1$, $f_4 = \frac{\sqrt{2}}{2}$ és $f_5 = 0$. A regressziós egyenlet alakja most (ellenőrizzük!):

$$\begin{pmatrix} 5 & 0 & 2.5 \\ 0 & 2.5 & 0 \\ 2.5 & 0 & 2.125 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1+\sqrt{2} \\ 0 \\ \frac{\sqrt{2}}{4} \end{pmatrix} = \begin{pmatrix} 2.41421 \\ 0 \\ 0.35355 \end{pmatrix},$$

melynek egyetlen megoldása $\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0.97059 \\ 0 \\ -0.97549 \end{pmatrix}$, azaz a regressziós polinom most $p(x) = a_0 + a_1 x + a_2 x^2 = 0.97059 - 0.97549 x^2$.

Melyik közelítés a jobb? A válasz nem magától értetődő, az alkalmazott alappontrendszertől függ. A (-1), 0, 1 alappontrendszert tekintve az első, 3-pontos közelítés a jobb, mert ez esetben az alapponti függvényértékeket a regressziós polinom *pontosan* reprodukálja, így az eltérés-négyzetösszeg pontosan 0, míg a második, 5-pontos közelítés esetén ez $(a_0+a_2-0)^2+(a_0-1)^2+(a_0+a_2-0)^2 = 0.000913$. A (-1), (-0.5), 0, 0.5, 1 5-pontos alappont rendszer mellett kiszámítva az eltérés-négyzetösszegeket az adódik, hogy az első közelítés esetén az eltérés-négyzetösszeg 0.0036797, míg a második

közelítés esetén ez a szám 0.0016821. Tehát ezen az alappontrendszeren a második közelítés a jobb.

2) *Az L_2-eltérés minimalizálása*: Itt az

$$\int_{-1}^{1} (a_0 + a_1 x + a_2 x^2 - f(x))^2 \, dx$$

négyzetintegrált minimalizáljuk. A korábbi szakasz eredményei alapján az a_0, a_1, a_2 együtthatók az alábbi egyenletrendszert elégítik ki:

$$\begin{pmatrix} \int_{-1}^{1} 1\,dx & \int_{-1}^{1} x\,dx & \int_{-1}^{1} x^2\,dx \\ \int_{-1}^{1} x\,dx & \int_{-1}^{1} x^2\,dx & \int_{-1}^{1} x^3\,dx \\ \int_{-1}^{1} x^2\,dx & \int_{-1}^{1} x^3\,dx & \int_{-1}^{1} x^4\,dx \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \int_{-1}^{1} f(x)\,dx \\ \int_{-1}^{1} f(x) \cdot x\,dx \\ \int_{-1}^{1} f(x) \cdot x^2\,dx \end{pmatrix}$$
$$= \begin{pmatrix} 2.41421 \\ 0 \\ 0.35355 \end{pmatrix},$$

Az integrálásokat elvégezve, az egyenletrendszer konkrét alakja:

$$\begin{pmatrix} 2 & 0 & \frac{2}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{2}{3} & 0 & \frac{2}{5} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1.27324 \\ 0 \\ 0.24119 \end{pmatrix},$$

melynek egyetlen megoldása: $\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0.98016 \\ 0 \\ -1.03063 \end{pmatrix}$, ami kissé eltér a kvadratikus regresszió eredményétől. Most

$$p(x) = a_0 + a_1 x + a_2 x^2 = 0.98016 - 1.03063 x^2.$$

Ugyanakkor viszont a módszer nem használ intervallumfelbontást, így a közelítés – a kvadratikus regressziótól eltérően – alappont-független.

2.5 Feladatok

1. Mutassuk meg, hogy tetszőleges $x = (x_1, x_2, ..., x_N) \in \mathbf{R}^N$ esetén érvényes a következő egyenlőtlenség:

$$\left| \sum_{j=1}^{N} x_j \right| \leq \sqrt{N} \cdot ||x||$$

Megoldás: Jelölje $e := (1, 1, ..., 1) \in \mathbf{R}^N$ a csupa 1-es komponensekből álló vektort. Akkor, a *Cauchy-egyenlőtlenséget* felhasználva:

$$\left| \sum_{j=1}^{N} x_j \right| = \left| \sum_{j=1}^{N} 1 \cdot x_j \right| = |\langle e, x \rangle| \leq$$

$$\leq ||e|| \cdot ||x|| = \sqrt{\sum_{j=1}^{N} 1^2} \cdot ||x|| = \sqrt{N} \cdot ||x||$$

2. Mutassuk meg, hogy ha az $x = (x_1, x_2, ..., x_N) \in \mathbf{R}^N$, $y = (y_1, y_2, ..., y_N) \in \mathbf{R}^N$ vektorokra $||x|| = ||y||$ teljesül, akkor:

$$\langle x + y, x - y \rangle = 0.$$

Megoldás: Felhasználva a skaláris szorzatra vonatkozó azonosságokat:

$$\langle x + y, x - y \rangle = \langle x, x \rangle - \langle x, y \rangle + \langle y, x \rangle - \langle y, y \rangle = ||x||^2 - ||y||^2 = 0.$$

3. Legyenek adottak a síkbeli $(x_1, y_1), (x_2, y_2), ...(x_N, y_N), \in \mathbf{R}^2$ pontok. Határozzuk meg azt az (x, y) pontot a síkon, melyre az előző pontoktól mért távolságaik négyzetösszege minimális.

Megoldás: Tetszőleges $(x, y) \in \mathbf{R}^2$ pontra a szóban forgó távolságnégyzetösszeg:

$$f(x, y) = \sum_{j=1}^{N} ((x - x_j)^2 + (y - y_j)^2)$$

Ennek minimuma csak ott lehet, ahol mindkét változó szerinti parciális derivált zérus, azaz:

$$\frac{\partial f}{\partial x} = \sum_{j=1}^{N} 2 \cdot (x - x_j) = 0$$

$$\frac{\partial f}{\partial y} = \sum_{j=1}^{N} 2 \cdot (y - y_j) = 0$$

Innen az (x, y) minimumhely csak a következő lehet:

$$x = \frac{1}{N} \sum_{j=1}^{N} x_j, \qquad y = \frac{1}{N} \sum_{j=1}^{N} y_j$$

azaz a pontrendszer súlypontja. (Ellenőrizzük, hogy itt tényleg minimuma van f-nek!)

4. Legyenek adottak az $x_1 := -2$, $x_2 := -1$, $x_3 := 1$, $x_4 := 2$ alappontok és a hozzájuk rendelt $f_1 := 5$, $f_2 := 7$, $f_3 := 11$, $f_4 := 13$ értékek. Határozzuk meg a fenti adatokra illeszkedő lineáris regressziós függvényt.

   Megoldás: Az adatokkal: $\sum_{j=1}^{4} 1 = 4$, $\sum_{j=1}^{4} x_j = 0$, $\sum_{j=1}^{4} x_j^2 = 10$, $\sum_{j=1}^{4} f_j = 36$, $\sum_{j=1}^{4} f_j x_j = 20$. A $p(x) := a_0 + a_1 x$ regressziós függvény a_0, a_1 együtthatói tehát kielégítik az alábbi egyenletrendszert:

$$\begin{pmatrix} 4 & 0 \\ 0 & 10 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 36 \\ 20 \end{pmatrix},$$

   ahonnan $a_0 = 9$, $a_1 = 2$, tehát a regressziós polinom:

$$p(x) = 9 + 2x.$$

   Vegyük észre, hogy a regressziós polinom *pontosan* illeszkedik az adatokra!

5. Tekintsük az előző feladat $x_1 := -2$, $x_2 := -1$, $x_3 := 1$, $x_4 := 2$ alappontjait és a hozzájuk rendelt $f_1 := 5$, $f_2 := 7$, $f_3 := 11$, $f_4 := 13$ értékeket. Határozzuk meg a fenti adatokra illeszkedő kvadratikus regressziós függvényt.

Megoldás: Az adatokkal: $\sum_{j=1}^4 1 = 4$, $\sum_{j=1}^4 x_j = 0$, $\sum_{j=1}^4 x_j^2 = 10$, $\sum_{j=1}^4 x_j^3 = 0$, $\sum_{j=1}^4 x_j^4 = 34$, $\sum_{j=1}^4 f_j = 36$, $\sum_{j=1}^4 f_j x_j = 20$, $\sum_{j=1}^4 f_j x_j^2 = 90$. A $p(x) := a_0 + a_1 x + a_2 x^2$ regressziós függvény a_0, a_1, a_2 együtthatói tehát kielégítik az alábbi egyenletrendszert:

$$\begin{pmatrix} 4 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 36 \\ 20 \\ 90 \end{pmatrix},$$

ahonnan $a_0 = 9$, $a_1 = 2$, $a_2 = 0$. Így a regressziós polinom:

$$p(x) = 9 + 2x,$$

pontosan ugyanaz, mint az előző feladatban. Ez nem is meglepő, mert az előző feladatban már az elsőfokú regressziós polinom is pontosan illeszkedett az adatokra, úgyhogy a regressziós polinom fokszámának növelése ezen nyilván nem javít már.

6. Tekintsük ismét az előző két feladat alappontjait és a hozzájuk rendelt értékeket, de f_4 értékét kissé változtassuk meg: $f_4 := 14$; tehát legyenek $x_1 := -2$, $x_2 := -1$, $x_3 := 1$, $x_4 := 2$, $f_1 := 5$, $f_2 := 7$, $f_3 := 11$, $f_4 := 14$. Határozzuk meg a fenti adatokra illeszkedő lineáris és kvadratikus regressziós függvényt.

Megoldás: Az adatokkal: $\sum_{j=1}^4 1 = 4$, $\sum_{j=1}^4 x_j = 0$, $\sum_{j=1}^4 x_j^2 = 10$, $\sum_{j=1}^4 x_j^3 = 0$, $\sum_{j=1}^4 x_j^4 = 34$, $\sum_{j=1}^4 f_j = 37$, $\sum_{j=1}^4 f_j x_j = 22$, $\sum_{j=1}^4 f_j x_j^2 = 94$. A $p(x) := a_0 + a_1 x$ regressziós függvény a_0, a_1 együtthatói kielégítik az alábbi egyenletrendszert:

$$\begin{pmatrix} 4 & 0 \\ 0 & 10 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 37 \\ 22 \end{pmatrix},$$

ahonnan $a_0 = 9.25$, $a_1 = 2.2$. Így a regressziós polinom:

$$p(x) = 9.25 + 2.2x,$$

A kvadratikus regresszió esetében a $p(x) := a_0 + a_1 x + a_2 x^2$ regressziós függvény a_0, a_1, a_2 együtthatói pedig a következő egyenletrendszert elégítik ki:

$$\begin{pmatrix} 4 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 37 \\ 22 \\ 94 \end{pmatrix},$$

ahonnan $a_0 = 8.8333$, $a_1 = 2.2000$, $a_2 = 0.1666$. Így a regressziós polinom:

$$p(x) = 8.8333 + 2.2x + 0.1666x^2.$$

7. Legyenek x_1, x_2, ..., x_M különböző alappontok, tartozzanak hozzájuk az y_1, y_2, ..., y_M értékek ($M \geq 2$). Mutassuk meg, hogy a lineáris regresszió $a := \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$ együtthatóit meghatározó egyenletrendszer épp az $Aa = b$ túlhatározott egyenletrendszer Gauss-féle normálegyenlete, ahol

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ ... & ... \\ 1 & x_M \end{pmatrix}, \qquad b = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_M \end{pmatrix}.$$

Megoldás: Nyilván: $A^* = \begin{pmatrix} 1 & 1 & ... & 1 \\ x_1 & x_2 & ... & x_M \end{pmatrix}$, ahonnan:

$$A^*A = \begin{pmatrix} \sum_{k=1}^M 1 & \sum_{k=1}^M x_k \\ \sum_{k=1}^M x_k & \sum_{k=1}^M x_k^2 \end{pmatrix}, \qquad A^*b = \begin{pmatrix} \sum_{k=1}^M y_k \\ \sum_{k=1}^M x_k y_k \end{pmatrix}.$$

8. Legyenek x_1, x_2, ..., x_M különböző alappontok, tartozzanak hozzájuk az y_1, y_2, ..., y_M értékek ($M \geq 3$). Mutassuk meg, hogy a kvadratikus regresszió $a := \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}$ együtthatóit meghatározó egyenletrendszer épp az $Aa = b$ túlhatározott egyenletrendszer Gauss-féle normálegyenlete, ahol

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ ... & ... & ... \\ 1 & x_M & x_M^2 \end{pmatrix}, \qquad b = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_M \end{pmatrix}.$$

Megoldás: Nyilván: $A^* = \begin{pmatrix} 1 & 1 & ... & 1 \\ x_1 & x_2 & ... & x_M \\ x_1^2 & x_2^2 & ... & x_M^2 \end{pmatrix}$, ahonnan:

$$A^*A = \begin{pmatrix} \sum_{k=1}^M 1 & \sum_{k=1}^M x_k & \sum_{k=1}^M x_k^2 \\ \sum_{k=1}^M x_k & \sum_{k=1}^M x_k^2 & \sum_{k=1}^M x_k^3 \\ \sum_{k=1}^M x_k^2 & \sum_{k=1}^M x_k^3 & \sum_{k=1}^M x_k^4 \end{pmatrix}, \qquad A^*b = \begin{pmatrix} \sum_{k=1}^M y_k \\ \sum_{k=1}^M x_k y_k \\ \sum_{k=1}^M x_k^2 y_k \end{pmatrix}.$$

9. Keressük meg az $Ax = b$ egyenletrendszer legkisebb négyzetes megoldását, ahol $A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \\ 4 & 5 \end{pmatrix}$, $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $b = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 2 \end{pmatrix}$.

   Megoldás: A Gauss-féle normálegyenlet: $A^*Ax = A^*b$, azaz

$$\begin{pmatrix} 30 & 40 \\ 40 & 54 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 16 \\ 22 \end{pmatrix}$$

   Ennek egyetlen megoldása: $x_1 = -0.8$, $x_2 = 1$.

10. Tekintsük az $f(x) := e^x$ exponenciális függvényt a $[0, 1]$ intervallumon. Határozzuk meg azt a legfeljebb harmadfokú p polinomot, mely f-et az $L_2(0, 1)$ norma szerint a legjobban közelíti.

    Megoldás: A $p(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ polinom együttható az alábbi egyenletrendszert elégítik ki:

$$a_0 \int_0^1 1\,dx + a_1 \int_0^1 x\,dx + a_2 \int_0^1 x^2\,dx + a_3 \int_0^1 x^3\,dx = \int_0^1 e^x\,dx$$

$$a_0 \int_0^1 x\,dx + a_1 \int_0^1 x^2\,dx + a_2 \int_0^1 x^3\,dx + a_3 \int_0^1 x^4\,dx = \int_0^1 xe^x\,dx$$

$$a_0 \int_0^1 x^2\,dx + a_1 \int_0^1 x^3\,dx + a_2 \int_0^1 x^4\,dx + a_3 \int_0^1 x^5\,dx = \int_0^1 x^2e^x\,dx$$

$$a_0 \int_0^1 x^3\,dx + a_1 \int_0^1 x^4\,dx + a_2 \int_0^1 x^5\,dx + a_3 \int_0^1 x^6\,dx = \int_0^1 x^3e^x\,dx$$

   Az integrálásokat a bal és a jobb oldalakon elvégezve (ellenőrizzük!):

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} e-1 \\ 1 \\ e-2 \\ 6-2e \end{pmatrix}$$

   Ennek egyetlen megoldása: $a_0 = 0.99906$, $a_1 = 1.01830$, $a_2 = 0.42125$, $a_3 = 0.27863$. A legjobban közelítő polinom tehát:

$$p(x) = 0.99906 + 1.01830x + 0.42125x^2 + 0.27863x^3$$

Érdekességképp megjegyezzük, hogy ez *nem egyezik* a harmadfokú Taylor-polinommal ($1 + x + 0.5x^2 + 0.16666x^3$), bár az együtthatók közel vannak a Taylor-polinom együtthatóihoz.

3 Interpoláció

3.1 Motiváció

A gyakorlatban megoldandó problémák közt az egyik leggyakoribb az interpoláció – még ha ezt nem is mindig mondjuk ki explicite. Durván szólva, adottak bizonyos *helyek* (egyenesen vagy síkon vagy térben), ezek az *interpolációs alappontok*. Minden interpolációs alapponthoz tartozik egy hozzárendelt érték (ez lehet szám, de akár vektor is). Az interpoláció alapproblémája: keressünk olyan, előírt tulajdonságokkal rendelkező függvényt, melyek az adott helyeken az adott értékeket veszi fel.

Néhány jellegzetes példa:

- Az interpolációs alappontok földrajzi koordinátapárok, a hozzárendelt értékek pedig az illető (a Föld felszínén levő) pont tengerszint feletti magassága. Feladat: a fenti adatokból domborzati térkép generálása, azaz olyan – kellően sima – kétváltozós függvény előállítása, mely illeszkedik az adatokra.

- Az interpolációs alappontok egy vízfelület – folyó, tó vagy tenger – adott földrajzi koordinátájú pontjai, a hozzárendelt értékek pedig az ott érvényes vízmélységek, pontosabban, az ezekből előállított fenékszintek. Feladat: a fenti adatokból megbízható vízmélység-adatokat előállítani ott is, ahol mélységmérés nem történt. Egy ilyen térképnek a hajózhatóság szempontjából különös fontossága van.

- Számítógépes modellek, szoftverek általában rengeteg adatot igényelnek, amelyek valamiféle anyagi állandót jelentenek más és más pontokban. Ilyen lehet például egy áramlási modellben a vízmélység, egy hőterjedési modellben a hővezetési tényező stb. Ha valahol ez nem áll rendelkezésre, akkor – jobb híján – azt a környező helyeken érvényes adatokból kell közelíteni: ez is egy speciális interpoláció.

- Maradva az áramlási modelleknél, tegyük fel, hogy egy tóban a szél hatására kialakuló cirkulációkat szeretnénk modellezni. Ehhez persze szükségünk van a szélsebességekre, amik vektorok. Ugyanakkor konkrét mérést csak néhány meteorológiai állomáson végeznek: másutt az itt mért adatokból kell interpolálni. Ez egy példa a vektor-interpolációra is. Jellemzően ekkor azt is meg lehet (vagy kell) követelni, hogy az interpolált vektormező tegyen eleget valamilyen differenciálegyenletnek is (pl. legyen divergenciamentes).

- Interpolációs problémára vezet az is, ha adott pontokat összekötő sima görbét (vagy adott pontokra illeszkedő sima felületet) szeretnénk előállítani (görbe- ill. felületillesztés).

- stb. stb. stb...

A fenti durva megfogalmazásban az interpoláció erősen alulhatározott probléma: nagyon sok olyan függvény konstruálható, mely előírt helyeken előírt értékeket vesz fel. Szükséges tehát pontosabban specifikálni vagy azt a függvényhalmazt, ahonnan az interpolációs függvényt választjuk, vagy azt a tulajdonságot, amit megkövetelünk az interpolációs függvénytől.

A továbbiakban csak a legegyszerűbb, klasszikus interpolációs módszerekkel foglalkozunk, ahol az interpolációs alappontok egy egyenes mentén helyezkednek el. Ez eleve egyfajta természetes rendezést jelent az alappontok közt, melyet a bemutatott interpolációs módszerek ki is használnak. Jóval nehezebb probléma a *szórt alappontú interpoláció*, ahol az alappontok síkban vagy térben helyezkednek el, mindenféle struktúra és rendezés nélkül. A fejezet végén – érintőlegesen – ilyen módszerekről is lesz szó.

3.2 A Lagrange-interpoláció

Legyenek az interpolációs alappontok mind különbözők: $a \leq x_0 < x_1 < x_2 < ... < x_N \leq b$, a hozzárendelt értékek pedig rendre $f_0, f_1, ..., f_N \in \mathbf{R}$ tetszőleges számok.

Probléma: Keressünk olyan $L_N(x) := a_0 + a_1 x + ... + a_N x^N$ legfeljebb N-edfokú polinomot, mely teljesíti az *interpolációs feltételeket*, azaz $L_N(x_k) = f_k$ ($k = 0, 1, ..., N$-re).

Miután $(N + 1)$ db interpolációs feltétel van, és összesen szintén $(N + 1)$ egyelőre ismeretlen együttható, ezért várható, hogy az interpolációs problémának létezik, mégpedig feltehetően egyértelmű megoldása. Ez valóban így is van. Az egyértelműség megmutatásával kezdjük: ha L_N és P_N mindketten legfeljebb N-edfokú polinomok, melyek kielégítik az interpolációs feltételeket, akkor kettőjük különbsége, $L_N - P_N$ is legfeljebb N-edfokú polinom, mely minden alappontban – tehát $(N + 1)$ különböző helyen – zérussal egyenlő. $L_N - P_N$ gyökeinek száma tehát meghaladja a fokszámot. Az algebra alaptétele miatt ez csak úgy lehetséges, ha $L_N - P_N \equiv 0$, azaz $P_N \equiv L_N$. Az interpolációs polinom tehát valóban egyértelmű. Másféle interpolációs technikáktól való megkülönböztetésképp, ezt nevezzük *Lagrange-interpolációs polinomnak*.

Ami az interpolációs polinom konkrét előállítását illeti, több módszert is mutatunk. A legegyszerűbb az interpolációs feltételeket megkövetelni, ami az ismeretlen együtthatókra egy $(N+1)$-ismeretlenes lineáris egyenletrendszert jelent:

$$L_N(x_k) = a_0 + a_1 x_k + ... + a_N x_k^N = \sum_{j=0}^{N} x_k^j \cdot a_j = f_k \qquad (k = 0, 1, ..., N)$$

Az egyenletrendszer mátrixa:

$$A = \begin{pmatrix} 1 & x_0 & x_0^2 & ... & x_0^N \\ 1 & x_1 & x_1^2 & ... & x_1^N \\ 1 & x_2 & x_2^2 & ... & x_2^N \\ ... & ... & ... & ... & ... \\ 1 & x_N & x_N^2 & ... & x_N^N \end{pmatrix} \in \mathbf{M}_{(N+1)\times(N+1)}$$

Megmutatható, hogy ha az alappontok mind különbözők, akkor az A mátrix reguláris, így az interpolációs egyenletrendszernek pontosan egy megoldása van.

A Lagrange-interpolációs polinom egyenletmegoldás nélkül, explicit formulákkal is előállítható. Adott alappontrendszer és N fokszám mellett jelölje $\ell_j^{(N)}$ a *Lagrange-féle alappolinomokat*:

$$\ell_j^{(N)}(x) := \frac{(x-x_0)(x-x_1)...(x-x_{j-1})(x-x_{j+1})...(x-x_N)}{(x_j-x_0)(x_j-x_1)...(x_j-x_{j-1})(x_j-x_{j+1})...(x_j-x_N)}$$

$(j = 0, 1, ..., N)$. Mindegyik Lagrange-féle alappolinom *pontosan* N-edfokú, és egyszerű behelyettesítéssel adódik, hogy

$$\ell_j^{(N)}(x_k) = \begin{cases} 1 & (j = k) \\ \\ 0 & (j \neq k) \end{cases}$$

Innen pedig azonnal következik, hogy a következő polinom legfeljebb N-edfokú, eleget tesz az interpolációs feltételeknek, tehát megegyezik a Lagrange-féle interpolációs polinommal:

$$L_N(x) := \sum_{j=0}^{N} f_j \cdot \ell_j^{(N)}(x)$$

Ha ugyanis x valamilyen alapponttal egyezik, pl. $x = x_k$, akkor a fenti összeg minden tagja eltűnik, egy kivételével (a k-adik tag kivételével), és $L_N(x_k) = f_k \cdot \ell_k^{(N)}(x_k) = f_k$.

Most tehát a Lagrange-interpolációs polinom előállításához nem kellett egyenletrendszert megoldani, viszont szorzat alakú polinomokat kell kezelni, ami magasabb fokszám esetén eléggé kényelmetlen lehet.

Kézi számolásokra jó kompromisszum lehet az *osztott differenciák* használata.

Adott $x_0, x_1, ..., x_N$ alappontrendszer és $f : [a, b] \to \mathbf{R}$ függvény mellett definiáljuk az *elsőrendű osztott differenciákat* a következő módon:

$$f[x_j, x_{j+1}] := \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j}$$

($j = 0, 1, ..., N-1$). Ezek tehát közönséges különbségi hányadosok. Definiáljuk rekurzív módon a *magasabb rendű osztott differenciákat*:

$$f[x_j, x_{j+1}, ..., x_{j+k}] := \frac{f[x_{j+1}, ..., x_{j+k}] - f[x_j, ..., x_{j+k-1}]}{x_{j+k} - x_j}$$

Ezekután tetszőleges $x_0, x_1, ..., x_N$ alappontrendszer és az erre vonatkozó $f(x_0), f(x_1), ..., f(x_N)$ függvényértékek ismeretéből most már gépiesen számíthatók az osztott differenciák. Példaképp, ha $N = 3$, akkor a következő táblázat elemei közvetlen számolásokkal adódnak:

x_0	$\boxed{f(x_0)}$			
		$\boxed{f[x_0, x_1]}$		
x_1	$f(x_1)$		$\boxed{f[x_0, x_1, x_2]}$	
		$f[x_1, x_2]$		$\boxed{f[x_0, x_1, x_2, x_3]}$
x_2	$f(x_2)$		$f[x_1, x_2, x_3]$	
		$f[x_2, x_3]$		
x_3	$f(x_3)$			

Mivel a különbségi hányadosok (elsőrendű osztott differenciák) az elsőrendű deriváltakkal kapcsolatosak, várható, hogy a magasabb rendű osztott differenciák a magasabb rendű deriváltakkal mutatnak hasonlatosságot. Ez valóban így is van. A részletes számításokat mellőzve, az osztott differenciák táblázatából a Lagrange-interpolációs polinom az alábbi formulával áll elő:

$$L_N(x) = f(x_0) + f[x_0, x_1] \cdot (x - x_0) + f[x_0, x_1, x_2] \cdot (x - x_0)(x - x_1) + ...$$
$$+ f[x_0, x_1, x_2, ..., x_N] \cdot (x - x_0)(x - x_1)...(x - x_{N-1})$$

Ehhez az osztott differenciák táblázatának csak a bekeretezett elemeit (felső sor) kell használni.

A módszer előnye mindkét előző módszerhez képest az, hogy ha az alappontrendszert egy újabb alapponttal bővítjük, nem kell elölről kezdenünk a számítást: elég csak a legmagasabb rendű osztott differenciát kiszámítani, és a Lagrange-interpolációs polinom formulája mindössze egy új taggal bővül.

A fenti technikákat a következő példán keresztül szemléltetjük.

Legyenek az alappontok: $x_0 := -1$, $x_1 := 0$, $x_2 := 1$, a hozzárendelt értékek pedig: $f_0 := 0$, $f_1 := 1$, $f_2 := 0$. Akkor a Lagrange-interpolációs polinom legfeljebb másodfokú.

Határozzuk meg a Lagrange-interpolációs polinomot mindhárom, fentebb bemutatott módszer szerint.

1) Írjuk fel az interpolációs polinomot $L_2(x) := a_0 + a_1 x + a_2 x^2$ alakban. Az interpolációs feltételeket megkövetelve, az ismeretlen a_0, a_1, a_2 együtthatókra az alábbi egyenletrendszert kapjuk:

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix},$$

melynek egyetlen megoldása: $\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$. Tehát a Lagrange-interpolációs polinom: $L_2(x) = a_0 + a_1 x + a_2 x^2 = 1 - x^2$.

2) A Lagrange-alappolinomok:

$$\ell_0^{(2)}(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{x(x-1)}{(-1)\cdot(-2)} = \frac{1}{2}x^2 - \frac{1}{2}x$$

$$\ell_1^{(2)}(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x+1)(x-1)}{1\cdot(-1)} = -x^2 + 1$$

$$\ell_2^{(2)}(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x+1)x}{2\cdot 1} = \frac{1}{2}x^2 + \frac{1}{2}x$$

Ezért:

$$L_2(x) = f_0 \cdot \ell_0^{(2)}(x) + f_1 \cdot \ell_1^{(2)}(x) + f_2 \cdot \ell_2^{(2)}(x) = -x^2 + 1$$

3) Osztott differenciákkal:

$$\begin{array}{cccc} j & x_j & f_j & \\ 0 & -1 & \boxed{0} & \\ & & \frac{1-0}{0-(-1)} = \boxed{1} & \\ 1 & 0 & 1 & \frac{-1-1}{1-(-1)} = \boxed{-1} \\ & & \frac{0-1}{1-0} = -1 & \\ 2 & 1 & 0 & \end{array}$$

Ezért:

$$L_2(x) = 0 + 1 \cdot (x+1) + (-1) \cdot (x+1)(x-0) = 1 - x^2$$

Alapvető fontosságú kérdés, hogy a Lagrange-interpolációs mennyire pontos abban az értelemben, hogy ha előre adott, sima függvényeket közelítünk a Lagrange-interpolációs polinomjukkal egy adott alappontrendszeren az itt felvett értékeikből, akkor a Lagrange-interpolációs polinom mennyire tér el az adott függvénytől A következő tétel erre ad választ.

Tétel: Ha $f \in C^{N+1}[a,b]$ (azaz $(N+1)$-szer folytonosan differenciálható az $[a,b]$ intervallumon), és L_N az $x_0, x_1, ..., x_N \in [a,b]$ alappontrendszerre és az $f(x_0), f(x_1), ..., f(x_N)$ alapponti adatokra támaszkodó Lagrange-interpolációs polinom, akkor tetszőleges $x \in [a,b]$ esetén:

$$|f(x) - L_N(x)| \leq \frac{\max\limits_{[a,b]} |f^{(N+1)}|}{(N+1)!} \cdot |\omega_N(x)|$$

ahol $\omega_N(x) := (x - x_0)(x - x_1)...(x - x_N)$. Innen pedig nyilván:

$$|f(x) - L_N(x)| \leq \max_{[a,b]} |f^{(N+1)}| \cdot \frac{(b-a)^{N+1}}{(N+1)!}$$

Bizonyítás: Legyen $x \in [a,b]$ tetszőleges, rögzített szám. Ha x interpolációs alappont, akkor a tétel a nyilvánvaló $0 \leq 0$ egyenlőtlenségre egyszerűsödik. Tegyük fel tehát, hogy x nem egyezik egyik alapponttal sem. Tekintsük az alábbi formulával definiált függvényt:

$$g(t) := f(t) - L_N(t) - \frac{\omega_N(t)}{\omega_N(x)} \cdot (f(x) - L_N(x))$$

Akkor g eltűnik az összes alappontban *és még a* $t = x$ *helyen is* (ellenőrizzük!), tehát $(N+2)$ különböző helyen. A differenciálszámítás középértéktételeiből

ismert Rolle-féle tétel szerint ezért a g' deriváltfüggvény eltűnik (legalább) $(N+1)$ különböző helyen, a g'' második deriváltfüggvény eltűnik (legalább) N különböző helyen, és így tovább, a $g^{(N+1)}$ deriváltfüggvény eltűnik legalább egy $\xi \in [a,b]$ helyen. Ámde $L_N^{(N+1)} \equiv 0$ (a deriválás rendje nagyobb a fokszámnál), és

$$\omega_N^{(N+1)}(t) = \frac{d^{N+1}}{dt^{N+1}}(t^{N+1} + ...) \equiv (N+1)!$$

miatt:

$$g^{(N+1)}(\xi) = f^{(N+1)}(\xi) - \frac{(N+1)!}{\omega_N(x)} \cdot (f(x) - L_N(x)) = 0$$

Az egyenlőséget átrendezve kapjuk, hogy minden $x \in [a,b]$-hez van oly $\xi \in [a,b]$, hogy

$$f(x) - L_N(x) = \frac{f^{(N+1)}(\xi)}{(N+1)!} \cdot \omega_N(x).$$

Mindkét oldal abszolút értékét véve, a tétel állítása innen már adódik.

A tétel tetszetős hibabecslés, ám a gyakorlatban óvatosan használandó. Ha $N \to \infty$, akkor az ismert $\frac{(b-a)^N}{N!} \to 0$ határérték azt sugallhatja, hogy a Lagrange-interpolációs polinomok L_N sorozatának hibája (azaz a $\max\limits_{x\in[a,b]} |f(x) - L_N(x)|$ szám) gyorsan 0-hoz tart, ha az interpolációs alappontok száma minden határon túl nő. Ez akkor van így, ha az f függvény deriváltjainak abszolút értéke a deriválás rendjével nem túl gyorsan növekszik, ami nem mindig áll fenn. Másrészt a gyakorlati esetek nagy részében f-et nem ismerjük, így a deriváltjait sem; csak az $f_0, f_1, ..., f_N$ értékek adottak, és nincs információnk arról, hogy ezek miféle függvény helyettesítési értékei. Egészen egyszerű esetekben is előfordulhat, hogy bár az interpolációs polinom pontosan kielégíti az interpolációs feltételeket az alappontokban, de két alappont között egészen szélsőséges értékeket is felvehet.

Tekintsük a következő példát (*Runge-példa*): legyen $x_0 := -5$, $x_1 := -4$, $x_2 := -3$, ..., $x_{10} := 5$ és $f_k := \frac{1}{1+x_k^2}$ $(k = 0, 1, ..., 10)$. Az alábbi ábrán ezen adatokra (folytonos vonallal) illesztedő 10-edfokú Lagrange-interpolációs polinom grafikonját látjuk (szaggatott vonal). Megfigyelhető, hogy a szélső alappontok közt az interpolációs polinom és az eredeti, $f(x) := \frac{1}{1+x^2}$ formulával értelmezett függvényt nagyon nagy hibával közelíti. Az alappontrendszert finomítva ($x_0 := -5$, $x_1 := -4.5$, $x_2 := -4$, ..., $x_{20} := 5$) a hiba nemhogy csökkenne, de sokkal nagyobb lett. Tanulságképp leszűrhető, hogy a Lagrange-interpolációt a gyakorlatban csak nem túlságosan magas fokszám mellett érdemes használni általában.

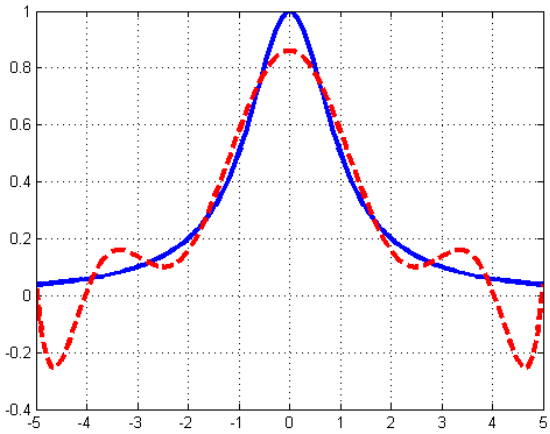

Figure 1: Az $x \to \frac{1}{1+x^2}$ függvény grafikonja (folytonos vonal), és a 10-edfokú Lagrange-interpolációs függvény grafikonja (szaggatott vonal)

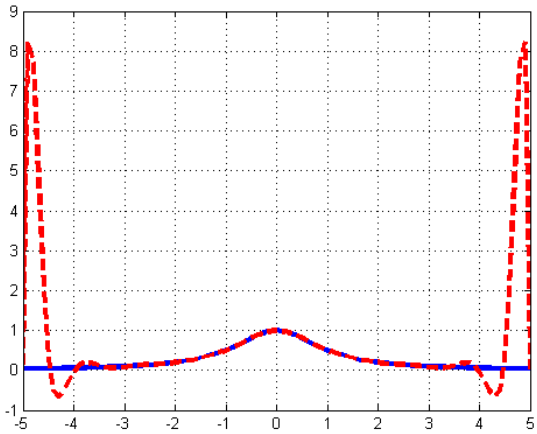

Figure 2: Az $x \to \frac{1}{1+x^2}$ függvény grafikonja (folytonos vonal), és a 20-adfokú Lagrange-interpolációs függvény grafikonja (szaggatott vonal)

3.3 Az Hermite-interpoláció

Ez a technika a Lagrange-interpolációt annyiban általánosítja, hogy az interpolációs polinomnak nemcsak az értékei, de a deriváltjai is előírhatók az alappontokban.

Legyenek $x_0, x_1, ..., x_N \in [a, b]$ adott interpolációs alappontok, m_0, m_1, ... , $m_N \geq 1$ egészek, és legyenek adva az

$$\begin{array}{c} f_0^{(0)}, f_0^{(1)}, ...f_0^{(m_0-1)} \\ f_1^{(0)}, f_1^{(1)}, ...f_1^{(m_1-1)} \\ ..... \\ f_N^{(0)}, f_N^{(1)}, ...f_N^{(m_N-1)} \end{array}$$

számok. Jelölje $m := m_0 + m_1 + ... + m_N$. Keressünk olyan, legfeljebb $(m-1)$-edfokú P_{m-1} polinomot, melyre teljesül, hogy

$$P_{m-1}^{(k)}(x_j) = f_j^{(k)} \qquad (k = 0, 1, ..., m_j - 1, \qquad j = 0, 1, ..., N)$$

Szemléletesen, az m_j számok azt mutatják, hogy az x_j alapponthoz hány adat (érték és derivált) van rendelve.

A probléma ebben a formában meglehetősen általános. Bizonyítás nélkül megemlítjük, hogy az Hermite-interpolációs probléma egyértelműen megoldható:

Tétel: Ilyen P_{m-1} polinom egyetlenegy létezik.

Az interpolációs polinom együtthatói az interpolációs feltételek által kikényszerített, az ismeretlen a_0, a_1, ..., a_{m-1} együtthatókra felírt m-ismeretlenes egyenletrendszer megoldásával határozható meg.

A problémával ilyen általánosságban nem foglalkozunk, de megemlítjük az alábbi speciális eseteket:

- Mindegyik $m_j = 1$, ekkor $m = N + 1$. Mindegyik x_j-hez csak az f_j érték tartozik (Lagrange-interpoláció).
- $N = 0$, $m = m_0$. x_0-hoz az $f_0^{(k)}$ $(k = 0, 1, ..., m-1)$ értékek tartoznak (Taylor-polinom):

$$P_{m-1}(x) = \sum_{k=0}^{m-1} \frac{f_0^{(k)}}{k!}(x - x_0)^k$$

- Mindegyik $m_j = 2$. Mindegyik x_j-hez az f_j érték és az f'_j deriváltérték tartozik (Hermite-Fejér-interpoláció).

A legutóbbinak is egy speciális esete lesz számunkra a fontos, ami megalapozza a gyakorlatban igen jól működő spline interpolációt (ld. később).

3.4 Kétpontos Hermite-interpoláció

Legyenek adottak az x_0, x_1 alappontok, és a hozzájuk tartozó f_0, f_1, f'_0, f'_1 értékek. Keressük azt a legfeljebb harmadfokú H polinomot, melyre $H(x_j) = f_j$ és $H'(x_j) = f'_j$ $(j = 0, 1)$. Az előző szakaszból már tudjuk, hogy ilyen polinom létezik és egyértelmű.

Jelölje $h := x_1 - x_0$, és keressük H-t ilyen alakban:

$$H(x) = A + B \cdot \frac{x - x_0}{h} + C \cdot \frac{(x - x_0)^2}{h^2} + D \cdot \frac{(x - x_0)^3}{h^3}$$

Akkor nyilván:

$$H'(x) = B \cdot \frac{1}{h} + 2C \cdot \frac{(x - x_0)}{h^2} + 3D \cdot \frac{(x - x_0)^2}{h^3}$$

Az interpolációs egyenletek:

$$\begin{aligned}
H(x_0) &= A &&= f_0 \\
H(x_1) &= A + B + C + D &&= f_1 \\
h \cdot H'(x_0) &= B &&= h \cdot f'_0 \\
h \cdot H'(x_1) &= B + 2C + 3D &&= h \cdot f'_0
\end{aligned}$$

Közvetlen számítással ellenőrizhető, hogy az egyenletrendszer megoldása a következő:

$$\begin{aligned}
A &= f_0 \\
B &= h f'_0 \\
C &= -3f_0 + 3f_1 - 2hf'_0 - hf'_1 \\
D &= 2f_0 - 2f_1 + hf'_0 + hf'_1
\end{aligned}$$

ami egy könnyen realizálható formulát ad a kétpontos Hermite-interpoláció alkalmazására.

3.5 Harmadfokú spline interpoláció

Legyenek adva az $a = x_0 < x_1 < ... < n_N = b$, interpolációs alappontok és a hozzájuk rendelt $f_0, f_1, ..., f_N$ értékek.

Alapötlet: Definiáljunk $f'_0, f'_1, ..., f'_N$ deriváltértékeket. Az f_k, f'_k, f_{k+1}, f'_{k+1} adatokkal minden $[x_k, x_{k+1}]$ részintervallumon egymástól függetlenül hajtsunk végre egy-egy kétpontos Hermite-interpolációt. Így kapjuk a H_0, H_1, ..., H_{N-1} harmadfokú polinomokat.

Ekkor az egyes részintervallumokon értelmezett H_k függvények, valamint azok deriváltjai is automatikusan *folytonosan csatlakoznak* az $x_1, ..., x_{N-1}$ pontokban, bárhogy is választjuk meg az f'_k deriváltértékeket.

A fő probléma az, hogy hogyan *célszerű* definiálni az f'_k deriváltértékeket. Néhány lehetséges út:

- f'_k-kat tetszőlegesen választjuk, pl. $f'_k := 0$. (Ez nagyon rossz definíció.)
- f'_k-kal az x_k-beli deriváltakat közelítjük az adatokból, pl. különbségi hányadosokkal: $f'_k := \frac{f_{k+1}-f_{k-1}}{x_{k+1}-x_{k-1}}$. (Ez jobb, mint az előző.)

De van még jobb megoldás: f'_k-kat úgy definiáljuk, hogy még a H_k függvények *második deriváltjai is folytonosan csatlakozzanak* az alappontokban, azaz

$$H''_{k-1}(x_k) = H''_k(x_k)$$

teljesüljön minden $k = 1, 2, ..., N-1$-re. Így kapjuk a harmadfokú spline interpolációt.

Jelölje $h_k := x_{k+1} - x_k$ $(k = 0, 1, ..., N-1)$. A kétpontos Hermite-interpoláció együtthatóira nyert formulákból hosszabb, de elemi számolás után kapjuk, hogy:

$$H''_{k-1}(x_k) = \frac{2}{h_{k-1}^2} \cdot \left(3f_{k-1} - 3f_k + h_{k-1}f'_{k-1} + 2h_{k-1}f'_k\right)$$

$$H''_k(x_k) = \frac{2}{h_k^2} \cdot \left(-3f_k + 3f_{k+1} - 2h_k f'_k - h_k f'_{k+1}\right)$$

A $H''_{k-1}(x_k) = H''_k(x_k)$ feltételek tehát biztosan teljesülnek, ha

$$\frac{2}{h_{k-1}^2} \cdot \left(3f_{k-1} - 3f_k + h_{k-1}f'_{k-1} + 2h_{k-1}f'_k\right) =$$

$$= \frac{2}{h_k^2} \cdot \left(-3f_k + 3f_{k+1} - 2h_k f'_k - h_k f'_{k+1}\right)$$

azaz, ha az $f'_0, ..., f'_N$ értékek kielégítik az alábbi egyenletrendszert:

$$\frac{1}{h_{k-1}} f'_{k-1} + \left(\frac{2}{h_{k-1}} + \frac{2}{h_k}\right) f'_k + \frac{1}{h_k} f'_{k+1} =$$

$$= -\frac{3}{h_{k-1}^2} f_{k-1} + \left(\frac{3}{h_{k-1}^2} - \frac{3}{h_k^2}\right) f_k + \frac{3}{h_k^2} f_{k+1}$$

($k = 1, 2, ..., N-1$). Ha az f'_0, f'_N deriváltértékek adottak, akkor az egyenletrendszer csak $N-1$ ismeretlent tartalmaz; ha nem, akkor ezek meghatározása további feltételek ("peremfeltételek") érvényesítésével történik.

Fontos speciális eset, ha az alappontok ekvidisztánsak, azaz $h_0 = h_1 = ... = h_{N-1} = h$. Ekkor az előző egyenletrendszer jelentősen leegyszerűsödik: könnyen ellenőrizhető, hogy ekkor

$$f'_{k-1} + 4f'_k + f'_{k+1} = \frac{-3f_{k-1} + 3f_{k+1}}{h} \qquad (k = 1, 2, ..., N-1)$$

Megjegyzés: A szélső, x_0 és x_N alappontokban tett feltételek (*peremfeltételek*) megválasztásának néhány szokásos módja:

- Szabadon előírjuk f'_0 és f'_N értékeit, pl. $f'_0 := f'_N := 0$.
- Az f'_0, f'_N deriváltértékeket a pontos deriváltértékekre állítjuk be (így nyerjük a *teljes spline függvényt*):

  $$f'_0 := f'(x_0), \qquad f'_N := f'(x_N)$$

  Ehhez persze szükséges $f'(x_0)$ és $f'(x_N)$ ismerete.
- Az f'_0, f'_N deriváltértékeket különbségi hányadosokkal közelítjük, pl.

  $$f'_0 := \frac{f_1 - f_0}{h_0}, \qquad f'_N := \frac{f_N - f_{N-1}}{h_{N-1}}$$

- Megköveteljük, hogy a spline függvény *második deriváltja* 0 legyen a végpontokban: $H''_0(x_0) := H''_{N-1}(x_N) := 0$. Ezt *természetes peremfeltételnek* nevezzük, és az alábbi egyenlőségek megkövetelésére vezet (ellenőrizzük!):

  $$2f'_0 + f'_1 = \frac{3f_1 - 3f_0}{h_0}, \qquad f'_{N-1} + 2f'_N = \frac{3f_N - 3f_{N-1}}{h_{N-1}}$$

Bizonyítás nélkül megemlítjük a spline interpoláció pontosságáról szóló tételt:

Tétel: Ha $f : [a, b] \to \mathbf{R}$ négyszer folytonosan differenciálható $[a, b]$-ben, és S_f jelöli az f függvény alapponti értékeiből meghatározott spline függvényt, az $f'_0 := f'(x_0)$, $f'_N := f'(x_N)$ peremfeltételek megkövetelése mellett, akkor f és S_f eltérésére teljesül, hogy

$$\max_{a \le x \le b} |f(x) - S_f(x)| = \mathcal{O}(h^4)$$

ahol h jelöli az alappontrendszerben előforduló legnagyobb lépésközt. Sőt, a spline deriváltjai is jól közelítik f deriváltjait:

$$\max_{a \le x \le b} |f'(x) - S'_f(x)| = \mathcal{O}(h^3)$$

$$\max_{a \le x \le b} |f''(x) - S''_f(x)| = \mathcal{O}(h^2)$$

$$\max_{a \le x \le b} |f'''(x) - S'''_f(x)| = \mathcal{O}(h)$$

Az ordókban szereplő konstansok f-től függenek, de az alappontrendszertől nem.

Megjegyezzük még, hogy a fenti eredmény az $f''_0 := f''_N := 0$ természetes peremfeltétel megkövetelése esetén is érvényben marad, amennyiben $f : [a, b] \to \mathbf{R}$ négyszer folytonosan differenciálható $[a, b]$-ben, és $f''(a) = f''(b) = 0$.

Megjegyzés: A spline eredetileg egy rajzeszköz volt, mellyel adott pontokon átmenő sima görbéket lehetett rajzolni (ld. a következő szakaszt). Különösen fontos volt ez a hajógyártásban. A spline maga egy rugalmas, fából vagy fémből készült hosszú szalagszerű eszköz volt, melyet hozzá lehetett igazítani néhány adott ponthoz; két pont között a saját rugalmassága által meghatározott alakot vett fel. Részletesebben ld:
`https://en.wikipedia.org/wiki/Flat_spline`

3.6 Görbeillesztés

Legyenek adva az (x_1, y_1), (x_2, y_2), ... ,$(x_N, y_N) \in \mathbf{R}^2$ adott, síkbeli pontok (sorrendjük lényeges). Keressünk olyan $\mathbf{R} \to \mathbf{R}^2$, $t \to (x(t), y(t))$ paraméterezésű síkgörbét, mely illeszkedik az adott pontokra.

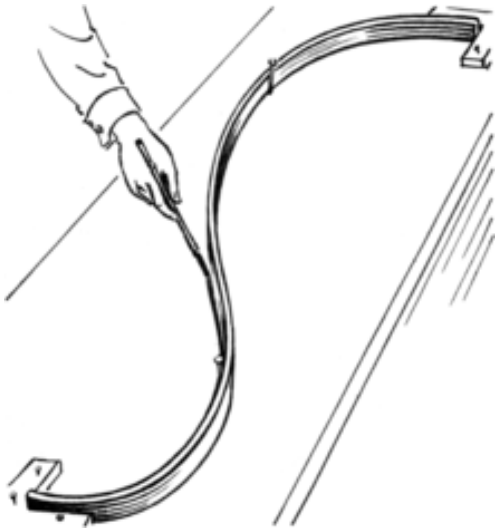

Figure 3: A spline mint rajzeszköz

A probléma két független, egyváltozós interpolációs problémára bontható. Vegyünk fel a paramétertartományban N db különböző paramétert: $t_1 < t_2 < ... < t_N$. (Egyéb információ híján lehet pl. a $t_k := k$ választással is élni.) Ezekután tekintsük a $t_1, t_2, ..., t_N$ és $x_1, x_2, ..., x_N$ adatok meghatározta interpolációs problémát, azaz keressünk olyan $x : \mathbf{R} \to \mathbf{R}$ függvényt, mely az adott paraméterértékekben az adott abszcisszaértékeket veszi fel, azaz $x(t_k) = x_k$ $(k = 1, 2, ., N)$. Hasonlóan, tekintsük a $t_1, t_2, ..., t_N$ és $y_1, y_2, ..., y_N$ adatok meghatározta interpolációs problémát, azaz keressünk olyan $y : \mathbf{R} \to \mathbf{R}$ függvényt, mely az adott paraméterértékekben az adott ordinátaértékeket veszi fel, azaz $y(t_k) = y_k$ $(k = 1, 2, ., N)$. A két interpolációs függvény által meghatározott $(x, y) : \mathbf{R} \to \mathbf{R}^2$ függvény épp a kívánt, az adott pontokon átmenő síkgörbe paraméterezése lesz. Az alkalmazott interpolációs technika elvileg tetszőleges lehet. Ha a pontok száma nem túl kicsi, akkor a Lagrange-interpoláció alkalmazása ellenjavallt, mert különösen a szélső pontok környékén a görbe lefutása nagyon extrém is lehet. Sokkal "szebb" a koordinátánkénti spline interpolációval előállított görbe.

Az ábrán 11, szabálytalanul elhelyezett pontra illesztett görbék láthatók. Az első esetben (10-edfokú) Lagrange-interpolációt alkalmaztunk mindkét koordinátára. Jól látható a görbe extrém lefutása. A második esetben spline interpolációt alkalmaztunk: a görbe az előzőnél sokkal realisztikusabb lett.

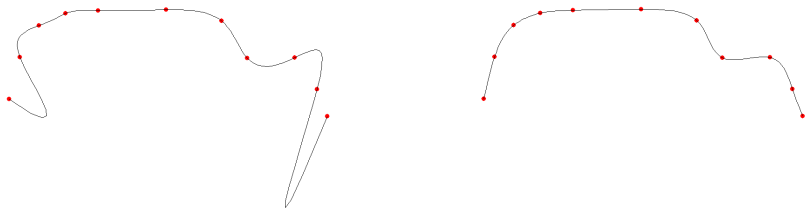

Figure 4: Görbeillesztés 11 pontra. Lagrange-interpoláció (balra) ill. spline interpoláció (jobbra)

Megjegyzés: Az itt vázolt görbeillesztési technika értelemszerű változtatásokkal *térgörbék* esetére is alkalmazható.

3.7 Szórt alappontú interpoláció

Legyenek $x_1, x_2, ..., x_N \in \mathbf{R}^2$ adott helyek *a síkon* (interpolációs alappontok), és $f_1, f_2, ..., f_N \in \mathbf{R}$ adott, az alappontokhoz rendelt értékek. Az alappontok elhelyezkedésében nem tételezünk fel semmilyen struktúrát (pl. rácsstruktúrát), sorrendjük is közömbös.

Az egyik legrégebbi és legegyszerűbb interpolációs technika a *Shepard-módszer* (más néven az inverz távolsággal súlyozás módszere). Itt az interpolációs függvényt a következő súlyozott összeg definiálja:

$$f(x) := \frac{\sum_{j=1}^{N} f_j \cdot \frac{1}{||x - x_j||^p}}{\sum_{j=1}^{N} \frac{1}{||x - x_j||^p}},$$

ha x nem interpolációs alappont. Itt $p > 1$ adott paraméter: szokásos értéke 2 vagy 4.

Az alappontokban f nincs értelmezve, de van határértéke: könnyen látható, hogy:

Állítás: Ha $x \to x_k$ valamely $k = 1, ..., N$-re, akkor $f(x) \to f_k$.

Bizonyítás: Feltehető, hogy $k = 1$. Legyen $x \to x_1$, akkor

$$f(x) = \frac{f_1 \cdot \frac{1}{||x-x_1||^p} + f_2 \cdot \frac{1}{||x-x_2||^p} + ... + f_N \cdot \frac{1}{||x-x_N||^p}}{\frac{1}{||x-x_1||^p} + \frac{1}{||x-x_2||^p} + ... + \frac{1}{||x-x_N||^p}} =$$

$$= \frac{f_1 + f_2 \cdot \frac{||x-x_1||^p}{||x-x_2||^p} + ... + f_N \cdot \frac{||x-x_1||^p}{||x-x_N||^p}}{1 + \frac{||x-x_1||^p}{||x-x_2||^p} + ... + \frac{||x-x_1||^p}{||x-x_N||^p}} \to \frac{f_1}{1} = f_1$$

Kicsit hosszabb számolásokkal az is megmutatható, hogy f deriválható az alappontokon kívül, és:

Állítás: Ha $x \to x_k$ valamely $k = 1, ..., N$-re, akkor f mindkét változó szerinti parciális deriváltja 0-hoz tart.

Ezt az állítást nem bizonyítjuk. Szemléletes jelentése: az interpolációs felület sima, de mindegyik alappontban vízszintes az érintősíkja, ami különösen szintvonalas ábrázoláskor lehet zavaró (az alappontok körül közel koncentrikus körök a szintvonalak).

A módszer egyszerű, könnyen programozható és numerikusan stabil: a Lagrange-interpolációtól eltérően, ha $||x|| \to +\infty$, akkor az $f(x)$ interpolált értékek korlátosak maradnak, és az is könnyen látható, hogy ekkor $f(x)$ az adatok $\frac{1}{N} \cdot \sum_{j=1}^{N} f_j$ számtani közepéhez tart (miért?). Viszont adott függvénynek az alappontokban felvett értékeiből való reprodukálása esetén meglehetősen pontatlan.

A módszert illusztrálandó, tekintsük az

$$f(x, y) := \sin \pi x \cdot \sin \pi y$$

formulával definiált függvényt az $\Omega := \{(x, y) \in \mathbf{R}^2 : \ 0 \leq x, y \leq 1\}$ egységnégyzeten értelmezve. Vegyünk fel 30 alappontot az egységnégyzetben véletlenszerűen, és az itt felvett függvényértékekből számítsuk ki a Shepard-interpolációs függvényt, a $p := 2$ paraméterválasztással. Ennek grafikonja látható az alábbi ábrán. Noha az alappontbeli értékek pontosan egyeznek az f függvény itt felvett értékeivel, az eredeti és az interpolált függvény meglehetősen eltér egymástól. A következő ábrán látható esetben szintén 30 véletlenszerűen megválasztott alappontot használtunk, de itt a p paraméter értékét 4-nek definiáltuk.

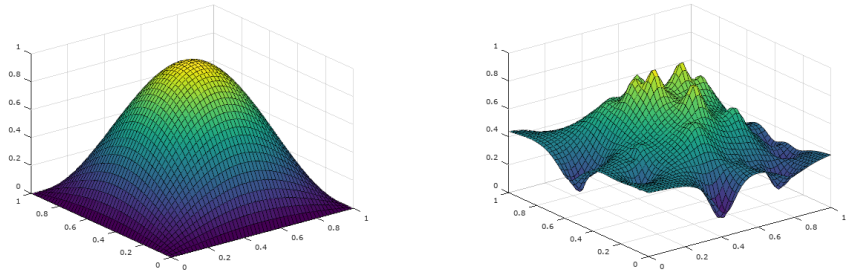

Figure 5: Szórt pontú interpoláció 30 véletlenszerűen választott alappontra. Tesztfelület (balra) ill. Shepard-interpoláció $p = 2$ mellett (jobbra)

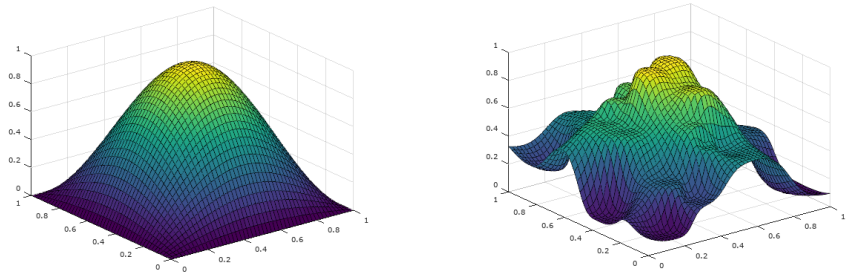

Figure 6: Szórt pontú interpoláció 30 véletlenszerűen választott alappontra. Tesztfelület (balra) ill. Shepard-interpoláció $p = 4$ mellett (jobbra)

Jóval pontosabb eljárás a *radiális bázisfüggvények módszere* (radial basis functions, RBF). Legyen $\Phi : \mathbf{R}^2 \to \mathbf{R}$ egy adott radiális (azaz körszimmetrikus) függvény: szemléletesen, $\Phi(x)$ értékei valójában csak az $||x||$ hossztól függenek. Keressük az interpolációs függvényt az alábbi alakban:

$$f(x) := \sum_{j=1}^{N} \alpha_j \Phi(x - x_j)$$

Az előre ismeretlen α_j együtthatókat az interpolációs feltételekből határozzuk meg:

$$\sum_{j=1}^{N} \alpha_j \Phi(x_k - x_j) = f_k \qquad (k = 1, 2, ...N)$$

ami egy lineáris egyenletrendszert jelent az ismeretlen együtthatókra nézve. Miután az α_j együtthatókat már kiszámítottuk, $f(x)$ kiértékelése tetszőleges

x mellett nehézség nélkül (és numerikusan olcsón, $\mathcal{O}(N)$ művelet árán) elvégezhető. Viszont az interpolációs egyenletrendszer megoldása általában $\mathcal{O}(N^3)$ műveletet igényel (Gauss-eliminációt alkalmazva), ami nagy N mellett sokszor elfogadhatatlanul sok. További numerikus nehézséget okoz, hogy az interpolációs egyenletrendszer sokszor nagyon rosszul kondícionált, így a kerekítési hibák erősen felhalmozódnak.

A legnépszerűbb radiális bázisfüggvények:

- Multikvadrikus módszer (method of multiquadrics, MQ):

  $$\Phi(x) := \sqrt{||x||^2 + c^2},$$

  ahol c adott paraméter (skálázó faktor).

- Inverz multikvadrikus módszer (iMQ):

  $$\Phi(x) := \frac{1}{\sqrt{||x||^2 + c^2}},$$

  ahol c adott paraméter (skálázó faktor).

- Vékony lemez módszer (thin plate splines, TPS):

  $$\Phi(x) := ||x||^2 \cdot \log ||x||,$$

  mely "önbeállós", azaz nem tartalmaz skálázó paramétert. A függvényt az origóban annak határértékeként értelmezzük: $\Phi(\mathbf{0}) := 0$, ahol felhasználtuk az elemi analízisből ismert $\lim\limits_{x \to 0} x^2 \log x = 0$ nevezetes határértéket.

- Gauss-függvények:

  $$\Phi(x) := e^{-c^2 \cdot ||x||^2},$$

  ahol c adott paraméter (skálázó faktor).

Összehasonlításképp meghatároztuk az előbbi, $f(x, y) := \sin \pi x \cdot \sin \pi y$ tesztfeladatra szintén 30 véletlenszerűen megválasztott interpolációs alappont esetén az MQ-interpolációval nyert interpolációs függvényt, $c := 1$ paraméterválasztással (ld. az ábrát). Látható, hogy az interpoláció pontossága sokkal jobb, mint a Shepard-interpoláció esetén. A pontosság ára az interpolációs egyenletrendszer felállításának és megoldásának szükségessége, ami nagy pontszám esetén komoly numerikus nehézségeket okozhat.

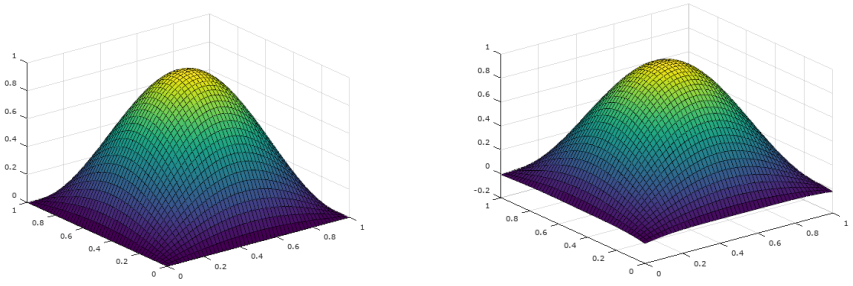

Figure 7: Szórt pontú interpoláció 30 véletlenszerűen választott alappontra. Tesztfelület (balra) ill. MQ-interpoláció $c = 1$ mellett (jobbra)

Megjegyzés: A vékony lemez módszer érdekes hasonlóságot mutat a harmadfokú spline-okkal. Megmutatható, hogy ha egy vékony, rugalmas, pl. acélból készült lemezt az interpolációs alappontokban adott magasságokban rögzítünk, az alappontok közt pedig hagyjuk olyan alakot felvenni, amit a saját rugalmassága enged, akkor az így kialakuló lemezfelület alakja nagy pontossággal épp a TPS-módszerrel kapott interpolációs függvény grafikonjával egyezik.

3.8 Feladatok

1. Adott $x_0 < x_1 < ... < x_N$ alappontrendszerhez tekintsük a megfelelő $\ell_k^{(N)}$ Lagrange-féle alappolinomokat $(k = 0, 1, ..., N)$. Mutassuk meg, hogy:

$$\sum_{k=0}^{N} \ell_k^{(N)}(x) \equiv 1$$

Megoldás: Az $f(x) :\equiv 1$ formulával definiált függvény megegyezik saját Lagrange-interpolációs polinomjával, ami kifejezhető a Lagrange-alappolinomok segítségével. Innen:

$$\sum_{k=0}^{N} 1 \cdot \ell_k^{(N)}(x) = f(x) \equiv 1$$

2. Adott $x_0 < x_1 < ... < x_N$ alappontrendszerhez, ahol $N \geq 1$, tekintsük a megfelelő $\ell_k^{(N)}$ Lagrange-féle alappolinomokat $(k = 0, 1, ..., N)$. Mutassuk meg, hogy:

$$\sum_{k=0}^{N} x_k \cdot \ell_k^{(N)}(x) \equiv x$$

Megoldás: Az $f(x) := x$ formulával definiált függvény megegyezik saját Lagrange-interpolációs polinomjával ($N \geq 1$-re), ami kifejezhető a Lagrange-alappolinomok segítségével. Innen:

$$\sum_{k=0}^{N} x_k \cdot \ell_k^{(N)}(x) = f(x) \equiv x.$$

3. Az $f(x) := x^6$ formulával definiált függvényt interpoláljuk az $x_0 := 0$, $x_1 := \frac{1}{3}$, $x_2 := \frac{2}{3}$, $x_3 := 1$ alappontokban, Lagrange-interpolációval. Az interpolációs polinom kiszámítása nélkül becsüljük meg a közelítés hibáját az $x := 0.8$ helyen. Majd határozzuk meg a Lagrange-interpolációs polinomot, és hasonlítsuk össze a tényleges hibát a becsléssel.

Megoldás: Használjuk a hibabecslő formulát (ez megtehető, mert most

f képlettel adott). Nyilván $f^{IV}(x) = 6 \cdot 5 \cdot 4 \cdot 3 \cdot x^2$, innen tetszőleges $x \in [0,1]$-re:

$$\frac{|f^{IV}(x)|}{4!} \leq \frac{6 \cdot 5 \cdot \ 4 \cdot 3}{4 \cdot 3 \cdot 2 \cdot 1} = 15.$$

Jelölje L_3 a szóbanforgó legfeljebb 3-adfokú Lagrange-interpolációs polinomot, akkor a hibabecslő formula szerint:

$$|f(0.8) - L_3(0.8)| \leq 15 \cdot (0.8-0) \cdot (0.8-\frac{1}{3}) \cdot (0.8-\frac{2}{3}) \cdot (1-0.8) = 0.14933$$

A hiba tehát legfeljebb 0.14933.

Most számítsuk ki L_3 interpolációs polinom a_0, a_1, a_2, a_3 együtthatóit. Tudjuk, hogy ezek kielégítik az alábbi egyenletrendszert:

$$\begin{pmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ f(x_3) \end{pmatrix},$$

melynek megoldása most (ellenőrizzük!): $a_0 = 0$, $a_1 = 0.61728$, $a_2 = -2.95062$, $a_3 = 3.3333$. Innen a tényleges hiba számítható:

$$|f(0.8) - L_3(0.8)| = 0.04995$$

ami kb. harmada a becsült értéknek. A hibabecslés tehát nem túl éles, de sokkal könnyebben számítható, mint a tényleges hiba.

4. Határozzuk meg azt a legfeljebb másodfokú polinomot, mely az $x_0 := -1$, $x_1 := 0$, $x_2 := 1$ helyeken rendre az $f_0 := 0$, $f_1 := 2$ ill. az $f_2 := 0$ értékeket veszi fel.

Megoldás: A feladat az adatokra illeszkedő legfeljebb másodfokú Lagrange-interpolációs polinom előállítása.

1. megoldás: Keressük az interpolációs polinomot

$$P(x) = a_0 + a_1 x + a_2 x^2$$

alakban. Az egyelőre ismeretlen a_0, a_1, a_2 együtthatók a $P(x_k) = f_k$ ($k = 0, 1, 2$) interpolációs egyenlőségekből határozhatók meg. Jelen esetben ez az egyenletrendszer:

$$\begin{aligned} a_0 - a_1 + a_2 &= 0 \\ a_0 &= 2 \\ a_0 + a_1 + a_2 &= 0 \end{aligned}$$

Az egyenletrendszer megoldása: $a_0 = 2$, $a_1 = 0$, $a_2 = -2$, így az interpolációs polinom: $P(x) = 2 - 2x^2$.

2. megoldás: Állítsuk elő az interpolációs polinomot

$$P(x) = f_0 \cdot \ell_0(x) + f_1 \cdot \ell_1(x) + f_2 \cdot \ell_2(x)$$

alakban, ahol ℓ_0, ℓ_1, ℓ_2 a Lagrange-féle alappolinomok:

$$\ell_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{x(x-1)}{(-1)\cdot(-2)} = \frac{1}{2}x^2 - \frac{1}{2}x$$

$$\ell_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(x+1)(x-1)}{1\cdot(-1)} = -x^2 + 1$$

$$\ell_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{(x+1)x}{2\cdot 1} = \frac{1}{2}x^2 + \frac{1}{2}x$$

Innen:

$$P(x) = 0 \cdot \left(\frac{1}{2}x^2 - \frac{1}{2}x\right) + 2 \cdot \left(-x^2 + 1\right) + 0 \cdot \left(\frac{1}{2}x^2 + \frac{1}{2}x\right) =$$

$$= -2x^2 + 2$$

3. megoldás: Osztott differenciák használata:

j	x_j	f_j		
0	-1	$\boxed{0}$		
			$\frac{2-0}{0-(-1)} = \boxed{2}$	
1	0	2		$\frac{-2-2}{1-(-1)} = \boxed{-2}$
			$\frac{0-2}{1-0} = -2$	
2	1	0		

ahonnan

$$P(x) = 0 + 2 \cdot (x+1) - 2 \cdot (x+1)(x-0) = 2 - 2x^2$$

5. Határozzuk meg azt a legfeljebb elsőfokú polinomot, mely az $x_0 := a$, $x_1 := b$ helyeken az $f_0 := f(a)$ ill. $f_1 := f(b)$ értékeket veszi fel.

Megoldás: A feladat az adatokra illeszkedő legfeljebb elsőfokú Lagrange-interpolációs polinom előállítása.

1. megoldás: Állítsuk elő az interpolációs polinomot

$$P(x) = f_0 \cdot \ell_0(x) + f_1 \cdot \ell_1(x)$$

alakban, ahol ℓ_0, ℓ_1 a Lagrange-féle alappolinomok:

$$\ell_0(x) = \frac{x-b}{a-b}$$

$$\ell_1(x) = \frac{x-a}{b-a}$$

Innen:

$$P(x) = f(a) \cdot \frac{x-b}{a-b} + f(b) \cdot \frac{x-a}{b-a}$$

2. megoldás: Osztott differenciák használata:

$$\begin{array}{llll} j & x_j & f_j & \\ 0 & a & \boxed{f(a)} & \\ & & & \boxed{\dfrac{f(b)-f(a)}{b-a}} \\ 1 & b & f(b) & \end{array}$$

ahonnan

$$P(x) = f(a) + \frac{f(b)-f(a)}{b-a} \cdot (x-a)$$

ami megegyezik az előző eredménnyel (ellenőrizzük!).

6. Határozzuk meg azt a legfeljebb másodfokú polinomot, mely az $x_0 := 1$, $x_1 := 4$, $x_2 := 5$ helyeken rendre az $f_0 := 1$, $f_1 := 19$ ill. az $f_2 := 29$ értékeket veszi fel.

Megoldás: A feladat az adatokra illeszkedő legfeljebb másodfokú Lagrange-interpolációs polinom előállítása. Használjunk osztott diffe-

renciákat (de ajánlatos végigszámolni a másik két módszerrel is a feladatot):

$$\begin{array}{cccc} j & x_j & f_j & \\ 0 & 1 & \boxed{1} & \\ & & & \frac{19-1}{4-1} = \boxed{6} \\ 1 & 4 & 19 & \qquad\qquad \frac{10-6}{5-1} = \boxed{1} \\ & & & \frac{29-19}{5-4} = 10 \\ 2 & 5 & 29 & \end{array}$$

ahonnan

$$P(x) = 1 + 6 \cdot (x-1) + 1 \cdot (x-1)(x-4) = -1 + x + x^2$$

7. Határozzuk meg azt a legfeljebb másodfokú polinomot, mely az $x_0 := -1$, $x_1 := 1$, $x_2 := 2$ helyeken rendre az $f_0 := -1$, $f_1 := 1$ ill. az $f_2 := 3$ értékeket veszi fel.

Megoldás: A feladat az adatokra illeszkedő legfeljebb másodfokú Lagrange-interpolációs polinom előállítása. Használjunk osztott differenciákat (de ajánlatos végigszámolni a másik két módszerrel is a feladatot):

$$\begin{array}{cccc} j & x_j & f_j & \\ 0 & -1 & \boxed{-1} & \\ & & & \frac{1-(-1)}{1-(-1)} = \boxed{1} \\ 1 & 1 & 1 & \qquad\qquad \frac{2-1}{2-(-1)} = \boxed{\frac{1}{3}} \\ & & & \frac{3-1}{2-1} = 2 \\ 2 & 2 & 3 & \end{array}$$

ahonnan

$$P(x) = -1 + 1 \cdot (x+1) + \frac{1}{3} \cdot (x+1)(x-1) = \frac{1}{3}x^2 + x - \frac{1}{3}.$$

8. Határozzuk meg azt a legfeljebb harmadfokú polinomot, mely az $x_0 := -2$, $x_1 := 0$, $x_2 := 1$, $x_3 := 3$ helyeken rendre az $f_0 := -15$, $f_1 := -1$, $f_2 := 3$ ill. az $f_3 := 80$ értékeket veszi fel.

   Megoldás: A feladat az adatokra illeszkedő legfeljebb harmadfokú Lagrange-interpolációs polinom előállítása. Használjunk osztott differenciákat (de ajánlatos végigszámolni a másik két módszerrel is a feladatot):

$$
\begin{array}{llll}
j & x_j & f_j & \\
0 & -2 & \boxed{-15} & \\
 & & \frac{-1+15}{2} = \boxed{7} & \\
1 & 0 & -1 & \frac{4-7}{3} = \boxed{-1} \\
 & & \frac{3+1}{1} = 4 & \frac{11.5+1}{5} = \boxed{2.5} \\
2 & 1 & 3 & \frac{38.5-4}{3} = 11.5 \\
 & & \frac{80-3}{2} = 38.5 & \\
3 & 3 & 80 &
\end{array}
$$

   ahonnan

$$P(x) = -15+7\cdot(x+2)-1\cdot(x+2)x+2.5\cdot(x+2)x(x-1) = 2.5x^3+1.5x^2-1.$$

9. Közelítsük $f(x) := \cos x$ formulával értelmezett függvényt a $\left[0, \frac{\pi}{2}\right]$ intervallumon kétpontos Hermite-interpolációs polinommal.

   Megoldás: Jelölje $x_0 := 0$, $x_1 := \frac{\pi}{2}$ (interpolációs alappontok). A végpontokban a függvényértékek és a deriváltértékek: $f_0 = f(x_0) = 1$, $f_1 = f(x_1) = 0$, $f_0' = f'(x_0) = 0$, $f_1' = f'(x_1) = -1$.

   A kétpontos, harmadfokú Hermite-interpolációs polinom:

$$H(x) = A + B \cdot \frac{x - x_0}{h} + C \cdot \frac{(x - x_0)^2}{h^2} + D \cdot \frac{(x - x_0)^3}{h^3},$$

   ahol $h := x_1 - x_0$. Az A, B, C, D együtthatók értéke a konkrét adatokkal:

$$
\begin{array}{lrrrrl}
A = & f_0 & & & & = 1 \\
B = & & & hf_0' & & = 0 \\
C = & -3f_0 & +3f_1 & -2hf_0' & -hf_1' & = -3 + h \\
D = & 2f_0 & -2f_1 & +hf_0' & +hf_1' & = 2 - h
\end{array}
$$

ahonnan:

$$H(x) = 1 + \frac{-3+h}{h^2} \cdot x^2 + \frac{2-h}{h^3} \cdot x^3 =$$
$$= 1 - 0.57923x^2 + 0.11074x^3.$$

Megjegyzés: Érdemes MATLAB-ban megjeleníteni egy ábrán az eredeti függvényt és az Hermite-interpolációs függvényt is.

10. Közelítsük $f(x) := \cos x$ formulával értelmezett függvényt az $x_0 := -\frac{\pi}{2}$, $x_1 := 0$, $x_2 := \frac{\pi}{2}$ alappontokon harmadfokú spline függvénnyel. (A végpontokban a deriváltértékek egyezzenek meg a fenti függvény deriváltjaival.)

Megoldás: Az interpolációs alappontokban a függvényértékek és a deriváltértékek: $f_0 = f(x_0) = 0$, $f_1 = f(x_1) = 1$, $f_2 = f(x_2) = 0$, $f_0' = f'(x_0) = 1$, f_1' ismeretlen, $f_2' = f'(x_2) = -1$.

Az ismeretlen f_1' deriváltértékre felírható egyenlet (ahol $h = \frac{\pi}{2}$):

$$f_0' + 4f_1' + f_2' = \frac{-3f_0 + 3f_2}{h} = 0,$$

ahonnan $f_1' = 0$. Innen a spline függvény az $[x_0, x_1]$ és az $[x_1, x_2]$ intervallumokban már egy-egy kétpontos Hermite-interpolációval határozható meg, az előző feladatok mintájára.

Eredmény: az $[x_0, x_1]$ intervallumon:

$$H_0(x) = (x+h) + \frac{3-2h}{h^2} \cdot (x+h)^2 + \frac{-2+h}{h^3} \cdot (x+h)^3$$

Az $[x_1, x_2]$ intervallumon pedig:

$$H_1(x) = 1 + \frac{-3+h}{h^2} \cdot x^2 + \frac{2-h}{h^3} \cdot x^3$$

Megjegyzés: Érdemes MATLAB-ban megjeleníteni egy ábrán az eredeti függvényt és a spline interpolációs függvényt is.

11. Közelítsük $f(x) := \cos 2x$ formulával értelmezett függvényt az $x_0 := -\frac{\pi}{4}$, $x_1 := 0$, $x_2 := \frac{\pi}{4}$ alappontokon harmadfokú spline függvénnyel. (A végpontokban a deriváltértékek egyezzenek meg a fenti függvény deriváltjaival.)

Megoldás: Az interpolációs alappontokban a függvényértékek és a deriváltértékek: $f_0 = f(x_0) = 0$, $f_1 = f(x_1) = 1$, $f_2 = f(x_2) = 0$, $f_0' = f'(x_0) = 2$, f_1' ismeretlen, $f_2' = f'(x_2) = -2$.

Az ismeretlen f_1' deriváltértékre felírható egyenlet (ahol $h = \frac{\pi}{4}$):

$$f_0' + 4f_1' + f_2' = \frac{-3f_0 + 3f_2}{h} = 0,$$

ahonnan $f_1' = 0$. Innen a spline függvény az $[x_0, x_1]$ és az $[x_1, x_2]$ intervallumokban már egy-egy kétpontos Hermite-interpolációval határozható meg, az előző feladatok mintájára.

Megjegyzés: Érdemes MATLAB-ban megjeleníteni egy ábrán az eredeti függvényt és a spline interpolációs függvényt is.

12. Közelítsük $f(x) := \cos 2x$ formulával értelmezett függvényt az $x_0 := -\frac{\pi}{4}$, $x_1 := -\frac{\pi}{12}$, $x_2 := \frac{\pi}{12}$, $x_3 := \frac{\pi}{4}$ alappontokon harmadfokú spline függvénnyel. (A végpontokban a deriváltértékek egyezzenek meg a fenti függvény deriváltjaival.)

Megoldás: Az interpolációs alappontokban a függvényértékek és a deriváltértékek: $f_0 = f(x_0) = 0$, $f_1 = f(x_1) = \frac{\sqrt{3}}{2}$, $f_2 = f(x_2) = \frac{\sqrt{3}}{2}$, $f_3 = f(x_3) = 0$, $f_0' = f'(x_0) = 2$, f_1' ismeretlen, f_2' ismeretlen, $f_3' = f'(x_3) = -2$.

Az ismeretlen f_1', f_2' deriváltértékre felírható egyenletrendszer (ahol $h = \frac{\pi}{6}$):

$$f_0' + 4f_1' + f_2' = \frac{-3f_0 + 3f_2}{h}$$

$$f_1' + 4f_2' + f_3' = \frac{-3f_1 + 3f_3}{h}$$

ahonnan $f_1' = \frac{3\sqrt{3}}{\pi} - \frac{2}{3}$, és $f_2' = -\frac{3\sqrt{3}}{\pi} + \frac{2}{3}$. Innen a spline függvény az $[x_0, x_1]$, az $[x_1, x_2]$ és az $[x_2, x_3]$ intervallumokban már egy-egy

kétpontos Hermite-interpolációval határozható meg, az előző feladatok mintájára.

Megjegyzés: Érdemes MATLAB-ban megjeleníteni egy ábrán az eredeti függvényt és a spline interpolációs függvényt is.

4 Közelítő integrálás

4.1 Motiváció

A gyakorlati matematikai modellezésben számtalanszor van szükség bizonyos határozott integrálok kiszámítására (akár többváltozós függvények esetében is). Ezek "pontos", azaz formulával való meghatározása legtöbbször még akkor sem lehetséges, ha az integrandusz maga is formulával adott, mert sokszor nincs elemi függvényekkel kifejezhető primitív függvény. Ha pedig mégis létezik primitív függvény, az sokszor annyira bonyolult, hogy nem éri meg a "pontos" Newton-Leibniz-tételt használni.

Ha már közelítő módszert használunk, természetes elvárás, hogy a közelítés pontosságát is meg tudjuk becsülni.

A Riemann-integrál definíciója maga is egy közelítő formulának tekinthető. Emlékeztetünk rá, hogy ha $f : [a, b] \to \mathbf{R}$ egy folytonos függvény és $a = x_1 < x_2 < ... < x_N = b$ az $[a, b]$ intervallum egy felbontása, $\xi_j \in [x_j, x_{j+1}]$ tetszőleges pontok ($j = 0, ..., N - 1$), akkor a

$$\sum_{j=0}^{N-1} f(\xi_j) \cdot (x_{j+1} - x_j)$$

Riemann-integrálközelítő összegek az $\int_a^b f(x)\,dx$ Riemann-integrálhoz tartanak, ha a felbontás minden határon túl finomodik. Elegendően finom felbontás esetén tehát egy konkrét Riemann-integrálközelítő összeg kiszámítása is tekinthető egy közelítő integrálási technikának. A probléma ezzel az, hogy a konvergencia esetleg nagyon lassú is lehet, és az eljárás hibája – egyéb információ híján – nagyon nehezen becsülhető. Így tehát a Riemann-integrálközelítő összegek használata integrálok közelítésére a gyakorlatban alkalmatlan. Ezért van szükség a témakör finomabb tanulmányozására, különös tekintettel az integrálközelítő formulák hibabecslésére.

Ebben a fejezetben néhány klasszikus integrálközelítő módszert (kvadratúraformulát) vezetünk be és vizsgálunk. A fejezet végén röviden érintjük a többváltozós integrálok közelítésének problémáját is.

4.2 Interpolációs kvadratúrák

Legyen $[a, b] \subset \mathbf{R}$ egy véges intervallum, $f : [a, b] \to \mathbf{R}$ pedig egy adott folytonos függvény. Korábbról már tudjuk, hogy akkor f biztosan Riemann-integrálható $[a, b]$-n.

Az *interpolációs kvadratúrák* alapötlete az, hogy az $[a, b]$ intervallumban felvéve egy $a \leq x_1 < x_1 < ... < x_N \leq b$ alappontrendszert, egy Lagrange-interpolációs L_N polinomot állítunk elő az $f_0 := f(x_0)$, $f_1 := f(x_1)$, ..., $f_N := f(x_N)$ függvényértékekre alapozva, és az $\int_a^b f(x)\,dx$ integrált az interpolációs polinom $\int_a^b L_N(x)\,dx$ integráljával közelítjük. Ez utóbbi integrált nyilván már pontosan ki tudjuk számítani.

Az L_N interpolációs polinom az $\ell_j^{(N)}$ Lagrange-féle alappolinomok segítségével explicite is felírható (ld. az Interpoláció c. fejezetet):

$$L_N(x) = \sum_{j=0}^{N} f(x_j \cdot \ell_j^{(N)}(x),$$

ahonnan:

$$\int_a^b L_N(x)\,dx = \sum_{j=0}^{N} f(x_j) \cdot \int_a^b \ell_j^{(N)}(x)\,dx =: \sum_{j=0}^{N} a_j f(x_j),$$

ahol $a_j := \int_a^b \ell_j^{(N)}\,dx$ a kvadratúra *súlyai*. Így a keresett integrál közelítése ez lesz (*interpolációs kvadratúra*):

$$\int_a^b f(x)\,dx \approx I_N(f) := \sum_{j=0}^{N} a_j f(x_j)$$

Állítás: Az I_N interpolációs kvadratúra pontos minden legfeljebb N-edfokú polinomra.

Ha ugyanis f egy legfeljebb N-edfokú polinom, akkor annak L_N Lagrange-interpolációs polinomja – a Lagrange-interpoláció egyértelműsége miatt – önmagával egyezik.

Következésképp az I_N interpolációs kvadratúra a_0, a_1, ..., a_N súlyai a nehézkes $a_j = \int_a^b \ell_j^{(N)}(x)\,dx$ formula helyett egy lineáris egyenletrendszer megoldásából is számíthatók. Elegendő a kvadratúraformulát az 1, x, x^2, ..., x^N alappolinomokra alkalmazni:

$$\sum_{j=0}^{N} x_j^k \cdot a_j = I_N(x^k) = \int_a^b x^k\,dx \qquad (k = 0, 1, ..., N)$$

Másképp felírva:

$$\begin{pmatrix} 1 & 1 & ... & 1 \\ x_0 & x_1 & ... & x_N \\ x_0^2 & x_1^2 & ... & x_N^2 \\ ... & ... & ... & ... \\ x_0^N & x_1^N & ... & x_N^N \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ ... \\ a_N \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \mu_1 \\ ... \\ \mu_N \end{pmatrix}, \tag{2}$$

ahol

$$\mu_k := \int_a^b x^k \, dx = \frac{b^{k+1} - a^{k+1}}{k+1} \qquad (k = 0, 1, ..., N)$$

Megjegyzések:

- Az, hogy egy kvadratúraformula milyen magas fokszámú polinomokra pontos, a közvetve kvadratúra hibájára van kihatással. Arról van szó ugyanis, hogy a f elég sima függvény (mondjuk $(n+1)$-szer folytonosan differenciálható), akkor véges Taylor-sorba fejthető a körül, $(n+1)$-edfokú maradéktaggal:

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + ... + \frac{f^{(n)}(a)}{n!} + \\ + (x-a)^n + \frac{f^{(n+1)}(\xi_x)}{(n+1)!}(x-a)^{n+1}$$

  alkalmas (x-től függő) $\xi_x \in [a, b]$ mellett. Ha most egy kvadratúraformula pontos minden, legfeljebb n-edfokú polinomra, akkor elég azt vizsgálni, hogy a maradéktag integrálját milyen jól közelíti a kvadratúra, mert a megelőző tagok integrálját pontosan adja. Általánosan elmondhatjuk, hogy minél magasabb fokszámú polinomra pontos a kvadratúra, annál kisebb a kvadratúra hibája (elegendően sima függvényekre).

- A Lagrange-interpolációnál elmondottakkal összhangban, általában itt is kerülni kell a magas fokszámú interpolációs polinomok használatát, mert az egyes alappontok közt az interpolációs polinom hibája nagyon nagy is lehet, ami a kvadratúra pontosságára is kihat. Az alappontok ügyes megválasztásával ez a jelenség tompítható, vagy akár teljesen meg is szüntethető. Ez vezet el az *ortogonális polinomrendszerek* alkalmazásához: ennek részleteivel azonban e jegyzet keretein belül nem foglalkozhatunk.

Speciális esetek: A leggyakrabban használt *elemi kvadratúraformulák* a 0-, 1- ill. 2-fokú interpolációs kvadratúrák:

1. Legyen $N = 0$, és az egyetlen x_0 alappont pedig $x_0 := \frac{a+b}{2}$. Akkor az (2) egyenletrendszer csak egyismeretlenes:

$$1 \cdot a_0 = \mu_0 = \int_a^b 1 \, dx = b - a.$$

Innen a kvadratúraformula:

$$\int_a^b f(x) \, dx \approx I_0(f) = f\left(\frac{a+b}{2}\right) \cdot (b-a) \tag{3}$$

Ezt *középpont szabálynak* vagy *érintőformulának* nevezzük. Az utóbbi elnevezést az indokolja, hogy ha az $x_0 := \frac{a+b}{2}$ helyen érintőt húzunk az f függvény grafikonjához, és az integrált az érintő alatti területtel közelítjük, akkor épp az érintőformulához jutunk (ellenőrizzük!).

Állítás: Az érintőformula pontos minden, legfeljebb *elsőfokú* polinomra. (Ez jobb eredmény a vártnál, tekintve, hogy most $N = 0$.)

Az állítást nyilván elég csak a $p_0(x) :\equiv 1$ és az $p_1(x) := x$ képletű alappolinomokra ellenőrizni:

$$I_0(p_0) = 1 \cdot (b-a) = \int_a^b 1 \, dx$$

$$I_0(p_1) = \frac{a+b}{2} \cdot (b-a) = \frac{b^2-a^2}{2} = \int_a^b x \, dx$$

2. Legyen $N = 1$, és az alappontok legyenek $x_0 := a$, $x_1 := b$. Akkor az (2) egyenletrendszer alakja most:

$$\begin{pmatrix} 1 & 1 \\ a & b \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} = \begin{pmatrix} b-a \\ \frac{b^2-a^2}{2} \end{pmatrix}$$

aminek egyetlen megoldása (ellenőrizzük!): $a_0 = a_1 = \frac{b-a}{2}$

Innen a kvadratúraformula:

$$\int_a^b f(x) \, dx \approx I_1(f) = \frac{f(a)+f(b)}{2} \cdot (b-a) \tag{4}$$

Ezt *trapézformulának* nevezzük. Az elnevezést az indokolja, hogy ha az $x = a$ és $x = b$ pontokból egy szelőt húzunk az f függvény grafikonjához, és az integrált a szelő alatti területtel (ami egy trapéz) közelítjük, akkor épp az trapézformulához jutunk (ellenőrizzük!).

Állítás: A trapézformula pontos minden, legfeljebb elsőfokú polinomra.

Az állítást nyilván elég csak a $p_0(x) :\equiv 1$ és az $p_1(x) := x$ képletű alappolinomokra ellenőrizni:

$$I_1(p_0) = \frac{1+1}{2} \cdot (b-a) = \int_a^b 1\,dx$$

$$I_1(p_1) = \frac{a+b}{2} \cdot (b-a) = \frac{b^2-a^2}{2} = \int_a^b x\,dx$$

3. Legyen $N = 2$, és az alappontok pedig legyenek $x_0 := a$, $x_1 := \frac{a+b}{2}$, $x_2 := b$. Akkor az (2) egyenletrendszer alakja most:

$$\begin{pmatrix} 1 & 1 & 1 \\ a & \frac{a+b}{2} & b \\ a^2 & \left(\frac{a+b}{2}\right)^2 & b^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} b-a \\ \frac{b^2-a^2}{2} \\ \frac{b^3-a^3}{3} \end{pmatrix}$$

Az egyenletrendszer egyetlen megoldása:

$$a_0 = \frac{b-a}{6}, \qquad a_1 = 4 \cdot \frac{b-a}{6}, \qquad a_2 = \frac{b-a}{6}$$

Valóban, a_0, a_1, a_2 kifejezéseit az egyenletrendszerbe visszahelyettesítve:

Első egyenlet:

$$a_0 + a_1 + a_2 = \frac{b-a}{6} \cdot (1+4+1) = b-a$$

Második egyenlet:

$$a \cdot a_0 + \frac{a+b}{2} \cdot a_1 + b \cdot a_2 = \frac{b-a}{6} \cdot \left(a + 4 \cdot \frac{a+b}{2} + b\right) =$$

$$= \frac{b-a}{6} \cdot (3a+3b) = \frac{b^2-a^2}{2}$$

Harmadik egyenlet:

$$a^2 \cdot a_0 + \left(\frac{a+b}{2}\right)^2 \cdot a_1 + b^2 \cdot a_2 = \frac{b-a}{6} \cdot \left(a^2 + 4 \cdot \frac{(a+b)^2}{4} + b^2\right) =$$

$$= \frac{b-a}{6} \cdot (2a^2 + 2ab + 2b^2) = \frac{b^3 - a^3}{3}$$

Innen a kvadratúraformula (*Simpson-formula*):

$$\int_a^b f(x)\,dx \approx I_2(f) = \frac{f(a) + 4 \cdot f(\frac{a+b}{2}) + f(b)}{6} \cdot (b-a) \qquad (5)$$

Állítás: Simpson-formula pontos minden, legfeljebb *harmadfokú* polinomra. (Ez jobb eredmény a vártnál, tekintve, hogy most $N = 2$.)

Az állítást nyilván elég csak a $p_0(x) :\equiv 1$, $p_1(x) := x$, $p_2(x) := x^2$ és a $p_3(x) := x^3$ képletű alappolinomokra ellenőrizni:

$$I_2(p_0) = \frac{1+4+1}{6} \cdot (b-a) = \int_a^b 1\,dx$$

$$I_2(p_1) = \frac{a + 4\frac{a+b}{2} + b}{6} \cdot (b-a) = \frac{3a+3b}{6} \cdot (b-a) =$$

$$= \frac{b^2 - a^2}{2} = \int_a^b x\,dx$$

$$I_2(p_2) = \frac{a^2 + 4(\frac{a+b}{2})^2 + b^2}{6} \cdot (b-a) = \frac{2a^2 + 2ab + 2b^2}{6} \cdot (b-a) =$$

$$= \frac{b^3 - a^3}{3} = \int_a^b x^2\,dx$$

$$I_2(p_3) = \frac{a^3 + 4(\frac{a+b}{2})^3 + b^3}{6} \cdot (b-a) = \frac{3a^3 + 3a^2b + 3ab^2 + 3b^3}{12} \cdot (b-a) =$$

$$= \frac{b^4 - a^4}{4} = \int_a^b x^3\,dx$$

Megjegyzés: Mivel a formulákban a koordinátarendszer megválasztása nyilván közömbös, feltehető, hogy az intervallum középpontja épp a 0, azaz $[a,b] = [-\frac{h}{2}, \frac{h}{2}]$, ahol $h := b - a$ jelöli az intervallum hosszát. Ekkor a Lagrange-interpolációs polinomok pl. osztott differenciákkal egyszerűen

meghatározhatók, ahonnan a fenti elemi kvadratúraformulák közvetlenül is levezethetők.

Érintőformula: Itt $N = 0$, $x_0 = 0$, és a Lagrange-interpolációs polinom nyilván

$$L_0(x) \equiv f_0 = f(x_0),$$

és ezért:

$$I_0(f) = \int_{-h/2}^{h/2} L_0(x)\,dx = \\ = f_0 \cdot h$$

Trapézformula: Itt $N = 1$, $x_0 = -\frac{h}{2}$, $x_1 = \frac{h}{2}$. Az osztott differenciák táblázata:

$$\begin{array}{ll|l} -\frac{h}{2} & \boxed{f_0} & \\ & & \boxed{\dfrac{f_1 - f_0}{h}} \\ \frac{h}{2} & f_1 & \end{array}$$

Ezért az interpolációs polinom:

$$L_1(x) = f_0 + \frac{f_1 - f_0}{h} \cdot \left(x + \frac{h}{2}\right)$$

Ezt integrálva a $[-\frac{h}{2}, \frac{h}{2}]$ intervallumon, a kvadratúraformulát kapjuk:

$$I_1(f) = \int_{-h/2}^{h/2} L_1(x)\,dx = \int_{-h/2}^{h/2} \left(f_0 + \frac{f_1 - f_0}{h} \cdot x + \frac{f_1 - f_0}{2}\right) dx = \\ = \frac{f_0 + f_1}{2} \cdot h$$

Simpson-formula: Itt $N = 2$, $x_0 = -\frac{h}{2}$, $x_1 = 0$, $x_2 = \frac{h}{2}$. Az osztott differenciák táblázata:

$$\begin{array}{ll|ll} -\frac{h}{2} & \boxed{f_0} & & \\ & & \boxed{\dfrac{2f_1 - 2f_0}{h}} & \\ 0 & f_1 & & \boxed{\dfrac{2f_2 - 4f_1 + 2f_0}{h^2}} \\ & & \dfrac{2f_2 - 2f_1}{h} & \\ \frac{h}{2} & f_2 & & \end{array}$$

Így az interpolációs polinom:

$$L_1(x) = f_0 + \frac{2f_1 - 2f_0}{h} \cdot \left(x + \frac{h}{2}\right) + \frac{2f_2 - 4f_1 + 2f_0}{h^2} \cdot \left(x + \frac{h}{2}\right) \cdot x$$

Ezt integrálva a $[-\frac{h}{2}, \frac{h}{2}]$ intervallumon, a kvadratúraformulát kapjuk:

$$I_2(f) = \int_{-h/2}^{h/2} L_2(x)\,dx =$$

$$= \int_{-h/2}^{h/2} \left(f_0 + \frac{2f_1 - 2f_0}{h} \cdot x + f_1 - f_0 + \frac{2f_2 - 4f_1 + 2f_0}{h^2} \cdot x^2 + \frac{f_2 - 2f_1 + f_0}{h} \cdot x\right) dx$$

$$= \left(f_0 + f_1 - f_0 + \frac{2f_2 - 4f_1 + 2f_0}{h^2} \cdot 2 \cdot \frac{h^2}{3 \cdot 8}\right) \cdot h =$$

$$= \left(f_0 + f_1 - f_0 + \frac{f_2}{6} - \frac{2f_1}{6} + \frac{f_0}{6}\right) \cdot h =$$

$$= \frac{f_0 + 4f_1 + f_2}{6} \cdot h.$$

Magasabb fokszámú interpolációs polinomot csak nagyon speciális alappontválasztás mellett szokás alkalmazni (a Lagrange-interpoláció már említett numerikus hátrányai miatt). Ehelyett az integrálás pontosság úgy növelhető, hogy az eredeti $[a, b]$ intervallumot (nem feltétlen egyenlő hosszúságú) részintervallumokra bontjuk, és mindegyik részintervallumon valamelyik előző "elemi" kvadratúraformulát alkalmazzuk. Így nyerjük az *összetett kvadratúraformulákat*.

4.3 Összetett kvadratúraformulák

Legyenek tehát $a = x_0 < x_1 < ... < x_N = b$ alappontok, és jelölje $f_j := f(x_j)$, és legyen $h_j := x_{j+1} - x_j$ a j-edik részintervallum hossza $(j = 0, 1, ..., N)$.

Összetett érintőformula:

$$\int_a^b f(x)\,dx = \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} f(x)\,dx \approx \sum_{j=0}^{N-1} f_{j+1/2} \cdot h_j$$

ahol $f_{j+1/2} := f(\frac{x_j + x_{j+1}}{2})$.

Összetett trapézformula:

$$\int_a^b f(x)\,dx = \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} f(x)\,dx \approx \sum_{j=0}^{N-1} \frac{f_j + f_{j+1}}{2} \cdot h_j$$

Összetett Simpson-formula:

$$\int_a^b f(x)\,dx = \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} f(x)\,dx \approx \sum_{j=0}^{N-1} \frac{f_j + 4f_{j+1/2} + f_{j+1}}{6} \cdot h_j$$

Speciálisan, ha a felbontás egyenközű, $h_j = h := \frac{b-a}{N}$, akkor a formulák leegyszerűsödnek:

- Összetett érintőformula:

$$\sum_{j=0}^{N-1} f_{j+1/2} \cdot h_j = (f_{1/2} + f_{3/2} + ... + f_{N-1/2}) \cdot h$$

- Összetett trapézformula:

$$\sum_{j=0}^{N-1} \frac{f_j + f_{j+1}}{2} \cdot h_j = \left(\frac{1}{2}f_0 + f_1 + f_2 + ... + f_{N-1} + \frac{1}{2}f_N\right) \cdot h$$

- Összetett Simpson-formula:

$$\sum_{j=0}^{N-1} \frac{f_j + 4f_{j+1/2} + f_{j+1}}{6} \cdot h_j =$$

$$= \left(\frac{1}{6}f_0 + \frac{2}{3}f_{1/2} + \frac{1}{3}f_1 + \frac{2}{3}f_{3/2} + \frac{1}{3}f_2 + ... + \frac{1}{6}f_N\right) \cdot h$$

4.4 A kvadratúraformulák hibája

Csak a korábban említett három "elemi" kvadratúraformula hibájával foglalkozunk, de azok összetett változatait is vizsgáljuk.

Legyen tehát $[a, b]$ adott intervallum, $h := b - a$. Mint korábban, az is feltehető, hogy $[a, b] = [-\frac{h}{2}, \frac{h}{2}]$.

A hibabecslések a 0 körüli Taylor-sorfejtésen alapulnak. Emlékeztetünk rá, hogy ha $f \in C^{n+1}[a,b]$, azaz $(n+1)$-szer folytonosan differenciálható az $[a,b]$ intervallumon, akkor minden $x \in [a,b]$-re érvényes az alábbi (véges) Taylor-sorfejtés:

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + ... + \frac{f^{(n)}(0)}{n!}x^n + \frac{f^{(n+1)}(\xi_x)}{(n+1)!}x^{n+1}$$

alkalmas (x-től függő) $\xi_x \in [a,b]$ mellett. (A jobb oldal utolsó tagja a *Lagrange-féle maradéktag.*)

Tétel: Ha $f \in C^2[a,b]$, akkor az érintőformula hibájára érvényes az alábbi becslés:

$$\left| \int_a^b f(x)\,dx - I_0(f) \right| = \left| \int_a^b f(x)\,dx - f\left(\frac{a+b}{2}\right) \cdot h \right| \leq \frac{1}{24} \cdot \max_{a \leq x \leq b} |f''(x)| \cdot h^3$$

Bizonyítás: Fejtsük f-et a 0 körül véges Taylor-sorba, másodfokú maradéktaggal:

$$f(x) = f(0) + f'(0) \cdot x + \frac{1}{2} f''(\xi_x) \cdot x^2$$

Integráljuk mindkét oldalt a és b (azaz $-\frac{h}{2}$ és $\frac{h}{2}$) közt. A jobb oldalon az elsőfokú tag integrálja zérus, így:

$$\int_a^b f(x)\,dx = f(0) \cdot h + \frac{1}{2} \int_{-h/2}^{h/2} f''(\xi_x) \cdot x^2\,dx$$

A jobb oldal első tagja épp $I_0(f)$. Innen:

$$\left| \int_a^b f(x)\,dx - I_0(f) \right| \leq \frac{1}{2} \cdot \max_{a \leq x \leq b} |f''(x)| \cdot \int_{-h/2}^{h/2} x^2\,dx$$

A jobb oldali integrál értéke pedig $2 \cdot \frac{1}{3} \cdot (\frac{h}{2})^3$, ahonnan az állítás már következik.

A trapézformula hibájára hasonló becslés érvényes:

Tétel: Ha $f \in C^2[a,b]$, akkor a trapézformula hibájára fennáll, hogy:

$$\left| \int_a^b f(x)\,dx - I_1(f) \right| = \left| \int_a^b f(x)\,dx - \frac{f(a)+f(b)}{2} \cdot h \right| \leq \frac{1}{12} \cdot \max_{a \leq x \leq b} |f''(x)| \cdot h^3$$

Bizonyítás: A trapézformula hibabecslése:

$$\left|\int_a^b f(x)\,dx - I_1(f)\right| = \left|\int_a^b (f(x) - L_1(x))\,dx\right| \le \int_a^b |f(x) - L_1(x)|\,dx$$

A jobb oldalon most használjuk fel a Lagrange-interpoláció hibabecslő formuláját (ld. 3. fejezet), mely szerint:

$$|f(x) - L_1(x)| \le \frac{1}{2} \cdot \max_{a\le x\le b} |f''(x)| \cdot |\omega_2(x)| = \frac{1}{2} \cdot \max_{a\le x\le b} |f''(x)| \cdot (x-a)\cdot(b-x)$$

Innen azt kapjuk, hogy:

$$\left|\int_a^b f(x)\,dx - I_1(f)\right| \le \frac{1}{2} \cdot \max_{a\le x\le b} |f''(x)| \cdot \int_a^b (x-a)\cdot(b-x)\,dx$$

A jobb oldali integrál nehézség nélkül kiszámítható:

$$\int_a^b (x-a)\cdot(b-x)\,dx = \int_a^b (-x^2+(a+b)x-ab)\,dx = \left[-\frac{x^3}{3} + (a+b)\frac{x^2}{2} - abx\right]_a^b =$$

$$= -\frac{b^3-a^3}{3} + \frac{(a+b)(b^2-a^2)}{2} - ab(b-a) =$$

$$= (b-a)\cdot\left(-\frac{b^2+ab+a^2}{3} + \frac{(b+a)^2}{2} - ab\right) =$$

$$= \frac{b-a}{6}\cdot(-2b^2-2ab-2a^2+3b^2+6ab+3a^2-6ab) =$$

$$= \frac{b-a}{6}\cdot(b^2-2ab+a^2) = \frac{(b-a)^3}{6},$$

ahonnan az állítás már következik.

A Simpson-formula hibája ennél bonyolultabb számítás útján kapható meg. A részleteket elhagyva, az eredmény:

Tétel: Ha $f \in C^4[a,b]$, akkor a Simpson-formula hibájára fennáll, hogy:

$$\left|\int_a^b f(x)\,dx - I_2(f)\right| = \left|\int_a^b f(x)\,dx - \frac{f(a)+4f(\frac{a+b}{2})+f(b)}{5}\cdot h\right| \le$$

$$\le \frac{1}{720} \cdot \max_{a\le x\le b} |f^{IV}(x)| \cdot h^5.$$

A fenti becslésekből most már könnyen adódnak a megfelelő *összetett* kvadratúraformulákra vonatkozó hibabecslések is. Egyszerűség kedvéért tegyük fel, hogy az $[a,b]$ intervallum felbontása egyenközű (ekvidisztáns), azaz $a = x_0 < x_1 < ... < x_N = b$ felbontásban $x_k = x_0 + k \cdot h$ $(k = 0, 1, ..., N)$, ahol $h = \frac{b-a}{N}$.

Az összetett érintőformula hibája:

$$\left| \int_a^b f(x)\,dx - \sum_{j=0}^{N-1} f_{j+1/2} \cdot h \right| = \left| \sum_{j=0}^{N-1} \left(\int_{x_j}^{x_{j+1}} f(x)\,dx - f_{j+1/2} \cdot h \right) \right| \leq$$

$$\leq \sum_{j=0}^{N-1} \left| \int_{x_j}^{x_{j+1}} f(x)\,dx - f_{j+1/2} \cdot h \right| \leq \sum_{j=0}^{N-1} \frac{1}{24} \max_{a \leq x \leq b} |f''(x)| \cdot h^3 =$$

$$= \frac{1}{24} \max_{a \leq x \leq b} |f''(x)| \cdot h^2 \cdot \sum_{j=0}^{N-1} h =$$

$$= \frac{b-a}{24} \max_{a \leq x \leq b} |f''(x)| \cdot h^2$$

Az összetett trapézformula hibája:

$$\left| \int_a^b f(x)\,dx - \sum_{j=0}^{N-1} \frac{f_j + f_{j+1}}{2} \cdot h \right| = \left| \sum_{j=0}^{N-1} \left(\int_{x_j}^{x_{j+1}} f(x)\,dx - \frac{f_j + f_{j+1}}{2} \cdot h \right) \right| \leq$$

$$\leq \sum_{j=0}^{N-1} \left| \int_{x_j}^{x_{j+1}} f(x)\,dx - \frac{f_j + f_{j+1}}{2} \cdot h \right| \leq \sum_{j=0}^{N-1} \frac{1}{12} \max_{a \leq x \leq b} |f''(x)| \cdot h^3 =$$

$$= \frac{1}{12} \max_{a \leq x \leq b} |f''(x)| \cdot h^2 \cdot \sum_{j=0}^{N-1} h =$$

$$= \frac{b-a}{12} \max_{a \leq x \leq b} |f''(x)| \cdot h^2$$

Az összetett Simpson-formula hibája:

$$\left| \int_a^b f(x)\,dx - \sum_{j=0}^{N-1} \frac{f_j + 4f_{j+1/2} + f_{j+1}}{6} \cdot h \right| =$$

$$= \left| \sum_{j=0}^{N-1} \left(\int_{x_j}^{x_{j+1}} f(x)\,dx - \frac{f_j + 4f_{j+1/2} + f_{j+1}}{6} \cdot h \right) \right| \leq$$

$$\leq \sum_{j=0}^{N-1} \left| \int_{x_j}^{x_{j+1}} f(x)\,dx - \frac{f_j + 4f_{j+1/2} + f_{j+1}}{6} \cdot h \right| \leq$$

$$\leq \sum_{j=0}^{N-1} \frac{1}{720} \max_{a \leq x \leq b} |f^{IV}(x)| \cdot h^5 =$$

$$= \frac{1}{720} \max_{a \leq x \leq b} |f^{IV}(x)| \cdot h^4 \cdot \sum_{j=0}^{N-1} h =$$

$$= \frac{b-a}{720} \max_{a \leq x \leq b} |f^{IV}(x)| \cdot h^4$$

Röviden, az összetett érintő- és az összetett trapézformula egyaránt $\mathcal{O}(h^2)$ pontosságú, míg az összetett Simpson-formula ennél sokkal pontosabb, hibája $\mathcal{O}(h^4)$.

Megjegyzés: Kvadratúraformulák konstrukciójában nemcsak a Lagrange-interpoláció jöhet szóba. Éppoly joggal vehetünk spline interpolációt is, tekintettel annak nagyon kedvező tulajdonságaira (ld. a 3. fejezetet). Már láttuk, hogy ha $f \in C^4[a,b]$, és S_f jelöli a h lépésközű ekvidisztáns alappontrendszeren felírt harmadfokú spline függvényt, mely az alappontokon ugyanazokat az értékeket veszi fel mint f, akkor az eredeti f függvény és az S_f spline eltérésére a következő becslés áll:

$$|f(x) - S_f(x)| \leq C \cdot h^4$$

alkalmas, a h lépésköztől és x-től független C szám mellett. Innen, az

$$I_{spline}(f) := \int_a^b S_f(x)dx$$

formulával definiált kvadratúra hibája:

$$\left| \int_a^b f(x)\,dx - I_{spline}(f) \right| \leq \int_a^b |f(x) - S_f(x)|\,dx \leq C \cdot (b-a) \cdot h^4,$$

azaz pontossága nem jobb, mint az összetett Simpson-formuláé, viszont az összetett Simpson-formula realizálása lényegesen egyszerűbb, mint a spline interpoláció.

4.5 Kitekintés

Ebben az utolsó szakaszban néhány általánosítási irányt érintünk, a teljességre való törekedés igénye nélkül.

4.5.1 Közelítő integrálás végtelen intervallumon

Tegyük fel, hogy $f : \mathbf{R} \to \mathbf{R}$ olyan folytonos függvény, melynek abszolút értéke elég gyorsan csökken az origótól távolodva: $|f(x)| = \mathcal{O}(\frac{1}{|x|^3})$, azaz alkalmas $C \geq 0$ konstans mellett $|f(x)| \leq C \cdot \frac{1}{|x|^3}$. A feladat az

$$\int_{-\infty}^{+\infty} f(x)\,dx$$

integrál közelítő kiszámítása.

Az ötlet a következő: végezzünk el egy $x := \operatorname{ctg} t$ helyettesítést, akkor $dx = -\frac{1}{\sin^2 t}dt$; az új t változó pedig π és 0 között halad (ekkor fut x a $-\infty$ és $+\infty$ határok közt). A helyettesítés után nyert integrál:

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{\pi}^{0} \frac{f(\operatorname{ctg} t)}{-\sin^2 t}\,dt = \int_{0}^{\pi} \frac{f(\operatorname{ctg} t)}{\sin^2 t}\,dt$$

Az új integrandusz határértéke 0-ban és π-ben egyaránt 0 (miért?). Így az összetett trapézformula nehézség nélkül alkalmazható a $t_j := \frac{j\pi}{N}$ ($j = 0, 1, ..., N$) ekvidisztáns alappontokon. Az integrandusz t_0-ban és t_N-ben eltűnik, így az összetett trapézformula még egyszerűbb lesz:

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{0}^{\pi} \frac{f(\operatorname{ctg} t)}{\sin^2 t}\,dt \approx \sum_{j=1}^{N-1} \frac{f(\operatorname{ctg} t_j)}{\sin^2 t_j} \cdot \frac{\pi}{N}$$

tehát

$$\int_{-\infty}^{+\infty} f(x)\,dx \approx \sum_{j=1}^{N-1} \frac{f(\operatorname{ctg} \frac{j\pi}{N})}{\sin^2 \frac{j\pi}{N}} \cdot \frac{\pi}{N}$$

4.5.2 Táblázattal adott függvény integrálása

Ha az f függvény értékei csak egy $a \leq x_0 < x_1 < ... < x_N \leq b$ alappontrendszeren ismertek, akkor célszerű ezen adatokra spline interpolációt alkalmazni, majd a spline függvényt integrálni (részintervallumonként).

4.5.3 Nem ekvidisztáns alappontrendszerek használata

Legyen az egyszerűség kedvéért $[a,b] := [-1,1]$. Próbáljunk meg az egyszerű, ekvidisztáns alappontrendszer helyett olyan alappontrendszert találni, melyre alapozott interpolációs kvadratúra *minél magasabbfokú polinomokon pontos* (ezért a kvadratúra – remélhetőleg – minél kisebb hibával bír).

Ha pl. $N = 1$, és $x_0 = -\frac{1}{\sqrt{3}}$, $x_0 = \frac{1}{\sqrt{3}}$, akkor az ezekre épülő interpolációs kvadratúra a_0, a_1 súlyai kielégítik a (2) egyenletrendszert:

$$\begin{pmatrix} 1 & 1 \\ -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix},$$

melynek egyetlen megoldása: $\begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Tehát a megfelelő kvadratúra (jelölje G_1):

$$G_1(f) = 1 \cdot f(x_0) + 1 \cdot f(x_1) = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

Már tudjuk, hogy G_1 pontos a legfeljebb elsőfokú polinomokon (az (2) egyenletrendszer épp ezt jelenti). Annál meglepőbb, hogy G_1 minden, legfeljebb *harmadfokú* polinomon is pontos. Valóban, a $p_2(x) := x^2$ alappolinomon:

$$G_1(p_2) = x_0^2 \cdot a_0 + x_1^2 \cdot a_1 = \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 1 = \frac{2}{3} = \int_{-1}^{1} x^2\,dx$$

Továbbá a $p_3(x) := x^3$ alappolinomon is:

$$G_1(p_3) = x_0^3 \cdot a_0 + x_1^3 \cdot a_1 = -\frac{1}{3\sqrt{3}} \cdot 1 + \frac{1}{3\sqrt{3}} \cdot 1 = 0 = \int_{-1}^{1} x^3\,dx$$

Más példa: legyen $N := 2$ és legyenek $x_0 := -\frac{3}{\sqrt{15}}$, $x_1 := 0$, $x_2 := \frac{3}{\sqrt{15}}$. Akkor a (2) egyenletrendszer most:

$$\begin{pmatrix} 1 & 1 & 1 \\ -\frac{3}{\sqrt{15}} & 0 & \frac{3}{\sqrt{15}} \\ \frac{9}{15} & 0 & \frac{9}{15} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ \frac{2}{3} \end{pmatrix},$$

melynek egyetlen megoldása: $\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \frac{5}{9} \\ \frac{8}{9} \\ \frac{5}{9} \end{pmatrix}$. Ezért a megfelelő kvadratúra (jelölje G_2):

$$G_2(f) = \frac{5}{9}\cdot f(x_0) + \frac{8}{9}\cdot f(x_1) + \frac{5}{9}\cdot f(x_2) = \frac{5}{9}\cdot f\left(-\frac{3}{\sqrt{15}}\right) + \frac{8}{9}\cdot f(0) + \frac{5}{9}\cdot f\left(-\frac{3}{\sqrt{15}}\right)$$

Hosszabb, de elemi számolásokkal megmutatható, hogy G_2 pontos a legfeljebb 5-fokú polinomokon.

Tovább általánosítva, ha az $x_0, x_1, ...x_N$ alappontrendszernek egy speciális polinomrendszer (*Legendre-polinomok*) gyökeit választjuk, akkor erre épített G_N interpolációs kvadratúra pontos a legfeljebb $(2N+1)$-edfokú polinomokon, tehát jóval pontosabb, mint az ekvidisztáns alappontrendszerre épülő interpolációs kvadratúra. Ezek a *Gauss-* (vagy Gauss-Legendre-) kvadratúrák.

4.5.4 Kétváltozós, tartományon vett integrálok közelítése

Többváltozós integrálok közelítő kiszámítása az egyváltozósokénál sokkal nehezebb probléma, egyebek közt azért, mert itt az Ω integrálási tartományt is megfelelő pontossággal közelíteni kell egyszerűbb tartományokkal.

Kétváltozós esetben egy egyszerű, általánosan elterjedt, és meglehetősen hatékony megközelítés a tartományt háromszögekre (háromváltozós esetben tetraéderekre) bontani. A felbontástól megköveteljük, hogy ne lapolják át egymást, és hézagmentesen fedjék le az Ω tartományt. Legyen a háromszögek rendszere T_1, T_2, ..., T_N (melyek egyesítése közel Ω-val egyezik), akkor az $\int_\Omega f(x,y)\,dxdy$ integrált így közelítjük:

$$\int_\Omega f(x,y)\,dxdy \approx \sum_{j=1}^{N} \int_{T_j} f(x,y)\,dxdy$$

A problémát így visszavezettük (egy általános) *háromszögön* vett integrál közelítő kiszámítására.

Háromszögön való közelítő integrálásra itt két egyszerű eljárást mutatunk. Jelölje P_1, P_2, P_3 a T háromszög csúcspontjait, S a súlypontját ($S = \frac{P_1+P_2+P_3}{3}$), $|T|$ pedig a területét.

A középpont szabály (érintőformula) általánosítása:

$$\int_T f(x,y)\,dxdy \approx f(S)\cdot|T| = f\left(\frac{P_1+P_2+P_3}{3}\right)\cdot|T|$$

A trapézformula általánosítása:

$$\int_T f(x,y)\,dxdy \approx \frac{f(P_1)+f(P_2)+f(P_3)}{3}\cdot|T|$$

Mindkét formula pontos a legfeljebb elsőfokú kétváltozós polinomokon. Mindkét formula összetett (tehát a teljes Ω tartomány háromszögfelbontására kiterjesztett) változatának hibája $\mathcal{O}(h^2)$, ahol most h jelöli az Ω-t lefedő háromszögrendszerben előforduló legnagyobb háromszögoldal hosszát.

A témakör messzire vezet, és sokkal nehezebb az egyváltozós kvadratúráknál. Érdekességképp megemlítjük, hogy lehetséges kétváltozós, tartományon vett integrálást visszavezetni egy, az eredeti tartomány *peremén* értelmezett másik integrál (vonalintegrál) kiszámítására, ami jóval egyszerűbb probléma.

4.6 Feladatok

1. Legfeljebb hányadfokú polinomokra pontos a $[0,1]$ intervallumon az

$$I(f) := \frac{f(0) + f(\frac{1}{2}) + f(1)}{3}$$

kvadratúra?

Megoldás: Elég a pontosságot a $p_0(x) := 1$, $p_1(x) := x$, $p_2(x) := x^2$, ... alappolinomokon ellenőrizni:

$$I(p_0) = \frac{1+1+1}{3} = 1 = \int_0^1 1\,dx,$$

$$I(p_1) = \frac{0+\frac{1}{2}+1}{3} = \frac{1}{2} = \int_0^1 x\,dx,$$

de

$$I(p_2) = \frac{0+\frac{1}{4}+1}{3} = \frac{5}{12} \neq \int_0^1 x^2\,dx = \frac{1}{3}$$

A kvadratúra tehát a legfeljebb elsőfokú polinomokra pontos.

2. Legfeljebb hányadfokú polinomokra pontos a $[0,1]$ intervallumon az

$$I(f) := \frac{f(0) + 2f(\frac{1}{2}) + f(1)}{4}$$

kvadratúra?

Megoldás: Elég a pontosságot a $p_0(x) := 1$, $p_1(x) := x$, $p_2(x) := x^2$, ... alappolinomokon ellenőrizni:

$$I(p_0) = \frac{1+2\cdot 1+1}{4} = 1 = \int_0^1 1\,dx,$$

$$I(p_1) = \frac{0+2\cdot\frac{1}{2}+1}{4} = \frac{1}{2} = \int_0^1 x\,dx,$$

de

$$I(p_2) = \frac{0+2\cdot\frac{1}{4}+1}{4} = \frac{3}{8} \neq \int_0^1 x^2\,dx = \frac{1}{3}$$

A kvadratúra tehát a legfeljebb elsőfokú polinomokra pontos.

3. Hogyan válasszuk meg a $c \geq 0$ számot úgy, hogy az alábbi kvadratúra a lehető legmagasabb fokszámú polinomokra pontos legyen a $[0,1]$ intervallumon? Optimális esetben hányadfokú polinomokra pontos a kvadratúra?

$$I(f) := \frac{f(0) + cf(\frac{1}{2}) + f(1)}{c+2}$$

Megoldás: Elég a pontosságot a $p_0(x) := 1$, $p_1(x) := x$, $p_2(x) := x^2$, ... alappolinomokon ellenőrizni:

$$I(p_0) = \frac{1+c+1}{c+2} = 1 = \int_0^1 1\,dx,$$

$$I(p_1) = \frac{0 + c \cdot \frac{1}{2} + 1}{c+2} = \frac{1}{2} = \int_0^1 x\,dx,$$

tehát a kvadratúra a legfeljebb elsőfokú polinomokra minden c esetén pontos. Továbbá:

$$I(p_2) = \frac{0 + c \cdot \frac{1}{4} + 1}{c+2} = \frac{1}{4} \cdot \frac{c+4}{c+2} \neq \int_0^1 x^2\,dx = \frac{1}{3}$$

Egyenlőség itt akkor teljesül, ha $c = 4$ (Simpson-formula). Ez tehát az optimális választás. Ellenőrizzük, hogy ekkor a még magasabb fokszámú polinomokon pontos-e:

$$I(p_3) = \frac{0 + 4 \cdot \frac{1}{8} + 1}{6} = \frac{1}{4} = \int_0^1 x^3\,dx$$

Tehát a kvadratúra még a harmadfokú polinomokon is pontos. Ámde:

$$I(p_4) = \frac{0 + 4 \cdot \frac{1}{16} + 1}{6} = \frac{5}{24} \neq \int_0^1 x^4\,dx = \frac{1}{5},$$

azaz itt már a kvadratúra nem pontos. Tehát az optimális választás: $c := 4$, és ekkor a kvadratúra pontos minden, legfeljebb harmadfokú polinomra.

4. Interpolációs típusú-e az alább kvadratúra a $[0,1]$ intervallumon?

$$I(f) := \frac{f(0) + 2f(\frac{1}{2}) + f(1)}{4}$$

Megoldás: Három alappont esetén a kérdés azzal ekvivalens, hogy a kvadratúra pontos-e a legfeljebb másodfokú polinomokra. Elég a pontosságot a $p_0(x) := 1$, $p_1(x) := x$, $p_2(x) := x^2$ alappolinomokon ellenőrizni:

$$I(p_0) = \frac{1+2+1}{4} = 1 = \int_0^1 1\,dx,$$

$$I(p_1) = \frac{0+2\cdot\frac{1}{2}+1}{4} = \frac{1}{2} = \int_0^1 x\,dx,$$

de

$$I(p_2) = \frac{0+2\cdot\frac{1}{4}+1}{4} = \frac{3}{8} \neq \int_0^1 x^2\,dx = \frac{1}{3}$$

A kvadratúra tehát nem interpolációs kvadratúra.

5. Interpolációs típusú-e az alább kvadratúra a $[0,1]$ intervallumon?

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

Megoldás: Két alappont esetén a kérdés azzal ekvivalens, hogy a kvadratúra pontos-e a legfeljebb elsőfokú polinomokra. Elég a pontosságot a $p_0(x) := 1$, $p_1(x) := x$ alappolinomokon ellenőrizni:

$$I(p_0) = \frac{1+1}{2} = 1 = \int_0^1 1\,dx,$$

$$I(p_1) = \frac{\frac{1}{3}+\frac{2}{3}}{2} = \frac{1}{2} = \int_0^1 x\,dx$$

A kvadratúra tehát interpolációs kvadratúra.

6. Legalább hány ekvidisztáns részre kell osztani az $[1,3]$ intervallumot ahhoz, hogy az összetett trapézformula legfeljebb 10^{-4} hibával közelítse az alábbi integrált?

$$\int_1^3 \frac{1}{x}\,dx \qquad (= \log 3)$$

Megoldás: Az összetett trapézformula hibabecslését használjuk. Jelölje

$f(x) := \frac{1}{x}$, akkor $f''(x) = \frac{2}{x^3}$, így ennek abszolút maximuma az $[1,3]$ intervallumon épp 2. Jelölje továbbá $a := 1$, $b := 3$, $h := \frac{b-a}{N}$. Ezt felhasználva, a hibabecslés:

$$\left| \int_a^b f(x)\,dx - \sum_{j=0}^{N-1} \frac{f(x_j) + f(x_{j+1})}{2} \cdot h \right| \leq$$

$$\leq \frac{1}{12} \max_{a \leq x \leq b} |f''(x)| \cdot h^2 \cdot (b-a) = \frac{1}{12} \cdot 2 \cdot \frac{4}{N^2} \cdot 2 =$$

$$= \frac{4}{3N^2}$$

A jobb oldal pedig akkor $\leq 10^{-4}$, ha $N^2 \geq \frac{4}{3} \cdot 10^4$, azaz, ha $N \geq 115.47$. Tehát 116 vagy ennél több részre osztva az $[1,3]$ intervallumot, az összetett trapézformula hibája biztosan 10^{-4} alatt marad.

7. Hogyan válasszuk meg a $[0,1]$ intervallumban a $0 < a < b < 1$ integrációs alappontokat úgy, hogy az alábbi kvadratúra a lehető legmagasabb fokszámú polinomokra pontos legyen a $[0,1]$ intervallumon? Optimális esetben hányadfokú polinomokra pontos a kvadratúra?

$$I(f) := \frac{f(a) + f(b)}{2}$$

Megoldás: Itt most a kvadratúra súlyai adottak, és a probléma az integrációs alappontok megválasztása. A pontosságot elég a $p_0(x) := 1$, $p_1(x) := x$, $p_2(x) := x^2$, ... alappolinomokon ellenőrizni:

$$I(p_0) = \frac{1+1}{2} = 1 = \int_0^1 1\,dx,$$

ez tehát mindig teljesül Továbbá:

$$I(p_1) = \frac{a+b}{2} = \int_0^1 x\,dx = \frac{1}{2},$$

tehát elsőfokú polinomokon akkor pontos a kvadratúra, ha $a + b = 1$, azaz az integrációs alappontok az intervallum középpontjára szimmetrikusan helyezkednek el. Nézzük a másodfokú alappolinom esetét

$$I(p_2) = \frac{a^2 + b^2}{2} = \int_0^1 x^2\,dx = \frac{1}{3}$$

Ez az egyenlőség akkor teljesül, ha $a^2 + b^2 = \frac{2}{3}$. Ebbe behelyettesítve az előzőleg kapott $a+b = 1$ egyenlőségből nyert $b = 1-a$ egyenlőséget:

$$a^2 + (1-a)^2 = \frac{2}{3}$$

Ennek megoldásai (ellenőrizzük!):

$$a = \frac{1 - \sqrt{1 - \frac{4}{6}}}{2} = \frac{1 - \sqrt{\frac{1}{3}}}{2}$$

$$a = \frac{1 + \sqrt{1 - \frac{4}{6}}}{2} = \frac{1 + \sqrt{\frac{1}{3}}}{2}$$

Ezek összege épp 1, tehát bármilyen választás esetén a másik a-érték épp a megfelelő b-értékkel egyezik. Mivel pedig a és b közül a-t jelöltük a kisebbnek, ezért a kapott eredmény úgy fogalmazható meg, hogy a kvadratúra akkor pontos a legfeljebb másodfokú polinomokon, ha

$$a = \frac{1}{2} - \frac{1}{2\sqrt{3}}, \qquad b = \frac{1}{2} + \frac{1}{2\sqrt{3}}$$

A dolog érdekessége, hogy ekkor a kvadratúra *automatikusan* pontos a harmadfokú polinomokon is:

$$I(p_3) = \frac{a^3 + b^3}{2} = \int_1^3 x^3\,dx = \frac{1}{4}$$

Az egyenlőség belátásához azt kell megvizsgálni, hogy $a^3 + b^3 = \frac{1}{2}$ teljesül-e. A bal oldal szorzattá bontható: $a^3 + b^3 = (a+b)(a^2 - ab + b^2)$. Mivel pedig $a + b = 1$, így elég azt belátni, hogy a, b fenti megválasztásával $a^2 - ab + b^2 = \frac{1}{2}$. Ez pedig egyszerű behelyettesítéssel már könnyen ellenőrizhető (tegyük ezt meg!).

Ugyancsak egyszerű számításokkal ellenőrizhető, hogy a kvadratúra a p_4 negyedfokú alappolinomra már nem pontos.

Tehát az integrációs alappontok optimális megválasztása: $a = \frac{1}{2} - \frac{1}{2\sqrt{3}}$, $b = \frac{1}{2} + \frac{1}{2\sqrt{3}}$, és ekkor a kvadratúra pontos minden, legfeljebb harmadfokú polinomra.

8. Legyen $h > 0$ tetszőleges lépésköz, és tekintsük a $[-\frac{h}{2}, \frac{h}{2}]$ intervallumon a

$$G_1(f) := \frac{f(-a) + f(a)}{2}$$

kvadratúrát, ahol $a := \frac{h}{2\sqrt{3}}$ (ld. az előző feladatot). Adjunk hibabecslést a kvadratúrára, feltéve, hogy f négyszer folytonosan differenciálható a $[-\frac{h}{2}, \frac{h}{2}]$ intervallumon.

Megoldás: Fejtsük f-et véges Taylor-sorba a 0 körül, negyedfokú maradéktaggal:

$$f(x) = f(0) + f'0)x + \frac{1}{2}f''(0)x^2 + \frac{1}{6}f'''(0)x^3 + \mathcal{O}(h^4)$$

Innen egyrészt (integrálva $-\frac{h}{2}$ és $\frac{h}{2}$ között):

$$\int_{-h/2}^{h/2} f(x)\,dx = f(0)h + \frac{0}{24}f''(1)h^3 + \mathcal{O}(h^5) \tag{6}$$

Másrészt, véve a helyettesítési értékeket az $x = a$ és $x = -a$ helyen:

$$f(a) = f(0) + f'0)a + \frac{1}{2}f''(0)a^2 + \frac{1}{6}f'''(0)a^3 + \mathcal{O}(h^4)$$

$$f(-a) = f(0) - f'0)a + \frac{1}{2}f''(0)a^2 - \frac{1}{6}f'''(0)a^3 + \mathcal{O}(h^4)$$

E két utóbbi egyenlőséget összeadva:

$$f(a)+f(-a) = 2f(0)+f''(0)a^2+\mathcal{O}(h^4) = 2f(0)+\frac{1}{12}f''(0)h^2+\mathcal{O}(h^4),$$

ahonnan nyilván:

$$\frac{f(a)+f(-a)}{2}\cdot h = f(0)h + \frac{1}{24}f''(0)h^3 + \mathcal{O}(h^5)$$

Ezt visszahelyettesítve a (6) egyenlőségbe, kapjuk a kvadratúra hibabecslését:

$$\left|\int_{-h/2}^{h/2} f(x)\,dx - \frac{f(a)+f(-a)}{2}\cdot h\right| = \mathcal{O}(h^5)$$

A kvadratúra tehát – h szerint – olyan pontos, mint a Simpson-formula, de 3 helyett csak 2 alappontot használ.

9. Legyen f egy négyszer folytonosan differenciálható függvény az $[a,b]$ intervallumon, és legyen $a = x_0 < x_1 < ... < x_N = b$ egy ekvidisztáns alappontrendszer: $x_k = a + k \cdot h$ $(k = 0, 1, ..., N)$, ahol $h := \frac{b-a}{N}$ a lépésköz. Tekintsük azt az összetett kvadratúraformulát melyben minden $[x_k, x_{k+1}]$ részintervallumon az előző feladatban szereplő kétpontos kvadratúraformulát alkalmazzuk:

$$\int_{x_k}^{x_{k+1}} f(x)\,dx \approx \frac{f(x_k^{(1)}) + f(x_k^{(2)})}{2} h,$$

ahol

$$x_k^{(1)} := x_k + \frac{h}{2} - \frac{h}{2\sqrt{3}} = \frac{\sqrt{3}-1}{2\sqrt{3}} h$$

$$x_k^{(2)} := x_k + \frac{h}{2} + \frac{h}{2\sqrt{3}} = \frac{\sqrt{3}+1}{2\sqrt{3}} h$$

h szerint hányadrendben pontos az így kapott összetett kvadratúraformula?

Megoldás: Az összetett kvadratúraformula (jelöljük G_N-nel):

$$G_N(f) = \sum_{k=0}^{N-1} \frac{f(x_k^{(1)}) + f(x_k^{(2)})}{2} h$$

Hibabecslése:

$$\left| \int_a^b f(x)\,dx - G_N(f) \right| = \left| \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x)\,dx - \frac{f(x_k^{(1)}) + f(x_k^{(2)})}{2} h \right| \leq$$

$$\leq \sum_{k=0}^{N-1} \left| \int_{x_k}^{x_{k+1}} f(x)\,dx - \frac{f(x_k^{(1)}) + f(x_k^{(2)})}{2} h \right|$$

A jobb oldalon használjuk az előző feladatban kapott ($\mathcal{O}(h^5)$-ös) hibabecsléseket. Eszerint tehát alkalmas $C > 0$ (f-től függő, de h-tól független) konstans mellett:

$$\left| \int_a^b f(x)\,dx - G_N(f) \right| \leq \sum_{k=0}^{N-1} C \cdot h^5 = C \cdot h^4 \sum_{k=0}^{N-1} h = C \cdot (b-a) \cdot h^4 =$$

$$= \mathcal{O}(h^4)$$

A kvadratúra tehát h szerint negyedrendben pontos, éppúgy, mint az összetett Simpson-formula.

5 Deriváltak közelítése

5.1 Motiváció

Noha a differenciálhányados definíciója viszonylag egyszerű (különbségi hányados határértéke), numerikus kiszámítása nem magától értetődő, különösen akkor, ha az illető függvény értékei hibákkal terheltek. Még fontosabb a differenciálhányados közelítése differenciálegyenletek numerikus megoldása során, ami igen sok fizikai-mérnöki probléma kapcsán természetes módon fellép. Nemcsak egyváltozós, hanem többváltozós függvények deriváltjainak (parciális deriváltjainak) kiszámításának is nagy a gyakorlati jelentősége, talán még nagyobb is, mint az egyváltozós függvényekéé: a parciális differenciálegyenletek numerikus megoldásának egy tekintélyes része erre épül.

Ebben a fejezetben egy- és többváltozós függvények közönséges és parciális deriváltjainak közelítésével foglalkozunk, beleértve a közelítések hibájának vizsgálatát is.

Az itt bemutatott technikák a Taylor-sorfejtésre épülnek. Ezért mindenekelőtt felidézzük az idevonatkozó, részben már ismert tételeket.

Taylor-sorfejtés (egyváltozós függvények):
Legyen $f : [x-\delta, x+\delta]$ egy $(n+1)$-szer folytonosan differenciálható függvény. Akkor minden $h \in [-\delta, \delta]$ esetén $f(x+h)$ előáll a következő alakban:

$$f(x+h) = f(x)+\frac{f'(x)}{1!}h+\frac{f''(x)}{2!}h^2+...+\frac{f^{(n)}(x)}{n!}h^n+\frac{f^{(n+1)}(\xi_x)}{(n+1)!}h^{n+1} \quad (7)$$

alkalmas, x-től függő $\xi_x \in [x-\delta, x+\delta]$ mellett.

Rögzített n mellett az utolsó tag (a *Lagrange-féle maradéktag*) $\mathcal{O}(|h|^{n+1})$ nagyságrendű.

Kétváltozós függvényekre a megfelelő formula bonyolultabb:

Taylor-sorfejtés (kétváltozós függvények):
Legyen $f : [x-\delta, x+\delta] \times [y-\delta, y+\delta]$ egy $(n+1)$-szer folytonosan differenciálható függvény. Akkor minden $h_x, h_y \in [-\delta, \delta]$ esetén $f(x+h_x, y+h_y)$ előáll a következő alakban:

$$f(x+h_x, y+h_y) = f(x,y) + \frac{1}{1!}\left(\frac{\partial f}{\partial x}h_x + \frac{\partial f}{\partial y}h_y\right) + \quad (8)$$

$$+\frac{1}{2!}\left(\frac{\partial^2 f}{\partial x^2}h_x^2 + 2\frac{\partial^2 f}{\partial x \partial y}h_x h_y + \frac{\partial^2 f}{\partial y^2}h_y^2\right) +$$

$$+\frac{1}{3!}\left(\frac{\partial^3 f}{\partial x^3}h_x^3+3\frac{\partial^3 f}{\partial x^2\partial y}h_x^2h_y+3\frac{\partial^3 f}{\partial x\partial y^2}h_xh_y^2+\frac{\partial^3 f}{\partial y^3}h_y^3\right)+...$$

$$+\frac{1}{n!}\cdot\sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right)\frac{\partial^n f}{\partial x^n\partial y^{n-k}}h_x^kh_y^{n-k}+\mathcal{O}(h^{n+1})$$

ahol $h := \max(|h_x|, |h_y|)$.

5.2 Közönséges differenciálhányadosok közelítése

Tegyük fel, hogy adott egy – az egyszerűség kedvéért ekvidisztáns – h lépésközű $x_0 < x_1 < ... < x_N$ alappontrendszer, ahol $x_k = x_0 + k \cdot h$ $(k = 0, 1, ..., N)$. Legyenek egy kellően sima f függvény estén adottak az $f_0 := f(x_0)$, ..., $f_N := f(x_N)$ függvényértékek.

5.2.1 Elsőrendű derivált közelítése

Probléma: az f' derivált közelítése az x_k alappontokban.

A legegyszerűbb, az $f'(x_k)$ deriváltat közelítő formulák a közönséges különbségi hányadosok.

Előrelépő séma:

$$f'(x_k) \approx \frac{f_{k+1} - f_k}{h}$$

Visszalépő séma:

$$f'(x_k) \approx \frac{f_k - f_{k-1}}{h}$$

Állítás: Mind az előre- mind a visszalépő séma *elsőrendű*, azaz a pontos $f'(x_k)$ deriváltat $\mathcal{O}(h)$ pontossággal közelítik.

Bizonyítás: Előrelépő séma: az (7) egyenlőséget alkalmazzuk másodrendű maradéktaggal. Alkalmas $\xi_x \in [x_k, x_{k+1}]$ mellett:

$$f_{k+1} = f(x_{k+1}) = f_k + f'(x_k) \cdot h + \frac{f''(\xi_x)}{2!}h^2,$$

ahonnan:

$$\left|\frac{f_{k+1} - f_k}{h} - f'(x_k)\right| = \frac{|f''(\xi_x)|}{2} \cdot h \leq \frac{\max|f''|}{2} \cdot h$$

A visszalépő séma esetén hasonlóan: alkalmas $\eta_x \in [x_{k-1}, x_k]$ mellett:

$$f_{k-1} = f(x_{k-1}) = f_k - f'(x_k) \cdot h + \frac{f''(\eta_x)}{2!} h^2,$$

ahonnan:

$$\left| \frac{f_k - f_{k-1}}{h} - f'(x_k) \right| = \frac{|f''(\eta_x)|}{2} \cdot h \leq \frac{\max |f''|}{2} \cdot h$$

Egy pontosabb séma a következő:

Centrális séma:

$$f'(x_k) \approx \frac{f_{k+1} - f_{k-1}}{2h}$$

Állítás: A centrális séma *másodrendű*, azaz a pontos $f'(x_k)$ deriváltat $\mathcal{O}(h^2)$ pontossággal közelíti.

Bizonyítás: Írjuk le újra az $f(x_{k-1})$, $f(x_{k+1})$ értékeket előállító Taylor-sorfejtéseket, de most harmadrendű maradéktaggal:

$$f_{k+1} = f(x_{k+1}) = f_k + f'(x_k) \cdot h + \frac{f''(x_k)}{2!} h^2 + \frac{f'''(\xi_x)}{3!} h^3$$

$$f_{k-1} = f(x_{k-1}) = f_k - f'(x_k) \cdot h + \frac{f''(x_k)}{2!} h^2 - \frac{f'''(\eta_x)}{3!} h^3$$

(ahol $\xi_x, \eta_x \in [x_{k-1}, x_{k+1}]$). A két egyenletet kivonva egymásból:

$$f_{k+1} - f_{k-1} = 2f'(x_k)h + \frac{f'''(\xi_x) + f'''(\eta_x)}{6} h^3,$$

ahonnan

$$\left| \frac{f_{k+1} - f_{k-1}}{2h} - f'(x_k) \right| \leq \frac{\max |f'''|}{3} \cdot h^2$$

Több alappontot bevonva, (elvileg) tetszőleges pontosságú sémák konstruálhatók (feltéve, hogy az f függvény elég sima, azaz elég sokszor folytonosan differenciálható).

Példaképp, tekintsük az x_{k-2}, x_{k-1}, x_k, x_{k+1}, x_{k+2} ekvidisztáns alappontok melletti Taylor-sorfejtéseket ötödrendű maradéktagokkal:

$$f_{k+1} = f_k + f'(x_k) \cdot h + \frac{f''(x_k)}{2!} h^2 + \frac{f'''(x_k)}{3!} h^3 + \frac{f^{IV}(x_k)}{4!} h^4 + \mathcal{O}(h^5)$$

$$f_{k-1} = f_k - f'(x_k) \cdot h + \frac{f''(x_k)}{2!}h^2 - \frac{f'''(x_k)}{3!}h^3 + \frac{f^{IV}(x_k)}{4!}h^4 + \mathcal{O}(h^5)$$

Kivonva:

$$f_{k+1} - f_{k-1} = 2f'(x_k)h + \frac{1}{3}f'''(x_k)h^3 + \mathcal{O}(h^5)$$

Hasonlóan:

$$f_{k+2} = f_k + f'(x_k) \cdot 2h + \frac{f''(x_k)}{2!}4h^2 + \frac{f'''(x_k)}{3!}8h^3 + \frac{f^{IV}(x_k)}{4!}16h^4 + \mathcal{O}(h^5)$$

$$f_{k-2} = f_k - f'(x_k) \cdot 2h + \frac{f''(x_k)}{2!}4h^2 - \frac{f'''(x_k)}{3!}8h^3 + \frac{f^{IV}(x_k)}{4!}16h^4 + \mathcal{O}(h^5)$$

Kivonva:

$$f_{k+2} - f_{k-2} = 4f'(x_k)h + \frac{8}{3}f'''(x_k)h^3 + \mathcal{O}(h^5)$$

Az $(f_{k+1} - f_{k-1})$-re kapott egyenlőség 8-szorosából ez utóbbi egyenlőséget kivonva, a harmadrendű deriváltat tartalmazó tagok is kiesnek, és a következőt kapjuk:

$$-f_{k+2} + 8f_{k+1} - 8f_{k-1} + f_{k-2} = 12f'(x_k) + \mathcal{O}(h^5),$$

ahonnan $f'(x_k)$ közelítésére egy *negyedrendű* sémát nyerünk:

$$f'(x_k) = \frac{f_{k-2} - 8f_{k-1} + 8f_{k+1} - f_{k+2}}{12} + \mathcal{O}(h^4)$$

5.2.2 Másodrendű derivált közelítése

A második derivált konzisztens közelítéséhez nyilván legalább 3 alappont szükséges. A legegyszerűbb, mégis jó pontosságú a *hárompontos centrális séma*. Ennek megkonstruálásához tekintsük az x_k alappont körüli következő Taylor-sorfejtéseket:

$$f_{k+1} = f_k + f'(x_k) \cdot h + \frac{f''(x_k)}{2!}h^2 + \frac{f'''(x_k)}{3!}h^3 + \mathcal{O}(h^4)$$

$$f_{k-1} = f_k - f'(x_k) \cdot h + \frac{f''(x_k)}{2!}h^2 - \frac{f'''(x_k)}{3!}h^3 + \mathcal{O}(h^4)$$

Összeadva a két egyenlőséget, az első- és a harmadrendű deriváltakat tartalmazó tagok kiesnek, és azt kapjuk, hogy:

$$f_{k+1} + f_{k-1} = 2f_k + f''(x_k)h^2 + \mathcal{O}(h^4)$$

Innen nyerjük a szóbanforgó sémát:

$$f''(x_k) \approx \frac{f_{k-1} - 2f_k + f_{k+1}}{h^2}$$

A séma hibája:

$$\left| f''(x_k) - \frac{f_{k-1} - 2f_k + f_{k+1}}{h^2} \right| = \mathcal{O}(h^2)$$

Bevonva még az x_{k-2}, x_{k+2} alappontokat, hasonló eljárással kapjuk a még pontosabb *ötpontos centrális sémát*. Az x_k körüli Taylor-sorfejtéseket a hatodrendű maradéktagig elvégezve:

$$f_{k+1} = f_k + f'(x_k)\cdot h + \frac{f''(x_k)}{2!}h^2 + \frac{f'''(x_k)}{3!}h^3 + \frac{f^{IV}(x_k)}{4!}h^4 + \frac{f^{V}(x_k)}{5!}h^5 + \mathcal{O}(h^6)$$

$$f_{k-1} = f_k - f'(x_k)\cdot h + \frac{f''(x_k)}{2!}h^2 - \frac{f'''(x_k)}{3!}h^3 + \frac{f^{IV}(x_k)}{4!}h^4 - \frac{f^{V}(x_k)}{5!}h^5 + \mathcal{O}(h^6)$$

Összeadva:

$$f_{k+1} + f_{k-1} = 2f_k + f''(x_k)h^2 + \frac{f^{IV}(x_k)}{12}h^4 + \mathcal{O}(h^6)$$

Hasonlóan:

$$f_{k+2} = f_k + f'(x_k)\cdot 2h + \frac{f''(x_k)}{2!}4h^2 + \frac{f'''(x_k)}{3!}8h^3 + \frac{f^{IV}(x_k)}{4!}16h^4 + \frac{f^{V}(x_k)}{5!}32h^5 + \\ + \mathcal{O}(h^6)$$

$$f_{k+2} = f_k - f'(x_k)\cdot 2h + \frac{f''(x_k)}{2!}4h^2 - \frac{f'''(x_k)}{3!}8h^3 + \frac{f^{IV}(x_k)}{4!}16h^4 - \frac{f^{V}(x_k)}{5!}32h^5 + \\ + \mathcal{O}(h^6)$$

Összeadva:

$$f_{k+2} + f_{k-2} = 2f_k + 4f''(x_k)h^2 + \frac{16}{12}f^{IV}(x_k)h^4 + \mathcal{O}(h^6)$$

Az $f_{k+1} + f_{k-1}$ összegre kapott egyenlőség 16-szorosából ez utóbbit kivonva:

$$16f_{k+1} + 16f_{k-1} - f_{k+2} - f_{k-2} = 30f_k + 12f''(x_k)h^2 + \mathcal{O}(h^6)$$

A séma tehát:

$$f''(x_k) \approx \frac{-f_{k-2} + 16f_{k-1} - 30f_k + 16f_{k+1} - f_{k+2}}{12h^2}$$

A séma hibája pedig:

$$\left| f''(x_k) - \frac{-f_{k-2} + 16f_{k-1} - 30f_k + 16f_{k+1} - f_{k+2}}{12h^2} \right| = \mathcal{O}(h^4)$$

5.2.3 Numerikus differenciálás nem feltétlen ekvidisztáns alappontrendszeren

Ha az alappontrendszer nem ekvidisztáns, akkor az előzőekben leírt sémakonstrukciók elbonyolódnak, és a pontosság is csökkenhet. E nehézség áthidalására egyszerű módszert kínál a *spline interpoláció* (ld. 3. fejezet) alkalmazása.

Legyen tehát $a = x_0 < x_1 < ... < x_N = b$ egy alappontrendszer, az alappontokhoz tartozzanak az f_0, f_1, ..., f_N függvényértékek. A spline interpolációnál leírtak szeint (3. fejezet) mindenekelőtt az f'_0, f'_1, ..., f'_N deriváltértékeket állítunk elő, a következő egyenletrendszer megoldásával:

$$\frac{1}{h_{k-1}} f'_{k-1} + \left(\frac{2}{h_{k-1}} + \frac{2}{h_k} \right) f'_k + \frac{1}{h_k} f'_{k+1} =$$

$$= -\frac{3}{h_{k-1}^2} f_{k-1} + \left(\frac{3}{h_{k-1}^2} - \frac{3}{h_k^2} \right) f_k + \frac{3}{h_k^2} f_{k+1}$$

$(k = 1, 2, ..., N-1)$. Az f'_0, f'_N deriváltértékre külön peremfeltételek megkövetelésével írunk fel további egyenleteket.

Ha csak az f függvény első deriváltjának közelítéseit kell előállítani az alappontokban, akkor az S_f spline függvényt nem is kell explicite előállítani, hiszen a fenti egyenletrendszer megoldása épp a spline függvény alapponti deriváltértékeit szolgáltatja. Továbbá, ha az f'_0, f'_N értékek adottak, akkor az így kapott teljes spline függvény deriváltja $\mathcal{O}(h^3)$ hibával közelíti az f függvény deriváltját, azaz ez a módszer pontosabb a centrális sémánál, és még csak az alappontok ekvidisztáns jellege sem szükséges ehhez. Ugyanez áll a természetes peremfeltétel esetében is, amennyiben f'' eltűnik a-ban és b-ben.

Ha f második vagy harmadik deriváltjait kell közelíteni, akkor már fel kell írni a spline függvényt az $[x_k, x_{k+1}]$ részintervallumokon. Emlékeztetünk rá, hogy ennek alakja a következő:

$$S_f(x) = H_k(x) = A + B \cdot \frac{x - x_k}{h_k} + C \cdot \frac{(x - x_k)^2}{h_k^2} + D \cdot \frac{(x - x_k)^3}{h_k^3}$$

ahol

$$\begin{aligned} A &= f_k \\ B &= h f'_k \\ C &= -3f_k + 3f_{k+1} - 2hf'_k - hf'_{k+1} \end{aligned}$$

$$D = \quad 2f_k - 2f_{k+1} + \quad hf'_k + hf'_{k+1}$$

Innen:

$$S''_f(x_k) = H''_k(x_k) = \frac{2C}{h_k^2}, \qquad S'''_f(x_k) = H'''_k(x_k) \equiv \frac{6D}{h_k^3}$$

amelyek az alapponti adatok (f_k) és a számított alapponti deriváltértékek (f'_k) ismeretében már nehézség nélkül számíthatók.

5.3 Alkalmazás közönséges differenciálegyenletek megoldására

Igen sok fizikai folyamat írható le közönséges differenciálegyenletek segítségével. Ezek egy tág osztálya a következő alakú: adott egy *kétváltozós* $f : \mathbf{R}^2 \to \mathbf{R}$ függvény, és keressünk olyan $y : [x_0, +\infty) \to \mathbf{R}$ függvényt, melyre

$$y'(x) = f(x, y(x))$$

teljesül minden $x > x_0$ mellett.

Ismeretes, hogy ennek a differenciálegyenletnek általában (végtelen) sok megoldása van, de az y függvény értékének előírása az x_0 helyen (*kezdeti feltétel*) már egyértelművé teszi a megoldást.

Nem célunk itt a közönséges differenciálegyenletek numerikus módszereinek témakörét részletesen tanulmányozni. Mindössze arra hozunk néhány egyszerű példát, hogy hogyan alkalmazható a közelítő differenciálás technikája ezen, a gyakorlat számára igen fontos területen.

Az

$$y'(x) = f(x, y(x)), \qquad y(x_0) = y_0 \tag{9}$$

kezdeti érték feladatot numerikusan egy véges, $[x_0, x_0 + L]$ intervallumon közelítjük. Vegyünk fel itt egy, az egyszerűség kedvéért ekvidisztáns, h lépésközű *számítási rácsot*:

$$x_k := x_0 + k \cdot h \qquad (k = 0, ..., N),$$

ahol $h = \frac{L}{N}$.

A véges differenciák módszerének alapötlete, hogy a (9) differenciálegyenletben előforduló $y'(x)$ deriváltakat egyszerű differenciasémákkal helyettesítjük. Csak a rácspontokban felvett (közelítő) y-értékeket számítjuk: két rácspont között y közelítése egyészen más (interpolációs) probléma. Ezzel itt nem foglalkozunk.

A két legegyszerűbb módszer a két legegyszerűbb differenciasémára épül. Legyen x_k egy tetszőleges rácspont ($k = 1, 2, ..., N-1$):

Előrelépő séma: Itt az

$$y'(x_k) \approx \frac{y(x_{k+1}) - y(x_k)}{h}$$

közelítést alkalmazzuk. Jelölje y_k az $y(x_k)$ függvényérték közelítését, akkor tehát az alábbi módszert nyerjük:

$$\frac{y_{k+1} - y_k}{h} = f(x_k, y_k)$$

azaz

$$y_{k+1} = y_k + h \cdot f(x_k, y_k) \qquad (k = 0, 1, ..., N-1) \tag{10}$$

Ez az *explicit Euler-módszer*: a módszert egy explicit rekurzió definiálja, a következő rácspontbeli y_{k+1} közelítő megoldásérték direkt számolással, egyenletmegoldás nélkül kapható.

Visszalépő séma: Itt az $\frac{y(x_{k+1})-y(x_k)}{h}$ különbségi hányadossal az x_{k+1}-beli $y'(x_{k+1})$ deriváltértéket közelítjük. Innen a megfelelő módszer:

$$\frac{y_{k+1} - y_k}{h} = f(x_{k+1}, y_{k+1})$$

azaz

$$y_{k+1} = y_k + h \cdot f(x_{k+1}, y_{k+1}) \qquad (k = 0, 1, ..., N-1) \tag{11}$$

Ez az *implicit Euler-módszer*. Itt minden egyes új y_{k+1} érték kiszámításához egy egyenletet kell megoldani, mivel y_{k+1} a jobb oldalon is előfordul. Ennek technikájával nem foglalkozunk, de megjegyezzük, hogy ha h elég kicsi, akkor ez az egyenlet igen általános feltételek mellett megoldható, tehát a feladat numerikus szempontból nem lényegesen nehezebb, mint az explicit Euler-módszer realizálása.

Megjegyzés: Felmerül a kérdés, miért foglalkozunk egyáltalán az implicit módszerekkel, mikor az explicit módszer sokkal egyszerűbb. Kiderült, hogy egy sor, a gyakorlat számára is fontos esetben ahhoz, hogy elég jól közelítő megoldást kapjunk, irreálisan kis h lépésközt kell alkalmazni, ami numerikus szempontból nagyon "megdrágítja" az explicit módszert az implicithez képest

– még úgy is, hogy az implicit módszer minden egyes lépésében egy (általában nemlineáris) egyenletet kell megoldani. Egy másik szempont: sok esetben fizikai meggondolásokból előre tudjuk, hogy a megoldás értékei csakis pozitív számok lehetnek, mert pl. hőmérsékletet, koncentrációt, energiát stb. jelentenek. Jogos elvárás, hogy a numerikus módszer adta közelítő megoldás is ilyen tulajdonságú legyen (pozitivitástartás). Ilyen kvalitatív tulajdonságokat az explicit módszerek (szemben az implicitekkel) nem, vagy nem mindig tartanak meg.

Igazolható (a részletekbe itt nem megyünk bele), hogy mind az explicit, mind az implicit Euler-módszer pontossága *elsőrendű*, azaz az $y(x_k) - y_k$ *globális hibára* fennáll az

$$|y(x_k) - y_k| = \mathcal{O}(h)$$

becslés: kicsit pontatlanul azt mondhatjuk, hogy a hiba a h lépésköz *első* hatványával arányos.

A pontosabb *centrális séma* a következőképp alkalmazható. Az $\frac{y(x_{k+1})-y(x_k)}{h}$ különbségi hányadossal most az $[x_k, x_{k+1}]$ intervallum $x_{k+1/2}$ középpontjában érvényes deriváltértéket közelítünk. Innen első lépésben a következő eljárást kapjuk:

$$\frac{y_{k+1} - y_k}{h} = f(x_{k+1/2}, y_{k+1/2})$$

A jobb oldalon további közelítések szükségesek, mert nem rácsponti értékek szerepelnek. Alkalmazva az

$$f(x_{k+1/2}, y_{k+1/2}) \approx \frac{f(x_k, y_k) + f(x_{k+1}, y_{k+1})}{2}$$

közelítést, egy újabb implicit módszert nyerünk (*Crank-Nicolson-séma*), mely azonban h szerint másodrendű közelítést ad (azaz hibája $\mathcal{O}(h^2)$), tehát pontosabb mind az explicit, mind az implicit Euler-módszernél:

$$\frac{y_{k+1} - y_k}{h} = \frac{f(x_k, y_k) + f(x_{k+1}, y_{k+1})}{2}$$

azaz

$$y_{k+1} = y_k + \frac{h}{2} \cdot (f(x_k, y_k) + f(x_{k+1}, y_{k+1})) \qquad (k = 0, 1, ..., N-1) \quad (12)$$

5.3.1 Kvadratúrával javított módszerek

Itt – igen röviden és részletes analízis nélkül – arra mutatunk két érdekes példát, hogyan lehet *differenciálegyenletek* numerikus megoldásában a *közelítő integrálási* technikákat (ld. 4. fejezet) felhasználni.

Tekintsük továbbra is az

$$y'(x) = f(x, y(x)), \qquad y(x_0) = y_0$$

kezdeti érték feladatot, és a differenciálegyenlet mindkét oldalát integráljuk az x_k és x_{k+1} rácspontok között. Kapjuk, hogy:

$$y(x_{k+1}) - y(x_k) = \int_{x_k}^{x_{k+1}} f(x, y(x))\, dx$$

A bal oldal közelítése nyilván az $y_{k+1} - y_k$ különbség. A jobb oldalon pedig egyszerű kvadratúraformulákat használhatunk fel.

A *trapézformulával* pl. a jobb oldali integrál így közelíthető:

$$\int_{x_k}^{x_{k+1}} f(x, y(x))\, dx = \frac{f(x_k, y(x_k)) + f(x_{k+1}, y(x_{k+1}))}{2} h + \mathcal{O}(h^3),$$

ahonnan visszakapjuk az

$$y_{k+1} - y_k = \frac{h}{2} \cdot (f(x_k, y_k) + f(x_{k+1}, y_{k+1}))$$

Crank-Nicolson-sémát. Sőt, explicit módszer konstruálására is van lehetőség. Felhasználva ui. az alábbi sorfejtéseket és becsléseket:

$$f(x_{k+1}, y(x_{k+1})) = f(x_{k+1}, y(x_k) + hy'(x_k) + \mathcal{O}(h^2)) =$$

$$= f(x_{k+1}, y(x_k) + hf(x_k, y(x_k) + \mathcal{O}(h^2))) =$$

$$= f(x_{k+1}, y(x_k) + hf(x_k, y(x_k))) + \mathcal{O}(h^2),$$

a következő sémát nyerjük:

$$y_{k+1} - y_k = \frac{h}{2} \cdot (f(x_k, y_k) + f(x_{k+1}, y_k + hf(x_k, y_k)))$$

Ez már explicit formula. Áttekinthetőbb az alábbi felírásban:

$$y^*_{k+1} := y_k + h \cdot f(x_k, y_k)$$

$$y_{k+1} := y_k + \frac{h}{2} \cdot (f(x_k, y_k) + f(x_{k+1}, y^*_{k+1}))$$

(*trapézformulával javított Euler-módszer*). Igazolható, hogy a módszer *másodrendű*, azaz érvényes a következő becslés:

$$|y(x_k) - y_k| = \mathcal{O}(h^2)$$

Egy másik lehetőség az *érintőformula* használata. Ekkor a jobb oldali integrál így közelíthető:

$$\int_{x_k}^{x_{k+1}} f(x, y(x))\, dx = f(x_{k+1/2}, y(x_{k+1/2})) \cdot h + \mathcal{O}(h^3)$$

A jobb oldalon $y(x_{k+1/2})$-et véges Taylor-sorral közelítve:

$$f(x_{k+1/2}, y(x_{k+1/2})) = f(x_{k+1/2}, y(x_k) + \frac{h}{2} y'(x_k) + \mathcal{O}(h^2)) =$$

$$= f(x_{k+1/2}, y(x_k) + \frac{h}{2} f(x_k, y(x_k)) + \mathcal{O}(h^2)) =$$

$$= f(x_{k+1/2}, y(x_k) + \frac{h}{2} f(x_k, y(x_k))) + \mathcal{O}(h^2) =$$

a következő sémát nyerjük:

$$y_{k+1} - y_k = h \cdot (f(x_{k+1/2}, y_k + \frac{h}{2} f(x_k, y_k))),$$

ami szintén explicit formula. Az alábbi felírásban áttekinthetőbb:

$$y_{k+1/2} := y_k + \frac{h}{2} \cdot f(x_k, y_k)$$

$$y_{k+1} := y_k + h \cdot f(x_{k+1/2}, y_{k+1/2})$$

(*érintőformulával javított Euler-módszer*). Megmutatható, hogy ez a módszer is másodrendű, azaz érvényes a következő becslés:

$$|y(x_k) - y_k| = \mathcal{O}(h^2)$$

5.4 Parciális deriváltak közelítése

Legyen $f : \mathbf{R}^2 \to \mathbf{R}$ egy kétváltozós függvény. A közönséges differenciálhányadosok közelítésére szolgáló eszköztárat természetes módon általánosítva, tekintsünk egy kétdimenziós rácspontrendszert:

$$x_0 < x_1 < ... < x_N, \qquad x_k := x_0 + kh \qquad (k = 0, 1, ..., N)$$

$$y_0 < y_1 < ... < y_M, \qquad y_j := y_0 + jh \qquad (j = 0, 1, ..., M)$$

Egyszerűség kedvéért mindkét változó szerint ekvidisztáns és azonos lépésközű rácsot használunk.

Probléma: f bizonyos parciális deriváltjainak közelítése a fenti rácspontokban.

A rövidség kedvéért a későbbiekben jelölje $f_{kj} := f(x_k, y_j)$.

Adott $e = (e_x, e_y)$ irányú derivált közelítése az (x_k, y_j) rácspontban. Ehhez szükséges a $\frac{\partial f}{\partial x}$ és a $\frac{\partial f}{\partial y}$ parciális deriváltak közelítése. Ezeket külön-külön, egy-egy centrális sémát alkalmazva közelíthetjük:

$$\frac{\partial f}{\partial e}(x_k, y_j) = \left(\frac{\partial f}{\partial x} \cdot e_x + \frac{\partial f}{\partial y} \cdot e_y\right)(x_k, y_j) \approx$$

$$\approx \frac{f_{k+1,j} - f_{k-1,j}}{2h} \cdot e_x + \frac{f_{k,j+1} - f_{k,j-1}}{2h} \cdot e_y$$

A közelítés pontossága $\mathcal{O}(h^2)$.

A gyakorlati alkalmazásokban kitüntetett szerepe van a $\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$ *Laplace-operátor* közelítésének. Mindkét változó szerint centrális sémát alkalmazva:

$$(\Delta f)(x_k, y_j) \approx \frac{f_{k+1,j} - 2f_{k,j} + f_{k-1,j}}{h^2} + \frac{f_{k,j+1} - 2f_{k,j} + f_{k,j-1}}{h^2}$$

A közelítés pontossága $\mathcal{O}(h^2)$.

Egy másik – olykor kényelmesebb – jelöléstechnika a kétindexes jelölések helyett az égtájak szerinti jelölések használata, a koordinátarendszert kelet-nyugati ill. észak-déli tájolásúnak tekintve. Így egy tetszőleges, centrálisnak (C) tekintett (k, j) indexű rácspont égtájak szerinti szomszédjai: NE (indexe $(k+1, j+1)$), N (indexe $(k, j+1)$), NW (indexe $(k-1, j+1)$), és így tovább. Ld. az ábrát.

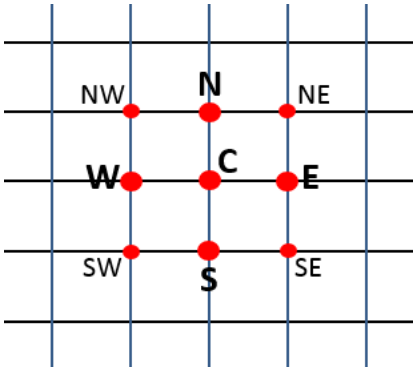

Figure 8: Számítási rács 2D-ben. Egy centrális pont és az égtájak szerinti szomszédjai

Ezzel a jelöléstechnikával a fenti, a Laplace-operátort közelítő 5-pontos centrális séma:

$$(\Delta f)_C \approx \frac{f_N + f_W + f_S + f_E - 4f_C}{h^2} \tag{13}$$

A Laplace-operátort az eddigi C, N, W, S, E alappontok helyett egy másik alappontkonfiguráción (C, NE, NW, SW, SE) is közelíthetjük. Ehhez viszont már a 2-változós Taylor-sorfejtés használata szükséges. Az alábbi táblázatban a 2-változós Taylor-sor egyes tagjainak együtthatóit mutatjuk,

a harmadrendű tagokkal bezárólag:

	f	f_x	f_y	f_{xx}	f_{xy}	f_{yy}	f_{xxx}	f_{xxy}	f_{xyy}	f_{yyy}
NE	1	h	h	$\frac{1}{2}h^2$	$\frac{2}{2}h^2$	$\frac{1}{2}h^2$	$\frac{1}{6}h^3$	$\frac{3}{6}h^3$	$\frac{3}{6}h^3$	$\frac{1}{6}h^3$
NW	1	$-h$	h	$\frac{1}{2}h^2$	$-\frac{2}{2}h^2$	$\frac{1}{2}h^2$	$-\frac{1}{6}h^3$	$\frac{3}{6}h^3$	$-\frac{3}{6}h^3$	$\frac{1}{6}h^3$
SW	1	$-h$	$-h$	$\frac{1}{2}h^2$	$\frac{2}{2}h^2$	$\frac{1}{2}h^2$	$-\frac{1}{6}h^3$	$-\frac{3}{6}h^3$	$-\frac{3}{6}h^3$	$-\frac{1}{6}h^3$
SE	1	h	$-h$	$\frac{1}{2}h^2$	$-\frac{2}{2}h^2$	$\frac{1}{2}h^2$	$\frac{1}{6}h^3$	$-\frac{3}{6}h^3$	$\frac{3}{6}h^3$	$-\frac{1}{6}h^3$

Az áttekinthetőség kedvéért a parciális deriválásokat indexekkel jelöltük. Mindegyik sornak megfelelő véges Taylor-sor hibája $\mathcal{O}(h^4)$, feltéve, hogy az f függvény *négyszer* folytonosan differenciálható.

A táblázat mind a négy sorát összeadva az első- és harmadrendű parciális deriváltakat valamint a másodrendű vegyes parciális deriváltat tartalmazó tagok kiesnek. A következőt kapjuk:

$$f_{NE} + f_{NW} + f_{SW} + f_{SE} = 4f_C + 2f_{xx}h^2 + 2f_{yy}h^2 + \mathcal{O}(h^4),$$

ahonnan a szintén $\mathcal{O}(h^2)$ pontosságú sémát kapjuk:

$$(\Delta f)_C = \frac{f_{NE} + f_{NW} + f_{SW} + f_{SE} - 4f_C}{2h^2} \tag{14}$$

Sőt, ha ugyanezen táblázat sorait *váltakozó előjellel* adjuk össze, akkor a másodrendű vegyes parciális deriváltra kapunk egy másodrendben pontos sémát:

$$f_{NE} - f_{NW} + f_{SW} - f_{SE} = 4f_{xy}h^2 + \mathcal{O}(h^4),$$

ahonnan:

$$\left(\frac{\partial^2 f}{\partial x \partial y}\right)_C = \frac{f_{NE} - f_{NW} + f_{SW} - f_{SE}}{4h^2} + \mathcal{O}(h^2)$$

A Laplace-operátort közelítő két 5-pontos sémának ((13) és (14)) egy kombinációja érdekes tulajdonságokkal rendelkezik. A 2-változós Taylor-sorfejtést a hatodrendű maradéktagig elvégezve hosszabb, de elemi számolások után a következő sémát nyerjük:

9-pontos centrális séma a Laplace-operátorra

$$(\Delta f)_C = \frac{f_{NW} + 4f_N + f_{NE} + 4f_E + f_{SE} + 4f_S + f_{SW} + 4f_W - 20f_c}{6h^2} -$$

$$-\frac{h^2}{12} \cdot (f_{xxxx} + 2f_{xxyy} + f_{yyyy}) + \mathcal{O}(h^4)$$

A jobb oldal első törtje egy *centrális 9-pontos séma*. A második tagja pedig könnyen ellenőrizhetően: $(-\frac{h^2}{12}(\Delta\Delta f)_C)$. A séma tehát általában $\mathcal{O}(h^2)$ pontosságú; ám, ha speciálisan $\Delta f \equiv 0$ (azaz a sémát a Laplace-egyenlet megoldásának közelítésére használjuk), akkor ez a tag eltűnik, és a 9-pontos séma pontossága $\mathcal{O}(h^4)$-re javul.

5.5 Feladatok

1. Legyen $f : \mathbf{R} \to \mathbf{R}$ háromszor folytonosan differenciálható. Legyenek $x_0, x_1, x_2 \in \mathbf{R}$, $h := x_1 - x_0 = x_2 - x_1$, és jelölje $f_0 := f(x_0)$, $f_1 := f(x_1)$, $f_2 := f(x_2)$. Konstruáljunk egy, az $f'(x_0)$ deriváltat h szerint másodrendben közelítő differenciasémát az f_0, f_1, f_2 függvényértékekből.

   Megoldás: Fejtsük f-et x_0 körül véges Taylor-sorba:

$$f_1 = f_0 + \frac{f'(x_0)}{1!} \cdot h + \frac{f''(x_0)}{2!} \cdot h^2 + \frac{f'''(\xi)}{3!} \cdot h^3$$

$$f_2 = f_0 + \frac{f'(x_0)}{1!} \cdot 2h + \frac{f''(x_0)}{2!} \cdot 4h^2 + \frac{f'''(\eta)}{3!} \cdot 8h^3$$

   Az első egyenlet 4-szereséből kivonva a másodikat, a másodrendű deriváltat tartalmazó tagok kiesnek, és azt kapjuk, hogy

$$4f_1 - f_2 = 3f_0 + \frac{f'(x_0)}{1!} \cdot 2h + \frac{4f'''(\xi) - 8f'''(\eta)}{3!} \cdot h^3$$

   Innen:

$$\left| \frac{-3f_0 + 4f_1 - f_2}{2h} - f'(x_0) \right| \leq ||f'''||_{\max} \cdot h^2$$

   A séma tehát:

$$f'(x_0) \approx \frac{-3f_0 + 4f_1 - f_2}{2h}$$

   és ez az $f'(x_0)$ deriváltat h szerint valóban másodrendben közelíti.

2. Legyen $f : \mathbf{R} \to \mathbf{R}$ elegendően sokszor folytonosan differenciálható. Legyenek $x_0 \in \mathbf{R}$, h adott, és $x_{-1} := x_0 - h$, $x_1 := x_0 + 2h$. Jelölje $f_0 := f(x_0)$, $f_{-1} := f(x_{-1})$, $f_1 := f(x_1)$. Konstruáljunk az $f''(x_0)$ deriváltat közelítő sémát az f_{-1}, f_0, f_1 függvényértékekből. Hányadrendű közelítés érhető el (h szerint)?

   Megoldás: Fejtsük f-et x_0 körül véges Taylor-sorba:

$$f_1 = f_0 + \frac{f'(x_0)}{1!} \cdot 2h + \frac{f''(x_0)}{2!} \cdot 4h^2 + \frac{f'''(\xi)}{3!} \cdot 8h^3$$

$$f_{-1} = f_0 - \frac{f'(x_0)}{1!} \cdot h + \frac{f''(x_0)}{2!} \cdot h^2 - \frac{f'''(\eta)}{3!} \cdot h^3$$

A első egyenlethez a második kétszeresét hozzáadva, az $f'(x_0)$-t tartalmazó tagok kiesnek, innen:

$$f_1 + 2f_{-1} = 3f_0 + \frac{f''(x_0)}{2!} \cdot 6h^2 + \frac{f'''(\xi)}{3!} \cdot 8h^3 - \frac{f'''(\eta)}{3!} \cdot 2h^3$$

Innen

$$\left| \frac{2f_{-1} - 3f_0 + f_1}{3h^2} - f''(x_0) \right| \leq \frac{10h^3 \cdot ||f'''||_{\max}}{3! \cdot 3h^2} = \frac{5 \cdot ||f'''||_{\max}}{9} \cdot h$$

A séma tehát:

$$\frac{2f_{-1} - 3f_0 + f_1}{3h^2},$$

és ez az $f''(x_0)$ deriváltat h szerint elsőrendben közelíti. (A közelítés rendje nem javul, ha a Taylor sorfejtést több tagra végezzük el – próbáljuk ki!)

3. Legyen $f : \mathbf{R} \to \mathbf{R}$ elegendően sokszor folytonosan differenciálható. Legyenek $x_0 \in \mathbf{R}$, h adott, és $x_k := x_0 + k \cdot h$, $k = -2, -1, 1, 2$-re. Jelölje $f_k := f(x_k)$ $(k = -2, -1, 0, 1, 2)$. Konstruáljunk az $f'''(x_0)$ deriváltat közelítő sémát az f_k ($k = -2$, -1, 0, 1, 2), függvényértékekből. Hányadrendű közelítés érhető el (h szerint)?

Megoldás: Fejtsük f-et x_0 körül véges Taylor-sorba:

$$f_1 = f_0 + \frac{f'(x_0)}{1!} \cdot h + \frac{f''(x_0)}{2!} \cdot h^2 + \frac{f'''(x_0)}{3!} \cdot h^3 + \frac{f^{IV}(x_0)}{4!} \cdot h^4 + \frac{f^V(\xi_1)}{5!} \cdot h^5$$

$$f_{-1} = f_0 - \frac{f'(x_0)}{1!} \cdot h + \frac{f''(x_0)}{2!} \cdot h^2 - \frac{f'''(x_0)}{3!} \cdot h^3 + \frac{f^{IV}(x_0)}{4!} \cdot h^4 - \frac{f^V(\xi_{-1})}{5!} \cdot h^5$$

Innen

$$f_1 - f_{-1} = \frac{f'(x_0)}{1!} \cdot 2h + \frac{f'''(x_0)}{3!} \cdot 2h^3 + \frac{f^V(\xi_1) + f^V(\xi_{-1})}{5!} \cdot h^5$$

Hasonlóan (h helyett $2h$-t szerepeltetve):

$$f_2 - f_{-2} = \frac{f'(x_0)}{1!} \cdot 4h + \frac{f'''(x_0)}{3!} \cdot 16h^3 + \frac{32f^V(\xi_2) + 32f^V(\xi_{-2})}{5!} \cdot h^5$$

Ebből kivonva az előző egyenlet 2-szeresét, az elsőrendű deriváltat tartalmazó tagok kiesnek:

$$f_2 - f_{-2} - 2f_1 + 2f_{-1} = f'''(x_0) \cdot 2h^3 +$$

$$+\frac{32f^V(\xi_2)+32f^V(\xi_{-2})-2f^V(\xi_1)-2f^V(\xi_{-1})}{5!}\cdot h^5$$

Innen

$$\left|\frac{-f_{-2}+2f_{-1}-2f_1+f_2}{2h^3}-f'''(x_0)\right|\le\frac{68h^5\cdot||f^V||_{\max}}{2h^3\cdot 5!}=$$

$$=\frac{17\cdot||f^V||_{\max}}{60}\cdot h^2$$

A formula tehát az $f'''(x_0)$ deriváltat közelíti, h szerint másodrendben.

4. Legyen $f:\mathbf{R}\to\mathbf{R}$ háromszor folytonosan differenciálható. Legyen $x_0\in\mathbf{R}$ tetszőleges, $h>0$ egy lépésköz, és $x_1:=x_0+h$, $x_2:=x_0+4h$. Jelölje $f_0:=f(x_0)$, $f_1:=f(x_1)$, $f_2:=f(x_2)$. A fenti alappontokra és alapponti értékekre támaszkodva, konstruáljunk differenciasémákat az $f'(x_0)$ elsőrendű és az $f''(x_0)$ másodrendű deriváltra. Hányadrendű közelítés érthető el?

Megoldás: Fejtsük f-et x_0 körül véges Taylor-sorba:

$$f_1=f_0+f'(x_0)\cdot h+\frac{1}{2}f''(x_0)\cdot h^2+\mathcal{O}(h^3)$$

$$f_2=f_0+f'(x_0)\cdot 4h+\frac{1}{2}f''(x_0)\cdot 16h^2+\mathcal{O}(h^3)$$

Az első egyenlet 16-szorosából kivonva a másodikat, a másodrendű deriváltat tartalmazó tagok kiesnek, és azt kapjuk, hogy

$$16f_1-f_2=15f_0+f'(x_0)\cdot 12h+\mathcal{O}(h^3)$$

Innen $f'(x_0)$ közelítése kifejezhető:

$$f'(x_0)\approx\frac{-15f_0+16f_1-f_2}{12h},$$

és a séma pontossága is adódik:

$$\left|f'(x_0)-\frac{-15f_0+16f_1-f_2}{12h}\right|=\mathcal{O}(h^2),$$

tehát a séma másodrendben közelíti $f'(x_0)$-t.

Az $f''(x_0)$ másodrendű derivált közelítése érdekében az elsőrendű $f'(x_0)$ deriváltat kell kiküszöbölni a fenti egyenlőségekből. Az első egyenlet 4-szereséből kivonva a másodikat, az elsőrendű deriváltat tartalmazó tagok kiesnek, és azt kapjuk, hogy

$$4f_1 - f_2 = 3f_0 + \frac{1}{2}f''(x_0) \cdot (-12h^2) + \mathcal{O}(h^3)$$

Innen $f''(x_0)$ közelítése kifejezhető:

$$f''(x_0) \approx \frac{3f_0 - 4f_1 + f_2}{6h^2},$$

és a séma pontossága is adódik:

$$\left| f''(x_0) - \frac{3f_0 - 4f_1 + f_2}{6h^2} \right| = \mathcal{O}(h),$$

tehát a séma elsőrendben közelíti $f''(x_0)$-t.

Megjegyezzük, hogy a pontosság nagyságrendje nem javul, ha a Taylor-sorból több tagot veszünk figyelembe (ellenőrizzük!).