MŰEGYETEM 1782

TYPOTEX

FARAGÓ ISTVÁN – HORVÁTH RÓBERT

NUMERIKUS MÓDSZEREK

2013

Az Olvasó most egy egyetemi jegyzetet tart a kezében, vagy néz a számítógépe képernyőjén. E jegyzetet a Budapesti Műszaki és Gazdaságtudományi Egyetemen illetve az Eötvös Loránd Tudományegyetemen tartott numerikus módszerek kurzusainkhoz írtuk. Az írás során mindvégig azt vettük figyelembe, hogy a jegyzet segítségével hallgatóink alapos ismereteket tudjanak elsajátítani a tárgy témájában és egyben eredményesebben tudjanak felkészülni a vizsgákra.

A jegyzet elején összefoglaljuk a szükséges előismereteket. Ezután a matematikai modellalkotással foglalkozunk, részletesen kitérve a számítógépes számábrázolásra és az ebből eredő hibákra. Ezután a klasszikus numerikus analízis egyes fejezeteit vesszük sorra: numerikus lineáris algebra, polinominterpoláció, numerikus deriválás és integrálás, közönséges differenciálegyenletek kezdeti- és peremérték-feladatai. A jegyzetet a parciális differenciálegyenletek véges differenciás megoldásainak bemutatásával zárjuk.

A jegyzetbe nem akartunk több dolgot belezsúfolni, mint amiről egy két féléves kurzus során az előadásokon is szó lehet, de igyekeztünk azért az érdeklődő hallgatóknak is kitekintést nyújtani az előadások anyagán túlmutató elméletek felvillantásával vagy az ezeket tárgyaló irodalom megadásával. Mivel ez a jegyzet elektronikus formában lesz elérhető, így kihasználtuk azokat a lehetőségeket is, amiket az elektronikus forma megenged. Így számos helyen megadtunk internethivatkozásokat valamilyen szemléltető programhoz, bővebb leíráshoz vagy életrajzhoz.

Kulcsszavak: numerikus módszerek, numerikus lineáris algebra, numerikus deriválás és integrálás, interpoláció, differenciálegyenletek numerikus megoldása

Támogatás:

Készült a TÁMOP-4.1.2-08/2/A/KMR-2009-0028 számú, a „Természettudományos (matematika és fizika) képzés a műszaki és informatikai felsőoktatásban” című projekt keretében.

Készült:

a BME TTK Matematika Intézet gondozásában

Szakmai felelős vezető:

Ferenczi Miklós

Lektorálta:

Havasi Ágnes

Az elektronikus kiadást előkészítette:

Horváth Róbert

Címlap grafikai terve:

Csépány Gergely László, Tóth Norbert

Második, javított kiadás, 2013

Tartalomjegyzék

Előszó

Az Olvasó most egy egyetemi jegyzetet tart a kezében vagy néz a számítógépe képernyőjén. E jegyzetet a Budapesti Műszaki és Gazdaságtudományi Egyetemen illetve az Eötvös Loránd Tudományegyetemen tartott numerikus módszerek kurzusainkhoz írtuk. Az írás során mindvégig azt vettük figyelembe, hogy a jegyzet segítségével hallgatóink alapos ismereteket tudjanak elsajátítani a tárgy témájában és egyben eredményesebben tudjanak felkészülni a vizsgákra. Ezt a célt szolgálják a magyarázó ábrák, a szemléltető példák, az ellenőrző kérdések, a gyakorló feladatok és a jegyzet végén található szószedet is. A jegyzetbe nem akartunk több dolgot belezsúfolni, mint amiről egy két féléves kurzus során az előadásokon is szó lehet, de igyekeztünk azért az érdeklődő hallgatóknak is kitekintést nyújtani az előadások anyagán túlmutató elméletek felvillantásával vagy az ezeket tárgyaló irodalom megadásával.

A jegyzetben a definíciókat és tételeket vastag vonallal emeltük ki. Azokat a példákat, amelyek a jobb megértést segítik bekeretezve közöljük. Szintén bekeretezve szedtük az egyes algoritmusokat és programrészleteket. A bizonyítások végét ■, a példák és megjegyzések végét pedig ⋄ jel zárja. A definíciók, a tételek, a következmények és a megjegyzések fejezetenként folytonosan sorszámozódnak. A fontosabb fogalmakat dőlt betűvel szedtük. Általában ezek kerültek a szószedetbe is.

Mivel ez a jegyzet elektronikus formában lesz elérhető, így kihasználtuk azokat a lehetőségeket is, amiket az elektronikus forma megenged. Így számos helyen megadtunk internethivatkozásokat valamilyen szemléltető programhoz, bővebb leíráshoz vagy életrajzhoz. Természetesen mivel ezek internetes tartalmak, a jövőben változhatnak és elérhetetlenné is válhatnak. A képletekre, tételekre vagy a szószedetbeli elemekre való hiperhivatkozások a pdf fájlban egy kattintással elérhetők, majd az ALT+← billentyűvel visszatérhetünk ez eredeti olvasási helyhez.

Köszönet illeti hallgatóinkat, akik az elmúlt félévek során alaposan átnézték a jegyzet korábbi változatait, megjegyzéseikkel hozzájárultak az anyag kialakulásához és végleges formába öntéséhez, és a korábbi változatokban lévő hibákra felhívták figyelmünket. Köszönet illeti Dr. Havasi Ágnest értékes javaslataiért, aki a tőle megszokott alapossággal nézte át a kéziratot. A jegyzet a TÁMOP – 4.1.2. – 08/2/A/KMR: Természettudományos (matematika és fizika) képzés a műszaki és informatikai felsőoktatásban pályázat támogatásával jött létre.

Budapest, 2011. január

A Szerzők

Nagyon köszönjük mindenkinek, hogy megosztotta velünk észrevételeit és javaslatait a jegyzettel kapcsolatban a hibabejelentő oldalon. A második, javított kiadásban már figyelembe vettük ezeket.

Budapest, 2013. augusztus

A Szerzők

1. Előismeretek

Ebben a fejezetben azokat az előismereteket gyűjtjük össze, amik nem tartoznak szorosan a numerikus módszerek tárgy témaköréhez, de ismeretük elengedhetetlen lesz a későbbiekben. Ezek az ismeretek főleg a lineáris algebra és a funkcionálanalízis tárgyhoz tartoznak. Bevezetjük a vektor- és mátrixnorma fogalmát, igazoljuk a Banach-féle fixponttételt, ismertetjük a Gram–Schmidt-féle ortogonalizációs eljárást, felsorolunk néhány nevezetes mátrixtípust és megvizsgáljuk a tulajdonságaikat. Szó lesz még a mátrixok sajátértékeiről és sajátvektorairól, ezek normákkal való kapcsolatáról, az M-mátrixokról ill. a diagonalizálható mátrixokról. Összehasonlítjuk a sorozatok és függvények konvergenciasebességét. A fejezetet a MATLAB programcsomag bemutatásával zárjuk.

Azok a hallgatók, akik tanultak lineáris algebrát és funkcionálanalízist e fejezet nagy részét átugorhatják az olvasás során. Bár a jelölések megismerésének érdekében érdemes minden fejezetet átszaladni, nekik csak a Gersgorin-tételt (1.2.14. tétel), a Banach-féle fixponttételt (1.1.18. tétel), a normák és sajátértékek kapcsolatáról szóló 1.2.3. fejezetet, az M-mátrixokról szóló 1.2.4. fejezetet és a konvergenciasebességről szóló 1.3. fejezetet érdemes alaposan átnézni.

1.1. Vektorterek

1.1.1. Valós és komplex vektorterek

Jelentse a továbbiakban $\mathbb{K}$ vagy a valós számok ($\mathbb{R}$) vagy a komplex számok ($\mathbb{C}$) testjét.

1.1.1. definíció.

Egy $V \neq \emptyset$ halmazt ($\mathbb{K} = \mathbb{R}$ esetén valós, $\mathbb{K} = \mathbb{C}$ estén komplex) *vektortérnek* nevezünk, ha értelmezve van rajta egy összeadás és egy számmal való szorzás művelet az alábbi tulajdonságokkal:

1. $x + y = y + x$, $\forall x, y \in V$,
2. $(x + y) + z = x + (y + z)$, $\forall x, y, z \in V$,
3. $\exists o \in V$, $x + o = x$, $\forall x \in V$,
4. $\forall x \in V$, $\exists \hat{x} \in V$, $x + \hat{x} = o$,
5. $1 \cdot x = x$, $\forall x \in V$,
6. $\alpha(x + y) = \alpha x + \alpha y$, $\forall x, y \in V$, $\forall \alpha \in \mathbb{K}$,
7. $(\alpha + \beta)x = \alpha x + \beta x$, $\forall x \in V$, $\forall \alpha, \beta \in \mathbb{K}$,
8. $\alpha(\beta x) = (\alpha\beta)x$, $\forall x \in V$, $\forall \alpha, \beta \in \mathbb{K}$.

A vektortér fenti axiómáiból könnyen nyerhetők az alábbi tulajdonságok: $0 \cdot x = o$ minden $x \in V$ esetén, $\alpha \cdot o = o$ minden $\alpha \in \mathbb{K}$ esetén és $\hat{x} = (-1) \cdot x$ minden $x \in V$ esetén. Ez utóbbi tulajdonság alapján az $x - y$ különbségen az $x + (-1) \cdot y$ összeget értjük.

Valós vektorteret alkotnak pl. a sík és a tér helyvektorai, az n-elemű valós oszlopvektorok halmaza ($\mathbb{R}^n$), az m-szer n-es valós mátrixok halmaza ($\mathbb{R}^{m\times n}$), az $[a,b]$ intervallumon folytonos függvények halmaza ($C[a,b]$), az $[a,b]$ intervallumon legalább k-szor folytonosan deriválható függvények halmaza ($C^k[a,b]$), a valós együtthatós polinomok halmaza (P_∞), a legfeljebb n-edfokú valós együtthatós polinomok halmaza (P_n) és ezek $[a,b]$ intervallumra vonatkozó leszorításai ($P_\infty[a,b]$, $P_n[a,b]$) a szokásos műveletek esetén[1].

Komplex vektorteret alkotnak pl. az n-elemű komplex oszlopvektorok halmaza ($\mathbb{C}^n$) és az m-szer n-es komplex mátrixok halmaza ($\mathbb{C}^{m\times n}$).

Ebben a fejezetben jelentsen a továbbiakban V egy adott (valós vagy komplex) vektorteret. V elemeit általánosan vektoroknak hívjuk.

1.1.2. definíció.

Egy $x \in V$ vektort az $x_1, \dots, x_k \in V$ vektorok *lineáris kombinációjának* hívunk, ha vannak olyan $\alpha_1, \dots, \alpha_k \in \mathbb{K}$ konstansok, hogy $x = \alpha_1 x_1 + \dots + \alpha_k x_k$.

1.1.3. definíció.

Egy V vektortér egy W részhalmazát a vektortér egy alterének hívjuk, ha W maga is vektortér a V-beli műveletekre nézve.

Például a legfeljebb harmadfokú polinomok vektorterében a legfeljebb másodfokú polinomok alteret alkotnak. Jelölje $\operatorname{lin}(x_1, x_2, \dots, x_n)$ az $x_1, x_2, \dots, x_n \in V$ vektorok összes lineáris kombinációjának halmazát. Ekkor $\operatorname{lin}(x_1, x_2, \dots, x_n)$ a V vektortér egy altere lesz a V-beli műveletekre nézve.

1.1.4. definíció.

Az $x_1, \dots, x_k \in V$ vektorrendszert *lineárisan függetlennek* mondjuk, ha az $\alpha_1 x_1 + \dots + \alpha_k x_k = o$ egyenlőségből $\alpha_i = 0$ $(i = 1, \dots, k)$ következik. Végtelen sok vektorból álló vektorrendszert akkor hívunk lineárisan függetlennek, ha bármely véges részhalmaza lineárisan független vektorokat tartalmaz. A nem lineárisan független vektorrendszereket lineárisan összefüggő rendszereknek hívjuk.

1.1.5. definíció.

Egy vektorrendszert a V vektortér *bázisának* hívunk, ha lineárisan független, és V minden eleme előállítható a vektorrendszer elemeinek lineáris kombinációjaként.

Bázisvektorok lineáris kombinációjaként minden V-beli vektor pontosan egyféleképpen írható fel. Ha V-nek van véges sok elemből álló bázisa, akkor V-t véges dimenziós vektortérnek hívjuk. Egy véges dimenziós vektortér minden bázisának egyforma az elemszáma. Ez a vektortér dimenziója.

[1] A jegyzetben használt vektorokkal és mátrixokkal kapcsolatos jelöléseket és elnevezéseket az 1.2. fejezetben foglaltuk össze.

1.1.2. Normált terek

1.1.6. definíció.
A $(V, \|\cdot\|)$ párt *normált térnek* hívjuk, ha V egy vektortér, és $\|\cdot\| : V \to \mathbb{R}$ egy adott függvény, ún. norma, az alábbi tulajdonságokkal:

1. $\|x\| = 0 \Leftrightarrow x = o$,
2. $\|\alpha x\| = |\alpha| \cdot \|x\|$, $\forall x \in V, \forall \alpha \in \mathbb{K}$,
3. $\|x + y\| \leq \|x\| + \|y\|$, $\forall x, y \in V$ (háromszög-egyenlőtlenség).

Mivel a sík- ill. a térvektorok vektorterében a vektorok hossza normát ad meg, ezért általánosan is szokás egy vektor normáját a vektor hosszának nevezni.

1.1.7. megjegyzés. Könnyen igazolható, hogy a norma csak nemnegatív értéket vehet fel. Vizsgáljuk ugyanis egy tetszőleges x elem esetén az $\|x - x\|$ értéket! A norma második és harmadik tulajdonságát felhasználva azt kapjuk, hogy

$$0 = \|o\| = \|x - x\| \leq \|x\| + \| - x\| = 2\|x\|,$$

amiből következik az állítás. $\diamond$

Most felsorolunk néhány fontos példát normált terekre.

- A sík és a tér helyvektorai, ha a $\|\vec{v}\|$ norma a vektor szokásos hossza.
- A $\mathbb{K}^n$ vektortér, ha egy $\overline{\mathbf{x}} = [x_1, \ldots, x_n]^T$ vektor esetén a normát pl. az

  $$\|\overline{\mathbf{x}}\|_p = \sqrt[p]{|x_1|^p + \cdots + |x_n|^p}$$

  képlettel értelmezzük $p = 1, 2, \ldots$ esetén. A leggyakrabban használt normák ezek közül az 1-es vagy oktaédernorma

  $$\|\overline{\mathbf{x}}\|_1 = |x_1| + \cdots + |x_n|$$

  és a 2-es vagy euklideszi norma

  $$\|\overline{\mathbf{x}}\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2},$$

  valamint a $p \to \infty$ határátmenettel nyert, ∞-nel jelölt maximumnorma

  $$\|\overline{\mathbf{x}}\|_\infty = \max\{|x_1|, \ldots, |x_n|\}.$$

  A $\mathbb{K}^n$ vektortéren megadott normákat *vektornormáknak* hívjuk.
- A $C[a, b]$ vektortér, ha a normát pl. az

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

  módon értelmezzük (maximumnorma), amely tulajdonképpen a függvénygrafikon x-tengelytől mért legnagyobb eltérésének nagyságát adja meg.

- A $\mathbb{K}^{m\times n}$ vektortér, ha a normát egy $\mathbf{A}=[a_{ij}]\in\mathbb{K}^{m\times n}$ mátrix esetén az

$$\|\mathbf{A}\| = \max_{i=1,\dots,m;\ j=1,\dots,n}\{|a_{ij}|\}$$

képlettel értelmezzük.

A $\mathbb{K}^{m\times n}$ vektortéren megadott normákat *mátrixnormáknak* hívjuk. Később majd látni fogunk más fontos mátrixnormákat is.

A norma alkalmas arra, hogy mérjük két folytonos függvény, két vektor vagy két mátrix "távolságát". Így mérni tudjuk, hogy pl. egy lineáris egyenletrendszer közelítő megoldása "milyen messze" van a pontos megoldástól. A távolság segítségével konvergenciát is definiálhatunk.

1.1.8. definíció.
Az $x,y\in(V,\|\cdot\|)$ vektorok *távolságán* az $\|x-y\|$ számot értjük.

A távolság elnevezés jogosságát az alábbi tétel mutatja.

1.1.9. tétel.
A fent definiált távolságra teljesülnek az alábbi tulajdonságok:
1. $\|x-y\|\geq 0$, $\forall x,y\in(V,\|\cdot\|)$, $\|x-y\|=0\Leftrightarrow x=y$,
2. $\|x-y\|=\|y-x\|$, $\forall x,y\in(V,\|\cdot\|)$,
3. $\|x-y\|\leq\|x-z\|+\|z-y\|$, $\forall x,y,z\in(V,\|\cdot\|)$ (háromszög-egyenlőtlenség).

A háromszög-egyenlőtlenség közvetlen következménye az alábbi tétel, ami azt mutatja, hogy két vektor normájának eltérése tetszőlegesen kicsi lehet, ha a két vektor távolságát elegendően kicsinek választjuk.

1.1.10. tétel.
Egy $(V,\|\cdot\|)$ normált térben $|\|x\|-\|y\||\leq\|x-y\|$ minden $x,y\in(V,\|\cdot\|)$ esetén.

Bizonyítás. Alkalmazzuk kétféleképpen a háromszög-egyenlőtlenséget:

$$\|y\| = \|(y-x)+x\| \leq \|y-x\|+\|x\|,$$

$$\|x\| = \|(x-y)+y\| \leq \|x-y\|+\|y\|.$$

Az első egyenlőtlenségből kapjuk, hogy $\|y\|-\|x\|\leq\|y-x\|$, a másikból pedig hogy $\|x\|-\|y\|\leq\|x-y\|$. Az utóbbi egyenlőséget az $\|y\|-\|x\|\geq-\|y-x\|$ alakba írva a két egyenlőség együttesen a

$$-\|y-x\|\leq\|y\|-\|x\|\leq\|y-x\|$$

alakot ölti, ami a bizonyítandó állítással ekvivalens. ■

1.1.11. definíció.
Azt mondjuk, hogy az $\{x_k\}\subset(V,\|\cdot\|)$ sorozat tart az $x\in(V,\|\cdot\|)$ elemhez (konvergens), ha az $\{\|x_k-x\|\}$ valós számsorozat nullához tart. Jelölés: $x_k\to x$.

Az x vektort a sorozat határértékének hívjuk. Könnyen igazolható, hogy a határérték egyértelmű.

1.1.12. definíció.

Azt mondjuk, hogy egy $H \subset (V, \|\cdot\|)$ *halmaz zárt*, ha minden olyan $\{x_k\} \subset H$ sorozatra, amely tart valamilyen $x \in (V, \|\cdot\|)$ elemhez, igaz, hogy $x \in H$. Egy $H \subset (V, \|\cdot\|)$ *halmaz nyílt*, ha komplementere zárt.

1.1.13. definíció.

Egy V vektortéren értelmezett $\|\cdot\|_\star$ és $\|\cdot\|_{\star\star}$ normákat ekvivalensnek nevezzük, ha vannak olyan $c_1, c_2 > 0$ konstansok, melyekre

$$c_1\|x\|_\star \leq \|x\|_{\star\star} \leq c_2\|x\|_\star, \ \forall x \in V.$$

Könnyen látható, hogy a normák ekvivalenciája ekvivalencia-reláció, azaz reflexív, szimmetrikus és tranzitív. Ekvivalens normák ugyanazt a konvergenciát definiálják. Ez azt jelenti, hogy ha egy sorozat az egyik normában tart egy adott elemhez, akkor a másik normában is ahhoz az elemhez fog tartani. A későbbiekben többször alkalmazzuk majd az alábbi tételt.

1.1.14. tétel.

Véges dimenziós vektorterekben minden norma ekvivalens.

Bizonyítás. Legyen V egy véges dimenziós vektortér a $v_1, \ldots, v_n$ bázissal. Ebben a vektortérben minden x vektor egyértelműen írható fel $x = \sum_{k=1}^n \alpha_k v_k$ alakban, ahol az α_k együtthatók $\mathbb{K}$-beli egyértelműen meghatározott konstansok. Ekkor a vektortérben a $\mu(x) = \sqrt{\sum_{k=1}^n |\alpha_k|^2}$ függvény normát definiál (1.6.10. feladat). Legyen $\|\cdot\|$ egy tetszőleges norma az adott V vektortéren. A tétel igazolásához elegendő megmutatnunk, hogy $\|\cdot\|$ és μ ekvivalens normák, mert a normák tranzitivitása miatt így bármely két norma ekvivalens lesz.

Legyen x egy tetszőleges V-beli vektor. Ekkor

$$\|x\| = \left\|\sum_{k=1}^n \alpha_k v_k\right\| \leq \sum_{k=1}^n |\alpha_k|\|v_k\| \leq \sqrt{\sum_{k=1}^n |\alpha_k|^2}\sqrt{\sum_{k=1}^n \|v_k\|^2} = c_2\mu(x),$$

ahol $c_2 = \sqrt{\sum_{k=1}^n \|v_k\|^2}$ egy, az x vektortól független konstans. Az utolsó becslésnél a Cauchy–Schwarz-egyenlőtlenséget használtuk. Így a $\|\cdot\|$ norma felülről becsülhető a μ norma konstansszorosával.

Az alsó becsléshez tekintsük az euklideszi normával ellátott $\mathbb{K}^n$ teret, melyen definiáljuk az $f : (\mathbb{K}^n, \|\cdot\|_2) \to \mathbb{R}$, $f(\overline{\chi}) = f(\chi_1, \ldots, \chi_n) = \|\sum_{k=1}^n \chi_k v_k\|$ függvényt. Ez a függvény folytonos, ugyanis az 1.1.10. tétel alapján tetszőleges $\overline{\gamma} = (\gamma_1, \ldots, \gamma_n), \overline{\beta} = (\beta_1, \ldots, \beta_n) \in \mathbb{K}^n$ vektorok esetén

$$|f(\overline{\gamma}) - f(\overline{\beta})| = \left|\left\|\sum_{k=1}^n \gamma_k v_k\right\| - \left\|\sum_{k=1}^n \beta_k v_k\right\|\right| \leq \left\|\sum_{k=1}^n (\gamma_k - \beta_k)v_k\right\| \leq c_2\|\overline{\gamma} - \overline{\beta}\|_2.$$

Mivel az f függvény tehát folytonos, így a

$$G = \{\overline{\chi} \in \mathbb{K}^n \mid \|\overline{\chi}\|_2 = 1\}$$

korlátos és zárt gömbhéjon van legkisebb értéke. Legyen ez a legkisebb érték $f^\star$. Az $f^\star$ érték nyilvánvalóan nagyobb nullánál, hiszen különben a $v_1, \ldots, v_n$ vektorok nem lennének függetlenek.

Mivel $x \neq o$ esetén $\mu(x/\mu(x)) = 1$, ezért $\|x/\mu(x)\| \geq f^\star$, amiből következik, hogy $\|x\| \geq f^\star \mu(x)$. Ez mutatja, hogy $c_1 = f^\star$ megfelelő választás. Ezt akartuk megmutatni. ■

1.1.15. definíció.

Azt mondjuk, hogy az $\{x_k\} \subset (V, \|\cdot\|)$ sorozat *Cauchy-sorozat*, ha minden $\varepsilon > 0$ számhoz van olyan $M \in \mathbb{N}$ szám, melyre $\|x_n - x_m\| < \varepsilon$ minden $n, m \geq M$ esetén.

1.1.16. tétel.

Minden $(V, \|\cdot\|)$ normált térbeli konvergens sorozat Cauchy-sorozat.

A tétel megfordítása nem igaz.

1.1.17. definíció.

Azt mondjuk, hogy a $(V, \|\cdot\|)$ normált tér *Banach*[2]*-tér* , ha minden $(V, \|\cdot\|)$-beli Cauchy-sorozat konvergens sorozat is egyben.

A normált terekre korábban felsorolt példák egyben példák Banach-terekre is. Tehát pl. $\mathbb{R}^n$ Banach-tér a felsorolt normákkal, és mivel ezen a vektortéren minden norma ekvivalens, ezért bármilyen más normával is. Ugyanakkor nem minden normált tér Banach-tér. Ha a $C[a, b]$ vektortéren a normát az $\|f\| = \int_a^b |f(x)|\,\mathrm{d}x$ módon definiáljuk, akkor az így nyert normált tér nem lesz Banach-tér. Most igazoljuk azt a tételt, amely a későbbi iterációs eljárások konvergenciáját fogja majd biztosítani.

1.1.18. tétel. (Banach-féle fixponttétel)

Legyen $(V, \|\cdot\|)$ egy Banach-tér, és $H \subset (V, \|\cdot\|)$ egy tetszőleges nem üres zárt részhalmaz. Tegyük fel, hogy az $F : H \to H$ leképezés kontrakció, azaz van olyan $0 \leq q < 1$ valós szám, mellyel

$$\|F(x) - F(y)\| \leq q\|x - y\|$$

bármely $x, y \in H$ elemek esetén.

- Ekkor F-nek egyértelműen létezik fixpontja H-ban, azaz egy olyan $x^\star \in H$ elem, mellyel $F(x^\star) = x^\star$.
- Tetszőleges $x_0 \in H$ kezdőelemmel az $x_{k+1} = F(x_k)$ módon előállított sorozat $x^\star$-hoz tart.
- Érvényes az

$$\|x^\star - x_m\| \leq \frac{q^m}{1-q}\|x_1 - x_0\| \tag{1.1.1}$$

becslés.

[2]Stefan Banach (1892 (Lvov)-1945), lengyel matematikus. A modern funkcionálanalízis megalapítója. Eredményei jelentősen hozzájárultak a topologikus vektorterek, a mértékelmélet, az integrálás és az ortogonális sorok elméletéhez is. Részletes angol nyelvű életrajz található pl. az http://www-history.mcs.st-and.ac.uk/Biographies/Banach.html oldalon.

Bizonyítás. Tekintsük egy tetszőleges $x_0 \in H$ elem esetén az $x_{k+1} = F(x_k)$ rekurzióval definiált sorozatot, melynek nyilvánvalóan mindegyik eleme H-ban található. Ekkor a kontrakciós tulajdonság miatt

$$\|x_{k+1} - x_k\| = \|F(x_k) - F(x_{k-1})\| \leq q\|x_k - x_{k-1}\| \leq \ldots \leq q^k\|x_1 - x_0\|.$$

Tetszőleges két $n > m$ természetes szám esetén

$$\begin{aligned}
\|x_n - x_m\| &= \|x_n - x_{n-1} + x_{n-1} - x_{n-2} + \ldots + x_{m+1} - x_m\| \\
&\leq \|x_n - x_{n-1}\| + \|x_{n-1} - x_{n-2}\| + \ldots + \|x_{m+1} - x_m\| \\
&\leq q^{n-1}\|x_1 - x_0\| + q^{n-2}\|x_1 - x_0\| + \ldots + q^m\|x_1 - x_0\| \\
&= (q^{n-1} + q^{n-2} + \ldots + q^m)\|x_1 - x_0\| \\
&= (q^{n-m-1} + q^{n-m-2} + \ldots + 1)q^m\|x_1 - x_0\| \\
&= \frac{q^{n-m} - 1}{q - 1}q^m\|x_1 - x_0\| \leq \frac{q^m}{1 - q}\|x_1 - x_0\|.
\end{aligned} \tag{1.1.2}$$

Ez mutatja, hogy $\{x_k\}$ egy H-beli Cauchy-sorozat, hiszen $0 \leq q < 1$, és $\varepsilon > 0$ esetén

$$M = \left\lceil \frac{\ln(\varepsilon(1-q)/\|x_1 - x_0\|)}{\ln q} \right\rceil$$

jó választás. Mivel Banach-terekben minden Cauchy-sorozat konvergens, ezért létezik olyan $x^\star \in (V, \|\cdot\|)$, melyre $x_k \to x^\star$. H zártsága miatt $x^\star \in H$ is igaz. Most azt fogjuk igazolni, hogy $x^\star$ fixpontja F-nek. Ha $x_1 = x_0$, akkor ez nyilvánvaló. Mivel

$$\|F(x^\star) - x_{k+1}\| = \|F(x^\star) - F(x_k)\| \leq q\|x^\star - x_k\| \to 0 \ (k \to \infty),$$

ezért $x_{k+1} \to F(x^\star)$. Mivel $x_{k+1} \to x^\star$ is igaz, így a határérték egyértelműségéből következik, hogy $F(x^\star) = x^\star$. Az egyértelműség igazolásához indirekt módon feltételezzük, hogy van legalább két különböző fixpont: $x^\star$ és $x^{\star\star}$. Ekkor

$$\|x^\star - x^{\star\star}\| = \|F(x^\star) - F(x^{\star\star})\| \leq q\|x^\star - x^{\star\star}\|,$$

ami nyilván csak úgy lehet $(q < 1)$, ha $x^\star = x^{\star\star}$, ami ellentmondás.

Az állítás harmadik részében szereplő becslés úgy igazolható, hogy az n indexszel végtelenhez tartunk az (1.1.2) becslésben. ■

A tételben természetesen H lehet a teljes $(V, \|\cdot\|)$ normált tér is. Vegyük észre, hogy a tétel második állítása gyakorlati útmutatást is ad arra, hogy a fixpontot hogy kell megkeresnünk. A harmadik részben szereplő becslés pedig a fixponthoz tartó sorozat első két elemének távolságával és a q konstanssal ad felső becslést arra, hogy a sorozat m-edik eleme milyen messze van a határértékétől. Vegyük észre azt is, hogy az $\{x_k\}$ sorozat kezdőeleme tetszőleges volt, így az $x^\star$ fixpont tetszőleges H-beli kezdőelemről induló iterációs sorozat határértékeként előállítható.

1.1.19. definíció.
Egy $F : (V_1, \|\cdot\|_\star) \to (V_2, \|\cdot\|_{\star\star})$ *leképezés folytonos* az $x^\star \in (V_1, \|\cdot\|_\star)$ pontban, ha minden $\{x_k\} \subset (V_1, \|\cdot\|_\star)$ sorozatra, melyre $x_k \to x^\star$, következik, hogy $F(x_k) \to F(x^\star)$ $(V_2, \|\cdot\|_{\star\star})$-ben. F folytonos, ha minden $x^\star \in (V_1, \|\cdot\|_\star)$ pontban folytonos.

Fontos példa, hogy az $F : (V, \|\cdot\|) \to (\mathbb{R}, |.|)$, $F(x) = \|x\|$ folytonos leképezés, hiszen tetszőleges $x_k \to x$ $(V, \|\cdot\|)$-beli sorozat esetén minden k indexre igaz, hogy $|\|x_k\| - \|x\|| \leq \|x_k - x\|$ (1.1.10. tétel), azaz $\|x_k\| \to \|x\|$.

1.1.20. definíció.
Egy $F : (V_1, \|\cdot\|_\star) \to (V_2, \|\cdot\|_{\star\star})$ *leképezés korlátos*, ha van olyan $K \in \mathbb{R}_0^+$ szám, melyre $\|F(x)\|_{\star\star} \leq K \cdot \|x\|_\star$ minden $x \in (V_1, \|\cdot\|_\star)$ esetén.

1.1.21. definíció.
Egy $F : (V_1, \|\cdot\|_\star) \to (V_2, \|\cdot\|_{\star\star})$ leképezést *lineáris operátornak* nevezünk, ha $F(\alpha x + \beta y) = \alpha F(x) + \beta F(y)$ minden $x, y \in (V_1, \|\cdot\|_\star)$, $\alpha, \beta \in \mathbb{K}$ esetén.

1.1.22. tétel.
Lineáris operátorokra a folytonosság és a korlátosság ekvivalens tulajdonságok. Ha egy lineáris operátor folytonos egy pontban, akkor folytonos $(V_1, \|\cdot\|_\star)$ minden pontjában.

Jelölje $B(V_1, V_2)$ az összes korlátos $L : (V_1, \|\cdot\|_\star) \to (V_2, \|\cdot\|_{\star\star})$ lineáris operátor vektorterét, ahol a műveleteket az

$$(L_1 + L_2)(x) = L_1(x) + L_2(x), \quad (\alpha L)(x) = \alpha \cdot L(x)$$

módon értelmezzük.

1.1.23. tétel.
Az

$$\|L\| := \sup_{x \neq o} \frac{\|L(x)\|_{\star\star}}{\|x\|_\star}$$

(a korlátosság miatt jól definiált) hozzárendelés normát ad meg a $B(V_1, V_2)$ vektortéren, így $B(V_1, V_2)$ normált tér. (Ha V_2 Banach-tér, akkor $B(V_1, V_2)$ is Banach-tér.)

Alkalmazzuk az előző tételt az $L : (\mathbb{K}^n, \|\cdot\|_\star) \to (\mathbb{K}^m, \|\cdot\|_{\star\star})$, $L(\overline{\mathbf{x}}) = \mathbf{A}\overline{\mathbf{x}}$ lineáris leképezésre, ahol $\mathbf{A} \in \mathbb{K}^{m \times n}$. A normák ekvivalenciája miatt (1.1.14. tétel) az L leképezés folytonos, azaz korlátos. Ekkor az előző tételt alkalmazva az

$$\|\mathbf{A}\| := \|L\| = \sup_{\overline{\mathbf{x}} \neq \overline{\mathbf{0}}} \frac{\|\mathbf{A}\overline{\mathbf{x}}\|_{\star\star}}{\|\overline{\mathbf{x}}\|_\star} \tag{1.1.3}$$

hozzárendelés mátrixnormát ad meg. A vektornormákból a fenti képlettel származtatott mátrixnormákat *indukált normáknak* hívjuk.

1.1.24. tétel.
Tegyük fel, hogy a $\mathbb{K}^n$ és $\mathbb{K}^m$ normált terekben is ugyanazt a vektornormát használjuk. Ekkor a korábban megismert vektornormák az alábbi mátrixnormákat indukálják:

- Oktaédernorma ($p = 1$): $\|\mathbf{A}\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^m |a_{ij}|$ (oszlopösszegnorma),
- Maximumnorma ($p = \infty$): $\|\mathbf{A}\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}|$ (sorösszegnorma),
- Euklideszi norma ($p = 2$): $\|\mathbf{A}\|_2 = \sqrt{\varrho(\mathbf{A}^H\mathbf{A})}$, ahol ϱ az $\mathbf{A}$ mátrix spektrálsugara, és $\mathbf{A}^H$ az $\mathbf{A}$ mátrix transzponált konjugáltja.

Bizonyítás. Alkalmazzuk az (1.1.3) képletet a mátrixnormára.
Oktaédernorma: Legyen $\mathbf{A} \in \mathbb{K}^{m\times n}$ egy adott mátrix és $\overline{\mathbf{x}} \in \mathbb{K}^n$ egy tetszőleges vektor. Ekkor

$$\|\mathbf{A}\overline{\mathbf{x}}\|_1 = \sum_{i=1}^{m}\left|\sum_{j=1}^{n} a_{ij}x_j\right| \le \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}||x_j| = \sum_{j=1}^{n}\sum_{i=1}^{m}|a_{ij}||x_j| = \sum_{j=1}^{n}\left(|x_j|\sum_{i=1}^{m}|a_{ij}|\right) \le$$

$$\le \left(\max_{j=1,\dots,n}\sum_{i=1}^{m}|a_{ij}|\right)\sum_{j=1}^{n}|x_j| = \left(\max_{j=1,\dots,n}\sum_{i=1}^{m}|a_{ij}|\right)\|\overline{\mathbf{x}}\|_1,$$

ami mutatja, hogy $\|\mathbf{A}\|_1 \le \max_{j=1,\dots,n}\sum_{i=1}^{m}|a_{ij}|$. Az egyenlőséghez azt kell megmutatni, hogy van olyan $\overline{\mathbf{x}}_0 \in \mathbb{K}^n$ vektor, mellyel a fenti becslésekben egyenlőségek szerepelnek. Tegyük fel, hogy a $\sum_{i=1}^{m}|a_{ij}|$ összeg a j_0 oszlopban a legnagyobb. Ekkor az $\overline{\mathbf{x}}_0 = \overline{\mathbf{e}}_{j_0}\sum_{i=1}^{m}|a_{ij_0}|$ választás megfelelő, ugyanis

$$\|\mathbf{A}\overline{\mathbf{x}}_0\|_1 = \left(\sum_{i=1}^{m}|a_{ij_0}|\right)\sum_{i=1}^{m}|a_{ij_0}| = \left(\max_{j=1,\dots,n}\sum_{i=1}^{m}|a_{ij}|\right)\|\overline{\mathbf{x}}_0\|_1.$$

Itt $\overline{\mathbf{e}}_{j_0}$ a j_0-adik egységvektort jelöli, azaz azt az n elemű vektort, melynek j_0-adik eleme 1, a többi pedig nulla.

Maximumnorma: Az oktaédernormához hasonlóan igazolható. Lásd az 1.6.2. feladatot a fejezet végén.

Euklideszi norma: Később igazoljuk (28. oldal). Most még nem áll rendelkezésünkre minden eszköz a bizonyításhoz. ■

1.1.25. megjegyzés. Unitér és ortogonális mátrixok 2-es normája 1, ugyanis $\|\mathbf{A}\|_2 = \sqrt{\varrho(\mathbf{A}^H\mathbf{A})} = \sqrt{\varrho(\mathbf{E})} = 1$. Unitér és ortogonális mátrixszal való szorzás nem változtatja meg egy mátrix 2-es normáját. Legyen $\mathbf{B}$ tetszőleges mátrix és $\mathbf{A}$ egy unitér mátrix, melyek összeszorozhatók $\mathbf{AB}$ alakban. Ekkor

$$\|\mathbf{AB}\|_2 = \sqrt{\varrho((\mathbf{AB})^H(\mathbf{AB}))} = \sqrt{\varrho(\mathbf{B}^H\mathbf{A}^H\mathbf{AB})} = \sqrt{\varrho(\mathbf{B}^H\mathbf{B})} = \|\mathbf{B}\|_2.$$

◇

1.1.26. megjegyzés. Diagonális mátrixok p-normája megegyezik a főátlóban lévő legnagyobb elemabszolútértékkel. ◇

1.1.27. tétel.

Tegyük fel, hogy a $\mathbb{K}^n$-beli $\|.\|_v$ vektornorma a $\mathbb{K}^{n\times n}$-beli $\|\cdot\|_m$ mátrixnormát indukálta. Ekkor igazak az alábbi tulajdonságok

- $\|\mathbf{Ax}\|_v \le \|\mathbf{A}\|_m \cdot \|\mathbf{x}\|_v$ minden $\mathbf{x} \in \mathbb{K}^n$ vektor és $\mathbf{A} \in \mathbb{K}^{n\times n}$ mátrix esetén (konzisztencia tulajdonság),
- Az $\mathbf{E}$ egységmátrixra $\|\mathbf{E}\|_m = 1$,
- $\|\mathbf{AB}\|_m \le \|\mathbf{A}\|_m \cdot \|\mathbf{B}\|_m$ minden $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{n\times n}$ mátrixok esetén (szubmultiplikativitási tulajdonság).

Az $\|\mathbf{A}\| = \max_{i,j}\{|a_{ij}|\}$ képlettel adott mátrixnorma nem indukált norma. Az $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j}|a_{ij}|^2}$ ún. Frobenius-norma sem indukált norma.

1.1.3. Euklideszi terek

1.1.28. definíció.
A $(V, \langle .,. \rangle)$ párt *euklideszi térnek* hívjuk, ha V egy vektortér, és $\langle .,. \rangle : (V \times V) \to \mathbb{K}$ egy adott függvény, ún. skaláris szorzat, az alábbi tulajdonságokkal:

1. $\langle x, y \rangle = \overline{\langle y, x \rangle}$ minden $x, y \in V$ esetén (a $^-$ föléhúzás a komplex konjugálást jelenti),
2. $\langle \alpha x + \beta y, z \rangle = \alpha\langle x, z \rangle + \beta\langle y, z \rangle$, minden $x, y, z \in V,\ \alpha, \beta \in \mathbb{K}$ esetén,
3. $\langle x, x \rangle > 0$, minden $o \neq x \in V$ esetén.

Tekintsünk két fontos példát euklideszi térre.

- A $\mathbb{K}^n$ oszlopvektorok terében, az $\overline{\mathbf{x}} = [x_1, \ldots, x_n]^T$ és $\overline{\mathbf{y}} = [y_1, \ldots, y_n]^T$ jelöléssel az $\langle \overline{\mathbf{x}}, \overline{\mathbf{y}} \rangle = \overline{x}_1 y_1 + \ldots + \overline{x}_n y_n$ hozzárendelés skaláris szorzást ad meg. Megjegyezzük, hogy az egyszerűség kedvéért a későbbiekben a mátrixszorzás szabályát használva az $\langle \overline{\mathbf{x}}, \overline{\mathbf{y}} \rangle = \overline{\mathbf{x}}^H \overline{\mathbf{y}}$ vagy valós vektorok esetén az $\langle \overline{\mathbf{x}}, \overline{\mathbf{y}} \rangle = \overline{\mathbf{x}}^T \overline{\mathbf{y}}$ írásmódot fogjuk használni, hiszen ezek olyan 1×1-es mátrixok, melyek egyetlen eleme éppen a skaláris szorzat értéke.

- A $C[a, b]$ vektortéren az
$$\langle f, g \rangle = \int_a^b s(x) f(x) g(x)\, \mathrm{d}x \tag{1.1.4}$$
hozzárendelés skaláris szorzást definiál minden s $C[a, b]$-beli pozitív ún. súlyfüggvény esetén.

Könnyen igazolható, hogy egy euklideszi térben az $\|x\| = \sqrt{\langle x, x \rangle}$ hozzárendelés normát definiál. Ezt a normát a skaláris szorzás által indukált normának nevezzük.

1.1.29. definíció.
Egy euklideszi tér x és y elemét ortogonálisnak hívjuk, ha $\langle x, y \rangle = 0$. Azt mondjuk, hogy az x elem normált, ha a skaláris szorzás által indukált normája 1. Egy vektorrendszer ortogonális, ha bármely két különböző eleme ortogonális. Egy vektorrendszer ortonormált, ha ortogonális és minden eleme normált.

Többször fontos szerepet fog játszani az az eljárás, mellyel lineárisan független vektorokból ortonormált rendszert lehet készíteni. Ezt az eljárást Gram–Schmidt-féle ortogonalizációs eljárásnak (röviden GS ortogonalizáció) hívjuk.

1.1.30. tétel. (Gram–Schmidt ortogonalizáció)
Egy euklideszi térben minden lineárisan független $x_1, \ldots, x_k$ vektorrendszerből előállítható egy olyan ortonormált $q_1, \ldots, q_k$ vektorrendszer, melyre $\mathrm{lin}(q_1, q_2, \ldots, q_l) = \mathrm{lin}(x_1, x_2, \ldots, x_l)$ minden $l = 1, \ldots, k$ index esetén.

Bizonyítás. Könnyen látható, hogy a $\hat{q}_1 = x_1$,

$$\hat{q}_l = x_l - \sum_{i=1}^{l-1} \frac{\langle \hat{q}_i, x_l \rangle}{\langle \hat{q}_i, \hat{q}_i \rangle} \hat{q}_i \ (l = 2, \dots, k)$$

módon előállított $\hat{q}_1, \dots, \hat{q}_k$ vektorok ortogonálisak, és az ezekből nyert

$$q_l = \frac{\hat{q}_l}{\|\hat{q}_l\|} \ (l = 1, \dots, k)$$

vektorok pedig ortonormált rendszert alkotnak. ■

A numerikus matematikában fontos szerepet játszanak az ortogonális polinomok. Az 1.1.29. definíció alkalmazásával két $p, q \in P_\infty[a, b]$ polinomot ortogonálisnak hívunk az $[a, b]$ intervallumon az s pozitív súlyfüggvényre nézve, ha

$$\int_a^b s(x)p(x)q(x)\,\mathrm{d}x = 0.$$

Az $1, x, x^2$ stb. polinomokból a $[-1, 1]$ intervallumon a Gram–Schmidt-ortogonalizációs eljárással készült polinomokat $s(x) \equiv 1$ súlyfüggvény esetén Legendre-polinomoknak, míg az $s(x) = 1/\sqrt{1-x^2}$ súlyfüggvény esetén Csebisev-polinomoknak nevezzük. Az első négy ortogonális polinomot adtuk meg az alábbi táblázatban.

Fokszám	Legendre	Csebisev
0	1	1
1	x	x
2	$(3x^2 - 1)/2$	$2x^2 - 1$
3	$(5x^3 - 3x)/2$	$4x^3 - 3x$
4	$(35x^4 - 30x^2 + 3)/8$	$8x^4 - 8x^2 + 1$

1.1.1. táblázat: Néhány Legendre- és Csebisev-polinom.

Az ortogonális polinomok fontos tulajdonságait foglalja össze az alábbi tétel.

1.1.31. tétel.

Tegyük fel, hogy a $p_0, p_1, \dots$ (az alsó index a fokszámot jelöli) polinomok páronként ortogonálisak az $[a, b]$ intervallumon egy adott s pozitív súlyfüggvényre. Ekkor a polinomoknak minden zérushelye valós, egyszeres és az $[a, b]$ intervallumba esik.

Bizonyítás. Tekintsük a p_l polinomot, és jelölje $z_1, \dots, z_k$ a páratlan multiplicitású különböző valós zérushelyeket $[a, b]$-ben. Ha $k = l$, akkor igaz az állítás, ha $k < l$, akkor tekintsük a $p(x) = (x - z_1) \dots (x - z_k)$ ($p \equiv 1$, ha $k = 0$) polinomot, amely k-ad fokú. A $p_l \cdot p$ polinom $(l+k)$-ad fokú, és nem vált előjelet $[a, b]$-ben (minden $[a, b]$ intervallumba eső valós gyöktényező páros hatványon szerepel). Így az

$$\int_a^b p_l(x)p(x)s(x)\,\mathrm{d}x = 0$$

feltétel nem teljesülhet. Ezzel igazoltuk az állítást. ■

1.2. Mátrixok

Ebben a jegyzetben az oszlopvektorokat föléhúzott félkövér kisbetűkkel, míg a mátrixokat félkövér nagybetűkkel jelöljük. Egy $\mathbf{A}$ mátrix elemeit általában a_{ij}-vel jelöljük, de abban az esetben, ha nem szeretnénk új jelölést bevezetni a mátrix elemeire, alkalmazzuk az $(\mathbf{A})_{ij}$ jelölést is. Hasonló módon járunk el a vektorok esetén is.

A jelölések megkönnyítése érdekében több esetben alkalmazzuk a MATLAB[3] programcsomag jelöléseit is. Pl. egy $\mathbf{A} \in \mathbb{K}^{m\times n}$ mátrix esetén

- $\mathbf{A}(:, c:d)$ az $\mathbf{A}$ mátrix $c, \ldots, d$ sorszámú oszlopait tartalmazó mátrix,
- $\mathbf{A}(a:b, c:d)$ az $\mathbf{A}(:, c:d)$ mátrix $a, \ldots, b$ sorszámú sorait tartalmazó mátrix,
- $\mathbf{A}(a:b, :)$ az $\mathbf{A}$ mátrix $a, \ldots, b$ sorait tartalmazó mátrix,
- $\mathrm{diag}(\mathbf{A})$ az $\mathbf{A}$ mátrix főátlóbeli elemeit tartalmazó oszlopvektor.

Vektorok esetén

- $\mathrm{diag}(\overline{\mathbf{v}})$ azt a diagonális mátrixot jelenti, melynek diagonális elemei rendre megegyeznek a $\overline{\mathbf{v}}$ vektor elemeivel.

Most felsorolunk néhány fontos mátrixtulajdonságot ill. elnevezést.

- Azokat a vektorokat ill. mátrixokat, melyek minden eleme nulla *nullvektornak* ill. *nullmátrixnak* nevezzük. Jelölésükre egységesen a $\mathbf{0}$ jelölést használjuk. Méretét külön nem jelöljük, az következik a képletekben szereplő többi mátrix méretéből.
- Azokat a mátrixokat, melyeknek ugyanannyi sora van ahány oszlopa, *négyzetes vagy kvadratikus* mátrixoknak nevezzük.
- Egy $\mathbf{A}$ valós mátrixot *szimmetrikusnak* hívunk, ha $\mathbf{A}^T = \mathbf{A}$, ahol $(\cdot)^T$ jelöli a transzponálás műveletét ($(\mathbf{A}^T)_{ij} = (\mathbf{A})_{ji}$). Egy $\mathbf{A}$ komplex mátrixot hermitikusnak[4] hívunk, ha $\mathbf{A}^H = \mathbf{A}$, ahol $(\cdot)^H$ jelöli a mátrix konjugáltjának transzponáltját.
- Egy $\mathbf{A} \in \mathbb{K}^{m\times n}$ mátrixról azt mondjuk, hogy *sávmátrix*, ha léteznek olyan $p, q \in \mathbb{N}$ konstansok, hogy $a_{ij} = 0$ ha $j < i-p$ és ha $j > i+q$. Az $1+p+q$ értéket a mátrix *sávszélességének* nevezzük.

  - Diagonális a mátrix, ha $p = 0,\ q = 0$. Speciális négyzetes diagonális mátrixként $\mathbf{E}$ fogja jelölni az egységmátrixot. A méretét általában nem jelöljük, mindig olyan méretűnek tekintjük, hogy elvégezhetőek legyenek vele a mátrixműveletek. Egy mátrix közvetlenül a főátló "feletti" ("alatti") elemeit a mátrix szuperdiagonálisának (szubdiagonálisának) nevezzük.

  - Felső háromszögmátrixról beszélünk, ha a főátló "alatti" elemek nullák ($p = 0$). Pl.

$$\begin{bmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}$$

[3] A MATLAB programcsomagról az 1.4. fejezetben írunk részletesebben.

[4] Charles Hermite (1822–1901), francia matematikusról elnevezett mátrixtípus. Ő mutatta meg, hogy az hermitikus mátrixok minden sajátértéke valós.

egy felső háromszögmátrix. A mátrixban a $*$ jel azt jelenti, hogy azon a helyen tetszőleges szám állhat mátrixelemként.

- Alsó háromszögmátrix: főátló "feletti" elemek nullák ($q = 0$).

- Felső Hessenberg-mátrix: a szubdiagonál "alatti" elemek nullák ($p = 1$).

- Alsó Hessenberg-mátrix: a szuperdiagonál "feletti" elemek nullák ($q = 1$).

- Tridiagonális mátrix: egyszerre alsó- és felső Hessenberg-mátrix ($p = 1, q = 1$). A tridiag(a, b, c) mátrix egy olyan tridiagonális mátrixot jelöl, melynek főátlójában b, tőle balra a és tőle jobbra c szerepel.

- Egy $\mathbf{A}$ négyzetes mátrix esetén azt a mátrixot, mellyel akár balról, akár jobbról szorozzuk $\mathbf{A}$-t, az egységmátrixot kapjuk eredményül, az $\mathbf{A}$ *mátrix inverzének* nevezzük. Jelölése: $\mathbf{A}^{-1}$ ($\mathbf{A}\mathbf{A}^{-1} = \mathbf{E}$). Azokat a mátrixokat, melyeknek van inverze reguláris vagy nemszinguláris mátrixoknak nevezzük. Legyenek az $\mathbf{A} \in \mathbb{R}^{k\times n}$ mátrix oszlopvektorai $\overline{\mathbf{a}}_1, \ldots, \overline{\mathbf{a}}_n$! Ekkor a lin$(\overline{\mathbf{a}}_1, \ldots, \overline{\mathbf{a}}_n)$ vektortér dimenzióját az $\mathbf{A}$ *mátrix rangjának* nevezzük és $r(\mathbf{A})$-val jelöljük. Egy $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix determinánsát a det$(\mathbf{A})$ módon jelöljük. Egy négyzetes mátrixnak pontosan akkor van inverze, ha determinánsa nullától különbözik, ami pontosan akkor teljesül, ha a mátrix rangja megegyezik oszlopainak számával. Fontos szabály a determinánsok szorzási szabálya, amely szerint két tetszőleges $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ mátrix esetén $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$.

- Az i-edik egységvektort $\overline{\mathbf{e}}_i$ fogja jelölni, azaz $\overline{\mathbf{e}}_i = [0, \ldots, 0, 1, 0, \ldots, 0]^T$, ahol az 1-es az i-edik elem a vektorban. A vektor méretét nem jelöljük, az mindig következik a képletekben szereplő többi mátrix méretéből.

- Két azonos méretű mátrix között az "$=$", "$<$", "$\leq$", "$>$", "$\geq$" relációkat elemenként értelmezzük. Egy $\mathbf{A}$ mátrixot nemnegatívnak ill. pozitívnak nevezünk, ha az $\mathbf{A} \geq \mathbf{0}$ ill. $\mathbf{A} > \mathbf{0}$ feltételek teljesülnek. A nempozitív ill. negatív tulajdonságokat hasonlóan definiáljuk. Ha $\mathbf{B}$ és $\mathbf{C}$ két azonos méretű mátrix, melyek balról szorozhatók az $\mathbf{A} \geq \mathbf{0}$ mátrixszal, akkor a $\mathbf{B} \leq \mathbf{C}$ feltételből következik az $\mathbf{AB} \leq \mathbf{AC}$ feltétel, ugyanis az $\mathbf{A}(\mathbf{C} - \mathbf{B})$ szorzatnak minden eleme nemnegatív és így a szorzat maga is egy nemnegatív mátrix lesz.

- Azokat a mátrixokat, melyek elemei mátrixok, *blokkmátrixoknak* hívjuk. Megadásukra a szokásos (a MATLAB-ban is használt) jelölést alkalmazzuk. Pl. ha $\overline{\mathbf{a}} = [1, 2, 3]^T$ és

$$\mathbf{B} = \begin{bmatrix} 3 & 4 & 4 \\ -4 & 0 & -1 \\ 5 & 2 & 2 \end{bmatrix},$$

akkor

$$\begin{bmatrix} 1 & \overline{\mathbf{a}}^T \\ \overline{\mathbf{a}} & \mathbf{B} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 & 3 \\ 1 & 3 & 4 & 4 \\ 2 & -4 & 0 & -1 \\ 3 & 5 & 2 & 2 \end{bmatrix}.$$

- Egy $\mathbf{A}$ valós mátrixot *ortogonálisnak* hívunk, ha van inverze, és $\mathbf{A}^{-1} = \mathbf{A}^T$. Az elnevezés onnét ered, hogy ebben az esetben a mátrix oszlopvektorai ortonormáltak a szokásos $\mathbb{R}^n$-beli skaláris szorzásra nézve. Egy $\mathbf{A}$ mátrixot *unitér* mátrixnak hívunk, ha van inverze és $\mathbf{A}^{-1} = \mathbf{A}^H$. Hermitikus ill. ortogonális mátrixok szorzata is hermitikus ill. ortogonális, hiszen ha pl. $\mathbf{A}$ és $\mathbf{B}$ ortogonálisak ($\mathbf{A}^{-1} = \mathbf{A}^T$ és $\mathbf{B}^{-1} = \mathbf{B}^T$), akkor $\mathbf{E} = (\mathbf{AB})(\mathbf{B}^T\mathbf{A}^T) = (\mathbf{AB})(\mathbf{AB})^T$, azaz $\mathbf{AB}$ inverze a transzponáltja, azaz $\mathbf{AB}$ ortogonális.

- Egy hermitikus $\mathbf{A}$ mátrix esetén az $\overline{\mathbf{x}} \mapsto \overline{\mathbf{x}}^H \mathbf{A}\mathbf{x}$ függvény minden oszlopvektorhoz hozzárendel egy valós számot. Ez abból következik, hogy az eredmény egy (1×1)-es mátrix, és $(\overline{\mathbf{x}}^H \mathbf{A}\overline{\mathbf{x}})^H = \overline{\mathbf{x}}^H \mathbf{A}\overline{\mathbf{x}}$, azaz $\overline{\mathbf{x}}^H \mathbf{A}\overline{\mathbf{x}}$ értéke valós.

1.2.1. definíció.

Legyen $\mathbf{A}$ egy adott hermitikus mátrix. Ekkor, ha tetszőleges $\overline{\mathbf{x}} \neq \mathbf{0}$ vektorra igaz, hogy

- $\overline{\mathbf{x}}^H \mathbf{A}\overline{\mathbf{x}} > 0$ ($\overline{\mathbf{x}}^H \mathbf{A}\overline{\mathbf{x}} < 0$), akkor az $\mathbf{A}$ mátrixot *pozitív (negatív) definit mátrixnak*,
- $\overline{\mathbf{x}}^H \mathbf{A}\overline{\mathbf{x}} \geq 0$ ($\overline{\mathbf{x}}^H \mathbf{A}\overline{\mathbf{x}} \leq 0$), akkor az $\mathbf{A}$ mátrixot *pozitív (negatív) szemidefinit mátrixnak*,
- $\overline{\mathbf{x}}^H \mathbf{A}\overline{\mathbf{x}}$ lehet pozitív és negatív is, akkor az $\mathbf{A}$ mátrixot *indefinit mátrixnak*

nevezzük.

Ha az $\mathbf{A}$ mátrix valós és szimmetrikus, akkor az adott tulajdonságok teljesüléséhez elegendő, ha a feltételek valós $\overline{\mathbf{x}}$ vektorokkal teljesülnek.

- Egy $\mathbf{P} = [\overline{\mathbf{e}}_{i_1}, \ldots, \overline{\mathbf{e}}_{i_n}] \in \mathbb{R}^{n \times n}$ alakú mátrixot, ahol $i_1, \ldots, i_n$ az $1, 2, \ldots, n$ számok egy permutációja, *permutációs mátrixnak* hívunk. Egy $\mathbf{A} \in \mathbb{R}^{n \times n}$ mátrix esetén az $\mathbf{AP}$ szorzat $\mathbf{A}$ oszlopainak $i_1, \ldots, i_n$ sorrendű átrendezését adja, míg a $\mathbf{P}^T\mathbf{A}$ szorzat $\mathbf{A}$ sorait rendezi át az említett sorrendbe. Érvényes továbbá a $\mathbf{P}\mathbf{P}^T = \mathbf{P}^T\mathbf{P} = \mathbf{E}$ reláció, azaz a permutációs mátrixok ortogonálisak.

1.2.1. Mátrixok sajátértékei és sajátvektorai

1.2.2. definíció.

Legyen $\mathbf{A} \in \mathbb{C}^{n \times n}$ egy tetszőleges négyzetes mátrix. Ha egy nullvektortól különböző $\mathbf{0} \neq \overline{\mathbf{v}} \in \mathbb{C}^n$ vektor és egy $\lambda \in \mathbb{C}$ szám esetén teljesül az

$$\mathbf{A}\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}$$

egyenlőség, akkor a $\overline{\mathbf{v}}$ vektort a mátrix sajátvektorának, és a λ számot a sajátvektorhoz tartozó sajátértéknek nevezzük (és fordítva). Egy összetartozó sajátértéket és sajátvektort *sajátpárnak* nevezünk.

Egyszerűen igazolhatók az alábbi tételek.

1.2.3. tétel.

Egy mátrix adott sajátértékhez tartozó sajátvektorai a nullvektorral kiegészítve $\mathbb{C}^n$ egy alterét alkotják.

Bizonyítás. Azt kell igazolnunk csak, hogy ha $\overline{\mathbf{v}}_1$ és $\overline{\mathbf{v}}_2$ két különböző, λ-hoz tartozó sajátvektor, akkor tetszőleges c_1, c_2 számok esetén $c_1\overline{\mathbf{v}}_1 + c_2\overline{\mathbf{v}}_2 \neq \mathbf{0}$ is sajátvektor. Ez következik az

$$\mathbf{A}(c_1\overline{\mathbf{v}}_1 + c_2\overline{\mathbf{v}}_2) = c_1\mathbf{A}\overline{\mathbf{v}}_1 + c_2\mathbf{A}\overline{\mathbf{v}}_2 = c_1\lambda\overline{\mathbf{v}}_1 + c_2\lambda\overline{\mathbf{v}}_2 = \lambda(c_1\overline{\mathbf{v}}_1 + c_2\overline{\mathbf{v}}_2)$$

egyenlőségből. ■

1.2.4. tétel.
Egy $\mathbf{A} \in \mathbb{C}^{n\times n}$ mátrixnak a multiplicitást is figyelembe véve pontosan n darab sajátértéke van. A sajátértékek a $\det(\mathbf{A}-\lambda\mathbf{E}) = 0$ egyenlet megoldásai. Egy adott λ sajátértékhez tartozó $\overline{\mathbf{v}}$ sajátvektorokat az $(\mathbf{A}-\lambda\mathbf{E})\overline{\mathbf{v}} = \mathbf{0}$ lineáris algebrai egyenletrendszer nullvektortól különböző megoldásai adják.

Bizonyítás. Az $\mathbf{A}\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}$ egyenlőséget átrendezve kapjuk, hogy a sajátvektornak az $(\mathbf{A} - \lambda\mathbf{E})\overline{\mathbf{v}} = \mathbf{0}$ egyenletrendszer nullvektortól különböző megoldásának kell lennie. Mivel ez az egyenletrendszer homogén, így biztosan van megoldása, hiszen a nullvektor megoldás lesz. Ahhoz, hogy λ sajátérték legyen, pontosan az kell, hogy legyen nullvektortól különböző megoldása is az egyenletrendszernek. Ez pontosan akkor teljesül, ha $\det(\mathbf{A}-\lambda\mathbf{E}) = 0$. Ez az egyenlet azt mutatja, hogy a sajátértékek a $p_{\mathbf{A}}(\lambda) := \det(\mathbf{A}-\lambda\mathbf{E})$ polinom zérushelyei. Ezekről pedig az algebra alaptételéből tudjuk, hogy multiplicitással együtt n darab van belőlük. ■

1.2.5. definíció.
Egy adott $\mathbf{A} \in \mathbb{C}^{n\times n}$ mátrix esetén a $p_{\mathbf{A}}(\lambda) = \det(\mathbf{A}-\lambda\mathbf{E})$ polinomot a mátrix *karakterisztikus polinomjának*, a $P_{\mathbf{A}}(\lambda) = 0$ egyenletet pedig *karakterisztikus egyenletnek* nevezzük.

1.2.6. megjegyzés. Ha $\mathbf{A}$ valós mátrix, λ pedig egy valós sajátértéke, akkor az $(\mathbf{A} - \lambda\mathbf{E})\overline{\mathbf{v}} = \mathbf{0}$ egyenlőség miatt választható a sajátértékhez valós $\overline{\mathbf{v}}$ sajátvektor. ◇

1.2.7. tétel.
Jelölje $\lambda_1, \ldots, \lambda_n$ az $\mathbf{A} \in \mathbb{C}^{n\times n}$ mátrix sajátértékeit. Ekkor

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i, \quad \mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i,$$

ahol $\mathrm{tr}(\mathbf{A})$ a mátrix nyoma (angolul trace), azaz a főátlóban szereplő elemek összege.

Bizonyítás. Ismert, egyébként könnyen igazolható, hogy egy $p(\lambda) = a_n\lambda^n + a_{n-1}\lambda^{n-1} + a_1\lambda + a_0$ n-edfokú polinom $\lambda_1, \ldots, \lambda_n$ zérushelyeire igazak az alábbi formulák (ún. Viéte-formulák):

$$\lambda_1\lambda_2\ldots\lambda_n = (-1)^n\frac{a_0}{a_n}, \;\; \lambda_1 + \lambda_2 + \ldots + \lambda_n = -\frac{a_{n-1}}{a_n}.$$

A karakterisztikus polinom legmagasabbfokú tagjának együtthatója $a_n = (-1)^n$,

$$a_{n-1} = (-1)^{n-1}(a_{11} + \ldots + a_{nn}) = (-1)^{n-1}(\mathrm{tr}(\mathbf{A}))$$

és a szabad tag $a_0 = \det(\mathbf{A})$. Ezekből az állítás közvetlenül adódik. ■

1.2.8. tétel.
Egy $\mathbf{A} \in \mathbb{C}^{n\times n}$ mátrixnak pontosan akkor nincs nulla sajátértéke, ha nemszinguláris, azaz ha van inverze.

Bizonyítás. Tekintsük az $\mathbf{A}\overline{\mathbf{v}} = \mathbf{0}$ homogén lineáris egyenletrendszert. Ennek pontosan akkor a nullvektor az egyetlen megoldása, ha $\det(\mathbf{A}) = \det(\mathbf{A} - 0\mathbf{E}) \neq 0$. Ez egyenértékű azzal, hogy a nulla nem sajátértéke a mátrixnak. ■

1.2.9. megjegyzés. Az előző tételt gyakran használjuk annak igazolására, hogy egy mátrix nemszinguláris. Ugyanis ehhez azt kell megmutatnunk, hogy nincs olyan nullától különböző vektor, mellyel a mátrixot megszorozva nullvektort kapunk. ◇

1.2.10. tétel.
Hermitikus mátrixok minden sajátértéke valós.

Bizonyítás. Legyen $\overline{\mathbf{v}}$ a mátrix egy sajátvektora λ sajátértékkel. Ekkor $\overline{\mathbf{v}}^H\mathbf{A}\overline{\mathbf{v}} = \overline{\mathbf{v}}^H\lambda\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}^H\overline{\mathbf{v}}$. Nyilván

$$(\overline{\mathbf{v}}^H\mathbf{A}\overline{\mathbf{v}})^H = \overline{\mathbf{v}}^H\mathbf{A}\overline{\mathbf{v}}, \quad (\overline{\mathbf{v}}^H\overline{\mathbf{v}})^H = \overline{\mathbf{v}}^H\overline{\mathbf{v}},$$

azaz ezek olyan (1×1)-es mátrixok, melyek transzponált konjugáltja önmaga, azaz valós számokat tartalmaznak. Így λ is valós kell legyen. ■

1.2.11. megjegyzés. Az előző tétel alapján a valós szimmetrikus mátrixoknak is minden sajátértéke valós, és ezekhez a sajátértékekhez valós sajátvektorok választhatók. ◇

1.2.12. tétel.
Hermitikus pozitív (szemi)definit mátrixok minden sajátértéke (nemnegatív) pozitív.

Bizonyítás. Legyen $\overline{\mathbf{v}}$ egy sajátvektora a mátrixnak λ sajátértékkel. Az előző tételből tudjuk, hogy λ valós. Ekkor $\overline{\mathbf{v}}^H\mathbf{A}\overline{\mathbf{v}} = \overline{\mathbf{v}}^H\lambda\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}^H\overline{\mathbf{v}} > 0$, és a $\overline{\mathbf{v}}^H\overline{\mathbf{v}} > 0$ egyenlőtlenségből következik az állítás (szemidefinitre hasonlóan). ■

1.2.13. definíció.
Egy $\mathbf{A} \in \mathbb{C}^{n\times n}$ mátrix legnagyobb abszolútértékű sajátértékének abszolút értékét $\mathbf{A}$ *spektrálsugarának* hívjuk. Jelölés: $\varrho(\mathbf{A})$. Azaz $\varrho(\mathbf{A}) = \max_{i=1,\dots,n}\{|\lambda_i| \,|\, \lambda_i \text{ sajátértéke } \mathbf{A}\text{-nak}\}$.

A sajátértékek komplex számsíkon való elhelyezkedésére ad becslést az alábbi ún. Gersgorin[5]-tétel.

1.2.14. tétel. (Gersgorin-tétel)
Tekintsük az $\mathbf{A} \in \mathbb{C}^{n\times n}$ mátrixot. Legyen K_i a komplex számsíkon az a zárt körlap, melynek középpontja a_{ii}, és sugara $\sum_{j=1,j\neq i}^{n} |a_{ij}|$ $(i = 1,\dots,n)$. Ekkor a mátrix sajátértékei az $\cup_{i=1,\dots,n} K_i$ halmazban találhatók.

Bizonyítás. Legyen λ egy sajátértéke a mátrixnak. Ha λ megegyezik valamelyik diagonális elemmel, akkor erre a sajátértékre igaz az állítás. Különben írjuk fel $\mathbf{A}$-t $\mathbf{A} = \mathbf{D} + \mathbf{T}$ alakban, ahol $\mathbf{D} = \operatorname{diag}(\operatorname{diag}(\mathbf{A}))$ az $\mathbf{A}$ diagonálisát tartalmazó mátrix. Az $\mathbf{A} - \lambda\mathbf{E}$ mátrix szinguláris, így van olyan $\overline{\mathbf{x}} \neq \mathbf{0}$ vektor, mellyel $(\mathbf{A} - \lambda\mathbf{E})\overline{\mathbf{x}} = \mathbf{0}$, azaz $(\mathbf{D} - \lambda\mathbf{E})\overline{\mathbf{x}} = -\mathbf{T}\overline{\mathbf{x}}$. A bal oldali mátrix invertálható, hiszen olyan diagonális mátrix, melynek egyik főátlóbeli eleme sem nulla. Így

$$\|\overline{\mathbf{x}}\|_\infty \leq \|(\mathbf{D} - \lambda\mathbf{E})^{-1}\mathbf{T}\|_\infty \|\overline{\mathbf{x}}\|_\infty,$$

[5] Szemjon Aranovics Gersgorin (1901–1933), belorusz matematikus. A sajátértékek becsléséről szóló cikkét 1931-ben publikálta. Bővebb életrajz:
http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Gershgorin.html

amiből $\|\overline{\mathbf{x}}\|_\infty$-val való osztás után kapjuk, hogy

$$1 \leq \frac{\sum_{j=1, j\neq k}^{n} |a_{kj}|}{|a_{kk} - \lambda|}$$

valamely $k = 1, \ldots, n$ indexre, azaz λa a K_k körlap belsejébe esik. ■

1.2.15. megjegyzés. Ha s darab körlap uniója diszjunkt a többi körlappal, akkor az unióban pontosan s darab sajátérték van. Ez az ún. második Gersgorin-tétel. ⋄

1.2.16. tétel.
Különböző sajátértékekhez tartozó sajátvektorok lineárisan függetlenek.

Bizonyítás. Elegendő a tételt csak két sajátértékre igazolni úgy, hogy megmutatjuk, hogy az egyikhez tartozó egy sajátvektor nem fejezhető ki a másikhoz tartozó sajátvektorok lineáris kombinációjaként. Tegyük fel tehát indirekt, hogy egy $\mathbf{A}$ mátrixnak $\lambda \neq \mu$ két sajátértéke, továbbá, hogy $\mathbf{A}\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}$ és $\mathbf{A}\overline{\mathbf{w}}_i = \mu\overline{\mathbf{w}}_i$ $(i = 1, \ldots, l)$ esetén a $\overline{\mathbf{v}}$ sajátvektor $\overline{\mathbf{v}} = \sum_{i=1}^{l} \alpha_i \overline{\mathbf{w}}_i$ alakban írható megfelelő α_i konstansokkal. Ekkor

$$\lambda\overline{\mathbf{v}} = \mathbf{A}\overline{\mathbf{v}} = \mathbf{A}\sum_{i=1}^{l} \alpha_i \overline{\mathbf{w}}_i = \mu \sum_{i=1}^{l} \alpha_i \overline{\mathbf{w}}_i = \mu\overline{\mathbf{v}},$$

ami csak úgy lehetne, ha $\lambda = \mu$. Ez ellentmondás. ■

1.2.17. következmény. A tétel közvetlen következménye, hogy ha egy $(n \times n)$-es mátrixnak minden sajátértéke különböző, akkor van n darab lineárisan független sajátvektorrendszere, azaz választható a sajátvektorai közül n darab lineárisan független vektor. ⋄

1.2.2. Diagonalizálhatóság

1.2.18. definíció.
Az $\mathbf{A}$ és $\mathbf{B}$ ugyanolyan méretű négyzetes mátrixokat *hasonlónak* hívjuk, ha van olyan $\mathbf{S}$ reguláris mátrix, melyre $\mathbf{B} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$.

A hasonlóság ekvivalenciareláció.

1.2.19. tétel.
Hasonló mátrixok sajátértékei megegyeznek.

Bizonyítás. Elég megmutatnunk, hogy a két hasonló $\mathbf{A}$ és $\mathbf{B}$ mátrix karakterisztikus polinomja ugyanaz. Legyen $\mathbf{B} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$. Mivel $\det(\mathbf{S})\det(\mathbf{S}^{-1}) = \det(\mathbf{E}) = 1$, ezért

$$\det(\mathbf{B} - \lambda\mathbf{E}) = \det(\mathbf{S}^{-1}\mathbf{A}\mathbf{S} - \lambda\mathbf{E})$$

$$= \det(\mathbf{S}^{-1})\det(\mathbf{A} - \lambda\mathbf{E})\det(\mathbf{S}) = \det(\mathbf{A} - \lambda\mathbf{E}). \blacksquare$$

1.2.20. megjegyzés. Könnyen látható, hogy hasonló mátrixok sajátvektorai között fennáll az alábbi összefüggés: ha $\overline{\mathbf{v}}$ sajátvektora $\mathbf{B}$-nek, akkor $\mathbf{S}\overline{\mathbf{v}}$ sajátvektora $\mathbf{A}$-nak. ⋄

1.2.21. definíció.
Egy $\mathbf{A}$ mátrixot *diagonalizálhatónak* hívunk, ha hasonló egy diagonális mátrixhoz.

1.2.22. megjegyzés. Nem diagonalizálható mátrix például az

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

mátrix. Ennek a mátrixnak kétszeres sajátértéke az 1, így az egységmátrixszal kellene hasonlónak lennie, de akkor valamilyen reguláris $\mathbf{S}$ mátrixszal $\mathbf{A} = \mathbf{S}^{-1}\mathbf{E}\mathbf{S} = \mathbf{E}$, ami nyilván nem igaz. ◇

1.2.23. tétel.
Egy $(n \times n)$-es mátrix pontosan akkor diagonalizálható, ha van n elemű lineárisan független sajátvektorrendszere.

Bizonyítás. Tegyük fel először, hogy az $(n \times n)$-es $\mathbf{A}$ mátrixnak van n darab lineárisan független sajátvektora, azaz $\mathbf{A}\overline{\mathbf{v}}_j = \lambda_j \overline{\mathbf{v}}_j$ $(j = 1, \ldots, n)$ úgy, hogy a $\overline{\mathbf{v}}_j$ sajátvektorok lineárisan függetlenek. Ekkor igaz az

$$\mathbf{A} \underbrace{\begin{bmatrix} \overline{\mathbf{v}}_1 & \ldots & \overline{\mathbf{v}}_n \end{bmatrix}}_{:=\mathbf{S}} = \begin{bmatrix} \overline{\mathbf{v}}_1 & \ldots & \overline{\mathbf{v}}_n \end{bmatrix} \underbrace{\begin{bmatrix} \lambda_1 & 0 & 0 & \ldots \\ 0 & \lambda_2 & 0 & \ldots \\ & & \ddots & \end{bmatrix}}_{:=\mathbf{\Lambda}}$$

egyenlőség. Bevezetve az $\mathbf{S} = \begin{bmatrix} \overline{\mathbf{v}}_1 & \ldots & \overline{\mathbf{v}}_n \end{bmatrix}$ jelölést arra a mátrixra, melynek oszopvektorai a sajátvektorok, írhatjuk, hogy $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{\Lambda}$, azaz a mátrix diagonalizálható.

A másik irány igazolásához tegyük fel, hogy van olyan reguláris $\mathbf{S}$ mátrix, mellyel $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{\Lambda}$, valamilyen $\mathbf{\Lambda}$ diagonális mátrixszal. Ekkor $\mathbf{A}$ sajátértékei megegyeznek $\mathbf{\Lambda}$ elemeivel. Mivel az $\overline{\mathbf{e}}_j$ rendszer sajátvektorrendszere a $\mathbf{\Lambda}$ mátrixnak, így $\mathbf{S}\overline{\mathbf{e}}_j$ sajátvektorrendszere $\mathbf{A}$-nak. Ezek $\mathbf{S}$ regularitása miatt lineárisan függetlenek vektorok. ■

1.2.24. definíció.
Egy $\mathbf{A}$ mátrixot *normális mátrixnak* hívunk, ha $\mathbf{A}^H\mathbf{A} = \mathbf{A}\mathbf{A}^H$.

1.2.25. megjegyzés. Valós normális mátrixok pl. a szimmetrikus és ortogonális mátrixok. Komplex normális mátrixok az hermitikus és unitér mátrixok. ◇

1.2.26. tétel.
Minden normális mátrix diagonalizálható.

Bizonyítás. Legyen $\mathbf{A}$ egy tetszőleges normális mátrix, és λ_1 és $\overline{\mathbf{v}}_1$ a mátrix egy sajátpárja. Legyen $\overline{\mathbf{v}}_1$ normált, azaz olyan, hogy $\overline{\mathbf{v}}_1^H\overline{\mathbf{v}}_1 = 1$. Egészítsük ki ezt a vektort a Gram–Schmidt-ortogonalizáció segítségével ortonormált rendszerré (unitér mátrixszá) a $\overline{\mathbf{v}}_2, \ldots, \overline{\mathbf{v}}_n$ vektorokkal.

Ekkor

$$\mathbf{A}\underbrace{\begin{bmatrix} \overline{\mathbf{v}}_1 & \dots & \overline{\mathbf{v}}_n \end{bmatrix}}_{=:\mathbf{S}_1\ (\text{unitér})} = \begin{bmatrix} \overline{\mathbf{v}}_1 & \dots & \overline{\mathbf{v}}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & * & * & \dots \\ 0 & * & * & \dots \\ \vdots & \vdots & \vdots & \\ 0 & * & * & \dots \end{bmatrix},$$

ahonnét, bevezetve az $\mathbf{S}_1 = [\overline{\mathbf{v}}_1 \dots \overline{\mathbf{v}}_n]$ jelölést és az $\mathbf{A}_2$ jelölést a jobb oldali második tényező $(2:n, 2:n)$ blokkjára, kapjuk, hogy

$$\mathbf{S}_1^H \mathbf{A} \mathbf{S}_1 = \begin{bmatrix} \lambda_1 & * \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}.$$

Hajtsuk végre az előző eljárást az $\mathbf{A}_2$ mátrixszal! Ehhez létezik olyan $\tilde{\mathbf{S}}_2$ unitér mátrix, mellyel

$$\tilde{\mathbf{S}}_2^H \mathbf{A}_2 \tilde{\mathbf{S}}_2 = \begin{bmatrix} \lambda_2 & * & * & \dots \\ 0 & * & * & \dots \\ & & \ddots & \\ 0 & * & * & \dots \end{bmatrix}.$$

Legyen

$$\mathbf{S}_2 = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{S}}_2 \end{bmatrix}.$$

Ekkor

$$\mathbf{S}_2^H \mathbf{S}_1^H \mathbf{A} \mathbf{S}_1 \mathbf{S}_2 = \begin{bmatrix} \lambda_1 & * & * & \dots \\ 0 & \lambda_2 & * & \dots \\ 0 & 0 & * & \dots \\ & & \vdots & \vdots \\ 0 & 0 & * & \dots \end{bmatrix}.$$

Hasonlóan folytatva nyerhetőek az $\mathbf{S}_3, \dots, \mathbf{S}_{n-1}$ unitér mátrixok, melyekkel

$$\mathbf{S}_{n-1}^H \dots \mathbf{S}_2^H \mathbf{S}_1^H \mathbf{A} \mathbf{S}_1 \mathbf{S}_2 \dots \mathbf{S}_{n-1} = \underbrace{\begin{bmatrix} \lambda_1 & * & * & \dots & * \\ 0 & \lambda_2 & * & \dots & * \\ & & \ddots & & \\ 0 & 0 & 0 & \dots & \lambda_n \end{bmatrix}}_{=:\mathbf{T}\ (\text{felső háromszög})}.$$

Legyen $\mathbf{S} = \mathbf{S}_1 \dots \mathbf{S}_{n-1}$. Ez nyilvánvalóan unitér mátrix. Vezessük be a $\mathbf{T}$ jelölést a fenti képletben szereplő felső háromszögmátrixra. Erre a mátrixra igaz, hogy

$$\mathbf{T}^H\mathbf{T} = \mathbf{S}^H\mathbf{A}^H\mathbf{S}\mathbf{S}^H\mathbf{A}\mathbf{S} = \mathbf{S}^H\mathbf{A}^H\mathbf{A}\mathbf{S} = \mathbf{S}^H\mathbf{A}\mathbf{A}^H\mathbf{S} = \mathbf{S}^H\mathbf{A}\mathbf{S}\mathbf{S}^H\mathbf{A}^H\mathbf{S} = \mathbf{T}\mathbf{T}^H,$$

így $\mathbf{T}$ normális mátrix. Mivel $\mathbf{T}$ felső háromszögmátrix, csak úgy lehet normális, ha diagonális (1.6.19. feladat). Vagyis az $\mathbf{A}$ normális mátrix unitér mátrixszal diagonalizálható. ■

1.2.27. következmény. A tétel bizonyításából közvetlenül következik, hogy minden $\mathbf{A}$ négyzetes mátrix felírható $\mathbf{A} = \mathbf{S}\mathbf{T}\mathbf{S}^H$ alakban, ahol $\mathbf{S}$ unitér mátrix, $\mathbf{T}$ pedig egy felső háromszögmátrix. Ezt az alakot a mátrixok *Schur-felbontásának* nevezzük. Vegyük észre, hogy a hasonlóság miatt a $\mathbf{T}$ mátrix főátlójában az $\mathbf{A}$ mátrix sajátértékei szerepelnek. ◇

1.2.28. következmény. Az előző tétel másik következménye, hogy egy mátrix akkor és csak akkor diagonalizálható unitér mátrixszal, ha normális. Az, hogy a normális mátrixok unitér mátrixszal diagonalizálhatók, következik az előző tétel bizonyításából. A másik irány igazolásához tegyük fel, hogy $\mathbf{A}$ unitér mátrixszal diagonalizálható, azaz van olyan $\mathbf{S}$ unitér mátrix, mellyel $\mathbf{S}^H\mathbf{A}\mathbf{S} = \mathbf{\Lambda}$, ahol $\mathbf{\Lambda}$ diagonális mátrix. Ekkor viszont $\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^H$, és

$$\begin{aligned}\mathbf{A}^H\mathbf{A} &= (\mathbf{S}\mathbf{\Lambda}\mathbf{S}^H)^H(\mathbf{S}\mathbf{\Lambda}\mathbf{S}^H) = \mathbf{S}\mathbf{\Lambda}^H\mathbf{S}^H\mathbf{S}\mathbf{\Lambda}\mathbf{S}^H = \mathbf{S}\mathbf{\Lambda}^H\mathbf{\Lambda}\mathbf{S}^H \\ &= \mathbf{S}\mathbf{\Lambda}\mathbf{\Lambda}^H\mathbf{S}^H = (\mathbf{S}\mathbf{\Lambda}\mathbf{S}^H)(\mathbf{S}\mathbf{\Lambda}^H\mathbf{S}^H) = \mathbf{A}\mathbf{A}^H.\end{aligned}$$

◇

1.2.29. tétel.
Egy valós mátrix akkor és csak akkor diagonalizálható ortogonális mátrixszal, ha szimmetrikus.

Bizonyítás. Igazoljuk először, hogy ha egy $\mathbf{A}$ valós mátrix ortogonális mátrixszal diagonalizálható, akkor az szimmetrikus. Legyen $\mathbf{S}$ ortogonális és $\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^T$. Ekkor $\mathbf{A}^T = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^T = \mathbf{A}$, azaz $\mathbf{A}$ szimmetrikus.

A másik irány igazolásához tudjuk, hogy a szimmetrikus mátrixok normálisak, ezért diagonalizálhatók. Így van lineárisan független sajátvektorrendszerük. Azt kell megmutatnunk, hogy ezek a vektorok választhatók ortonormáltan. Ha egy sajárértékhez több lineárisan független sajátvektor is tartozik, akkor ezek a Gram–Schmidt eljárással ortonormálhatók. Már csak azt kell megmutatnunk, hogy a különböző sajátértékekhez tartozó sajátvektorok nemcsak függetlenek, hanem ortogonálisak is. Legyen tehát $\overline{\mathbf{v}}_\lambda$ és $\overline{\mathbf{v}}_\mu$ két különböző sajátértékhez (λ és μ) tartozó sajátvektor. Mivel a mátrix szimmetrikus, így a sajátértékei és a sajátvektorai is valósak. A szimmetria miatt a

$$\overline{\mathbf{v}}_\lambda^T\mathbf{A}\overline{\mathbf{v}}_\mu = \overline{\mathbf{v}}_\lambda^T\mu\overline{\mathbf{v}}_\mu = \mu\overline{\mathbf{v}}_\lambda^T\overline{\mathbf{v}}_\mu,$$

$$\overline{\mathbf{v}}_\mu^T\mathbf{A}\overline{\mathbf{v}}_\lambda = \overline{\mathbf{v}}_\mu^T\lambda\overline{\mathbf{v}}_\lambda = \lambda\overline{\mathbf{v}}_\mu^T\overline{\mathbf{v}}_\lambda = \lambda\overline{\mathbf{v}}_\lambda^T\overline{\mathbf{v}}_\mu$$

értékeknek meg kell egyezniük. Ez csak úgy lehet, ha $\overline{\mathbf{v}}_\lambda^T\overline{\mathbf{v}}_\mu = 0$, azaz a különböző sajátértékekhez tartozó sajátvektorok ortogonálisak. Így választható ortonormált sajátvektorrendszer. A mátrix azzal az ortogonális mátrixszal diagonalizálható, melynek oszlopai az ortonormált sajátvektorok. ■

1.2.3. Normák és sajátértékek

Most azt vizsgáljuk meg, hogy milyen kapcsolat van egy mátrix normája és a sajátértékei között. Korábban már találkoztunk egy olyan tétellel (az 1.1.24. tételt akkor bizonyítás nélkül közöltük), ami összeköti a mátrixok normáját a sajátértékeivel. Ez a tétel azt mondta ki, hogy

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

Most már minden eszközünk megvan ezen állítás bizonyításához.

Bizonyítás. (Az 1.1.24. tétel harmadik állításának bizonyítása.) Az $\mathbf{A}^H\mathbf{A}$ mátrix hermitikus és pozitív szemidefinit. Az hermitikusság nyilvánvaló, a pozitív szemidefinitség következik az $\overline{\mathbf{x}}^H\mathbf{A}^H\mathbf{A}\overline{\mathbf{x}} = \|\mathbf{A}\overline{\mathbf{x}}\|_2^2 \geq 0$ egyenlőtlenségből. Az hermitikusság miatt a mátrix diagonalizálható, azaz $\mathbf{A}^H\mathbf{A}$ felírható $\mathbf{A}^H\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$ alakban, ahol $\mathbf{V}$ megfelelő unitér mátrix, $\mathbf{\Lambda}$ pedig a

nemnegatív valós sajátértékeket tartalmazó diagonális mátrix. Így

$$\begin{aligned}\frac{\|\mathbf{A}\overline{\mathbf{x}}\|_2^2}{\|\overline{\mathbf{x}}\|_2^2} &= \frac{\overline{\mathbf{x}}^H\mathbf{A}^H\mathbf{A}\overline{\mathbf{x}}}{\|\overline{\mathbf{x}}\|_2^2} = \frac{\overline{\mathbf{x}}^H\mathbf{V}\mathbf{\Lambda}\mathbf{V}^H\overline{\mathbf{x}}}{\|\overline{\mathbf{x}}\|_2^2} = \frac{\|\sqrt{\mathbf{\Lambda}}\mathbf{V}^H\overline{\mathbf{x}}\|_2^2}{\|\overline{\mathbf{x}}\|_2^2} \\ &\leq \frac{\|\sqrt{\mathbf{\Lambda}}\mathbf{V}^H\|_2^2\|\overline{\mathbf{x}}\|_2^2}{\|\overline{\mathbf{x}}\|_2^2} = \|\sqrt{\mathbf{\Lambda}}\|_2^2 = \varrho(\mathbf{A}^H\mathbf{A}).\end{aligned} \tag{1.2.1}$$

A $\sqrt{\mathbf{\Lambda}}$ mátrix az a diagonális mátrix, melynek főátlóbeli elemei $\mathbf{\Lambda}$ megfelelő elemeinek gyökei. Az $\mathbf{A}^H\mathbf{A}$ mátrix legnagyobb abszolútértékű sajátértékéhez tartozó sajátvektort választva $\overline{\mathbf{x}}$-nek pont egyenlőség van. Így az állítás valóban igaz. ■

A fenti tétel közvetlen következménye az alábbi tétel.

1.2.30. tétel.
Hermitikus (valós szimmetrikus) négyzetes mátrixok esetén $\|\mathbf{A}\|_2 = \varrho(\mathbf{A})$.

Bizonyítás. Mivel $\mathbf{A}$ hermitikus, ezért $\mathbf{A}^H\mathbf{A} = \mathbf{A}^2$, és minden sajátértéke valós. $\mathbf{A}^2$ sajátértékei az eredeti mátrix sajátértékeinek négyzetei, azaz $\mathbf{A}^2$ spektrálsugara megegyezik $\mathbf{A}$ spektrálsugarának négyzetével. Ebből következik az állítás. ■

1.2.31. tétel.
Négyzetes mátrixokra indukált normák esetén érvényes a $\varrho(\mathbf{A}) \leq \|\mathbf{A}\|$ becslés.

Bizonyítás. Legyen $\overline{\mathbf{x}} \neq \mathbf{0}$ egy sajátvektora $\mathbf{A}$-nak, és λ a hozzá tartozó sajátérték. Ekkor $|\lambda| \cdot \|\overline{\mathbf{x}}\| = \|\lambda\overline{\mathbf{x}}\| = \|\mathbf{A}\overline{\mathbf{x}}\| \leq \|\mathbf{A}\| \cdot \|\overline{\mathbf{x}}\|$, amiből az állítás következik. ■

1.2.32. tétel.
Adott $\mathbf{A} \in \mathbb{C}^{n\times n}$ mátrix esetén minden $\varepsilon > 0$ számhoz létezik olyan $\|\cdot\|$ indukált norma, mellyel $\|\mathbf{A}\| \leq \varrho(\mathbf{A}) + \varepsilon$.

Bizonyítás. Induljunk ki az $\mathbf{A}$ mátrix Schur-felbontásából. Eszerint $\mathbf{A}$ felírható $\mathbf{A} = \mathbf{STS}^H$ alakban, ahol $\mathbf{S}$ unitér mátrix, $\mathbf{T}$ pedig olyan felső háromszögmátrix, melynek diagonálisában az $\mathbf{A}$ mátrix $\lambda_1, \ldots, \lambda_n$ sajátértékei szerepelnek. Legyen $d > 0$ egy később megválasztandó értékű paraméter. Definiáljuk a $\mathbf{D} = \operatorname{diag}(d, d^2, \ldots, d^n)$ diagonális mátrixot. Ekkor a $\mathbf{DTD}^{-1}$ továbbra is felső háromszögmátrix, melynek diagonálisa megegyezik $\mathbf{T}$ diagonálisával. Számoljuk ki $\mathbf{DTD}^{-1}$ 1-es normáját. A j-edik oszlop abszolútérték-összege:

$$|\lambda_j| + \sum_{i=1}^{j-1} \frac{1}{d^j} d^i |t_{ij}| = |\lambda_j| + \sum_{i=1}^{j-1} d^{-(j-i)} |t_{ij}|.$$

Mivel d kitevője negatív, így ha d-t megfelelően nagynak választjuk, akkor elérhető, hogy

$$|\lambda_j| + \sum_{i=1}^{j-1} d^{-(j-i)} |t_{ij}| \leq \varrho(\mathbf{A}) + \varepsilon$$

teljesüljön minden $j = 1, \ldots, n$ oszlop esetén. Így tehát

$$\|\mathbf{DTD}^{-1}\|_1 \leq \varrho(\mathbf{A}) + \varepsilon. \tag{1.2.2}$$

Már csak azt kell megmutatnunk, hogy a bal oldalon álló norma az $\mathbf{A}$ mátrixnak valamilyen indukált mátrixnormája.

Definiáljuk az $\mathbf{S}$ és $\mathbf{D}$ mátrixok segítségével az $\|\overline{\mathbf{x}}\|_{\mathbf{S},\mathbf{D}} = \|(\mathbf{S}\mathbf{D}^{-1})^{-1}\overline{\mathbf{x}}\|_1$ vektornormát. Annak igazolását, hogy ez valóban vektornorma, az Olvasóra bízzuk (1.6.8. feladat). Határozzuk meg, hogy milyen mátrixnormát indukál ez a vektornorma! Legyen $\mathbf{X}$ tetszőleges $(n \times n)$-es mátrix.

$$\begin{aligned}
\|\mathbf{X}\|_{\mathbf{S},\mathbf{D}} &= \sup_{\overline{\mathbf{x}}\neq\mathbf{0}} \frac{\|\mathbf{X}\overline{\mathbf{x}}\|_{\mathbf{S},\mathbf{D}}}{\|\overline{\mathbf{x}}\|_{\mathbf{S},\mathbf{D}}} = \sup_{\overline{\mathbf{x}}\neq\mathbf{0}} \frac{\|(\mathbf{S}\mathbf{D}^{-1})^{-1}\mathbf{X}\overline{\mathbf{x}}\|_1}{\|(\mathbf{S}\mathbf{D}^{-1})^{-1}\overline{\mathbf{x}}\|_1} \\
&= \sup_{\overline{\mathbf{x}}\neq\mathbf{0}} \frac{\|(\mathbf{S}\mathbf{D}^{-1})^{-1}\mathbf{X}\mathbf{S}\mathbf{D}^{-1}(\mathbf{S}\mathbf{D}^{-1})^{-1}\overline{\mathbf{x}}\|_1}{\|(\mathbf{S}\mathbf{D}^{-1})^{-1}\overline{\mathbf{x}}\|_1} \\
&= \sup_{\overline{\mathbf{y}}:=(\mathbf{S}\mathbf{D}^{-1})^{-1}\overline{\mathbf{x}}\neq\mathbf{0}} \frac{\|(\mathbf{S}\mathbf{D}^{-1})^{-1}\mathbf{X}\mathbf{S}\mathbf{D}^{-1}\overline{\mathbf{y}}\|_1}{\|\overline{\mathbf{y}}\|_1} \\
&= \|(\mathbf{S}\mathbf{D}^{-1})^{-1}\mathbf{X}\mathbf{S}\mathbf{D}^{-1}\|_1.
\end{aligned}$$

Végül számítsuk ki az $\mathbf{A}$ mátrix normáját a fenti indukált mátrixnormában, és alkalmazzuk az (1.2.2) becslést.

$$\|\mathbf{A}\|_{\mathbf{S},\mathbf{D}} = \|(\mathbf{S}\mathbf{D}^{-1})^{-1}\mathbf{A}\mathbf{S}\mathbf{D}^{-1}\|_1 = \|\mathbf{D}\mathbf{S}^{-1}\mathbf{A}\mathbf{S}\mathbf{D}^{-1}\|_1 = \|\mathbf{D}\mathbf{T}\mathbf{D}^{-1}\|_1 \leq \varrho(\mathbf{A}) + \varepsilon.$$

Így tehát az $\|\cdot\|_{\mathbf{S},\mathbf{D}}$ indukált mátrixnormában az $\mathbf{A}$ mátrix normája valóban kisebb, mint $\varrho(\mathbf{A})+\varepsilon$. ■

1.2.33. következmény. A tétel következménye, hogy ha egy mátrix spektrálsugara 1-nél kisebb, akkor van olyan indukált mátrixnorma, melyben a mátrix normája is egynél kisebb. ◇

1.2.34. tétel.
Egy $\mathbf{A} \in \mathbb{C}^{n\times n}$ mátrix esetén pontosan akkor igaz, hogy $\mathbf{A}^k \to \mathbf{0}$ elemenként, ha $\varrho(\mathbf{A}) < 1$. Pontosan ugyanekkor lesz a

$$\sum_{k=0}^{\infty} \mathbf{A}^k$$

sor konvergens, és összege az $(\mathbf{E} - \mathbf{A})^{-1}$ mátrix.

Bizonyítás. Először igazoljuk azt az irányt, hogy a $\varrho(\mathbf{A}) < 1$ feltételből következik a másik két állítás. Mivel $\varrho(\mathbf{A}) < 1$, ezért van olyan $\|\cdot\|$ indukált mátrixnorma, mellyel $\|\mathbf{A}\| < 1$ (1.2.32. tétel). Emiatt $\|\mathbf{A}^k\| \leq \|\mathbf{A}\|^k \to 0$, ha $k \to \infty$. A normák ekvivalenciája miatt (1.1.14. tétel) ekkor $\|\mathbf{A}^k\|_\infty \to 0$ is igaz, ami azt jelenti hogy $\mathbf{A}^k$ elemenként is nullához tart.

A másik irányhoz legyen most $\overline{\mathbf{v}}$ a mátrix egy sajátvektora λ sajátértékkel. Ekkor $\mathbf{A}^k\overline{\mathbf{v}} = \lambda^k\overline{\mathbf{v}}$. Mivel $\mathbf{A}^k$ nullmátrixhoz tart, így ennek a vektornak nullvektorhoz kellene tartania. Ez csak akkor lehet, ha $|\lambda| < 1$. Így $\varrho(\mathbf{A}) < 1$ következik.

Tekintsük az alábbi azonosságot tetszőleges l természetes szám esetén:

$$(\mathbf{E} - \mathbf{A})(\mathbf{E} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^l) = \mathbf{E} - \mathbf{A}^{l+1}.$$

Az $\mathbf{E} - \mathbf{A}$ mátrix reguláris, mert sajátértékei nem lehetnek nullák. Így

$$\mathbf{E} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^l = (\mathbf{E} - \mathbf{A})^{-1}(\mathbf{E} - \mathbf{A}^{l+1}).$$

Így pontosan akkor van a sornak határértéke $l \to \infty$ esetén, ha az $\{\mathbf{A}^{l+1}\}$ mátrixsorozat nullához tart, azaz ha $\varrho(\mathbf{A}) < 1$, és ilyenkor az összeg valóban $(\mathbf{E} - \mathbf{A})^{-1}$. ■

1.2.35. megjegyzés. Az $\mathbf{A}^k \to \mathbf{0}$ $(k \to \infty)$ tulajdonsággal rendelkező $\mathbf{A}$ mátrixokat szokás konvergens mátrixoknak is nevezni. ◇

1.2.4. M-mátrixok

1.2.36. definíció.
Az olyan $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrixokat, melyek főátlón kívüli elemei nempozitívak, nemszingulárisak és inverzük nemnegatív, M-mátrixoknak nevezzük.

Az M-mátrixok elnevezésében az M betű a *m*onoton mátrix kezdőbetűjére utal. Ugyanis ha egy $\mathbf{A}$ mátrix M-mátrix, akkor az $\mathbf{A}\overline{\mathbf{x}} \geq \mathbf{A}\overline{\mathbf{y}}$ egyenlőségből következik az $\overline{\mathbf{x}} \geq \overline{\mathbf{y}}$ egyenlőség. Differenciálegyenletek numerikus megoldása során gyakran olyan egyenletrendszerekhez jutunk, melyek együtthatómátrixa M-mátrix.

1.2.37. példa. Az

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

mátrix például M-mátrix, hiszen a főátlón kívül nincs pozitív eleme, és inverze

$$\mathbf{A}^{-1} = \begin{bmatrix} 3/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 3/4 \end{bmatrix}.$$

◇

1.2.38. tétel.
Egy M-mátrix főátlója pozitív elemeket tartalmaz.

Bizonyítás. Jelöljük $\mathbf{A}$-val a mátrixot. Ha $a_{ii} \leq 0$ lenne valamilyen i indexre, akkor $\mathbf{A}\overline{\mathbf{e}}_i \leq \mathbf{0}$ lenne. De ekkor az egyenlőtlenséget az $\mathbf{A}^{-1} \geq \mathbf{0}$ mátrixszal szorozva azt kapjuk, hogy $\mathbf{A}^{-1}\mathbf{A}\overline{\mathbf{e}}_i = \overline{\mathbf{e}}_i \leq \mathbf{0}$, ami ellentmondás. ■

Az M-mátrix definíciójában szereplő feltételek nehezen ellenőrizhetők, hiszen ismerni kell hozzá a mátrix inverzét. A következő tétel egy könnyebben ellenőrizhető szükséges és elégséges feltételt ad.

1.2.39. tétel.
Legyen az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix olyan, hogy a főátlóján kívüli elemek nempozitívak. Ekkor $\mathbf{A}$ pontosan akkor M-mátrix, ha van olyan $\overline{\mathbf{g}} > \mathbf{0}$ vektor, mellyel $\mathbf{A}\overline{\mathbf{g}} > \mathbf{0}$.

Bizonyítás. Igazoljuk először a feltétel szükségességét. Legyen $\overline{\mathbf{e}} = [1, \ldots, 1]^T$. Ekkor $\overline{\mathbf{g}} = \mathbf{A}^{-1}\overline{\mathbf{e}}$ megfelelő választás, ugyanis az $\mathbf{A}^{-1} \geq \mathbf{0}$ mátrixnak nem lehet nulla sorösszege, azaz $\overline{\mathbf{g}} > \mathbf{0}$. Erre a $\overline{\mathbf{g}}$ vektorra igaz továbbá, hogy $\mathbf{A}\overline{\mathbf{g}} = \mathbf{A}\mathbf{A}^{-1}\overline{\mathbf{e}} = \overline{\mathbf{e}} > \mathbf{0}$, amit igazolni akartunk.

A feltétel elégségességének igazolásához legyen

$$\mathbf{G} = \operatorname{diag}(g_1, \ldots, g_n)$$

és

$$\mathbf{D} = \operatorname{diag}(a_{11}g_1, \ldots, a_{nn}g_n).$$

Nyilvánvalóan $\mathbf{A}$ minden főátlóbeli eleme pozitív, különben az $\mathbf{A}\overline{\mathbf{g}} > \mathbf{0}$ feltétel nem teljesülhetne pozitív $\overline{\mathbf{g}}$ vektorral. Így $\mathbf{D}$ minden főátlóbeli eleme is pozitív, ezért invertálható. Ekkor $\mathbf{D}^{-1}\mathbf{AG}$ továbbra is olyan mátrix, melynek nincs a főátlóján kívül pozitív eleme, továbbá a főátlójában egyesek állnak. Így felírható $\mathbf{D}^{-1}\mathbf{AG} = \mathbf{E} - \mathbf{B}$ alakban, ahol $\mathbf{B}$ egy nemnegatív mátrix nulla főátlóval. Mivel $\mathbf{D}^{-1}\mathbf{AG}\overline{\mathbf{e}} = \mathbf{D}^{-1}\mathbf{A}\overline{\mathbf{g}} > \mathbf{0}$, emiatt $(\mathbf{E} - \mathbf{B})\overline{\mathbf{e}} > \mathbf{0}$. Ez mutatja, hogy $\mathbf{B}$ maximumnormája kisebb 1-nél. Tehát a spektrálsugara is kisebb 1-nél. Így $\mathbf{E} - \mathbf{B}$ invertálható és (1.2.34. tétel) $\mathbf{0} \leq \mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \cdots = (\mathbf{E} - \mathbf{B})^{-1}$. Tehát $\mathbf{A}$-nak is van inverze, nevezetesen $\mathbf{A}^{-1} = \mathbf{G}(\mathbf{E} - \mathbf{B})^{-1}\mathbf{D}^{-1}$, és az nemnegatív, mert a jobb oldal minden mátrixa nemnegatív. ■

Az előző tétel szerint, ha egy $\mathbf{A}$ mátrixhoz, melynek nincsenek pozitív elemei a főátlóján kívül, sikerül olyan $\overline{\mathbf{g}} > \mathbf{0}$ vektort találnunk, melyre $\mathbf{A}\overline{\mathbf{g}} > \mathbf{0}$, akkor $\mathbf{A}$ M-mátrix. A következő tétel azt mutatja, hogy a $\overline{\mathbf{g}}$ és $\mathbf{A}\overline{\mathbf{g}}$ vektorok segítségével felső becslést adhatunk az $\mathbf{A}^{-1}$ mátrix maximumnormájára.

1.2.40. tétel.
Legyen $\mathbf{A}$ M-mátrix, és $\overline{\mathbf{g}} > \mathbf{0}$ egy olyan vektor, melyre $\mathbf{A}\overline{\mathbf{g}} > \mathbf{0}$ teljesül. Ekkor

$$\|\mathbf{A}^{-1}\|_\infty \leq \frac{\|\overline{\mathbf{g}}\|_\infty}{\min_i(\mathbf{A}\overline{\mathbf{g}})_i}.$$

Bizonyítás. Az állítás az alábbi becslésekből következik, figyelembe véve az $\mathbf{A}^{-1}$ mátrix nemnegativitását.

$$\|\mathbf{A}^{-1}\|_\infty \min_{i=1,\ldots,n}(\mathbf{A}\overline{\mathbf{g}})_i = \|\mathbf{A}^{-1}\overline{\mathbf{e}} \min_{i=1,\ldots,n}(\mathbf{A}\overline{\mathbf{g}})_i\|_\infty \leq \|\mathbf{A}^{-1}\mathbf{A}\overline{\mathbf{g}}\|_\infty = \|\overline{\mathbf{g}}\|_\infty. \blacksquare$$

1.2.41. példa. Az 1.2.37. példában szereplő

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

mátrix esetén $\overline{\mathbf{g}} = [2, 3, 2]^T > \mathbf{0}$ mellett $\mathbf{A}\overline{\mathbf{g}} = [1, 2, 1]^T > \mathbf{0}$. Így az inverz maximumnormájára egy felső becslés

$$\|\mathbf{A}^{-1}\|_\infty \leq \frac{\|\overline{\mathbf{g}}\|_\infty}{\min_i(\mathbf{A}\overline{\mathbf{g}})_i} = \frac{3}{1} = 3,$$

ami tényleg teljesül, hiszen $\|\mathbf{A}^{-1}\|_\infty = 2$. ◇

sorszám	első sorozat	második sorozat
0.	2.00000000000000	2.00000000000000000000000000000000000000000000000000
1.	1.80000000000000	1.50000000000000000000000000000000000000000000000000
2.	1.67600000000000	1.41666666666666666666666666666666666666666666666667
3.	1.59510240000000	1.41421568627450980392156862745098039215686274509 81
4.	1.54066723335142	1.41421356237468991062629557889013491011655962211 57
5.	1.50330168095915	1.41421356237309504880168962350253024361498192577 62
6.	1.47731008656169	1.41421356237309504880168872420969807856967187537 72
7.	1.45906557737600	1.41421356237309504880168872420969807856967187537 70
8.	1.44617834146764	1.41421356237309504880168872420969807856967187537 70
9.	1.43703516193463	1.41421356237309504880168872420969807856967187537 70
10.	1.43052815627098	1.41421356237309504880168872420969807856967187537 70
11.	1.42588707568258	1.41421356237309504880168872420969807856967187537 70
12.	1.42257168042272	1.41421356237309504880168872420969807856967187537 70
13.	1.42020066182865	1.41421356237309504880168872420969807856967187537 70
14.	1.41850366984279	1.41421356237309504880168872420969807856967187537 70
15.	1.41728840370705	1.41421356237309504880168872420969807856967187537 70
$\sqrt{2} = 1.41421356237309504880168872420969807856967187537 70\ldots$		

1.3.1. táblázat: Két $\sqrt{2}$-höz tartó, de más konvergenciasebességű, sorozat első 16 eleme. Az első elsőrendű, a második másodrendű.

1.3. Sorozatok és függvények konvergenciájának jellemzése

1.3.1. Sorozatok konvergenciasebessége

A numerikus módszerek egy nagy csoportja ún. iterációs módszer. Ez azt jelenti, hogy a feladat megoldása egy rekurzív eljárással előállított sorozat határértéke lesz. A megoldást úgy közelítjük, hogy az iteráció segítségével előállítjuk a sorozat egy megfelelően nagy sorszámú elemét. Hogy hányadik elemet kell előállítanunk, az azon múlik, hogy a sorozat milyen "gyorsan" konvergál a határértékéhez. Nyilvánvalóan egy gyorsan konvergáló sorozat sokkal kedvezőbb a számunkra, mint egy lassan konvergáló. Most megvizsgáljuk, hogy hogyan jellemezhető egy sorozat konvergenciasebessége.

Az 1.3.1. táblázat két $\sqrt{2}$-höz tartó sorozat első 15 elemét tartalmazza 14 ill. 49 tizedesjegyre kerekítve. Mindkét sorozatot egy-egy iterációs eljárással állítottuk elő. Látható, hogy a második sorozat sokkal "gyorsabban" konvergál. A hetedik elemtől kezdve nem is látszik változás, pedig 49 tizedesjegyre kerekítettünk. Az első sorozat pedig még a 15. lépés után is három ezrednyire van a határértéktől. A sorozatok elemeinek határértéktől való távolságát az 1.3.1. ábrán ábrázoltuk. Jelöljük egy adott konvergens sorozat elemeit $x^{(k)}$-val, és határértékét $x^\star$-gal. Jelölje $e^{(k)} := x^{(k)} - x^\star$ $(k = 0, 1, \ldots)$ a k-adik sorozatelem hibáját. Vizsgáljuk meg, hogy a fent vizsgált két sorozat esetében egy lépés során hogyan változik ez a hiba.

Az

$$\frac{\ln|e^{(k)}|}{\ln|e^{(k-1)}|}$$

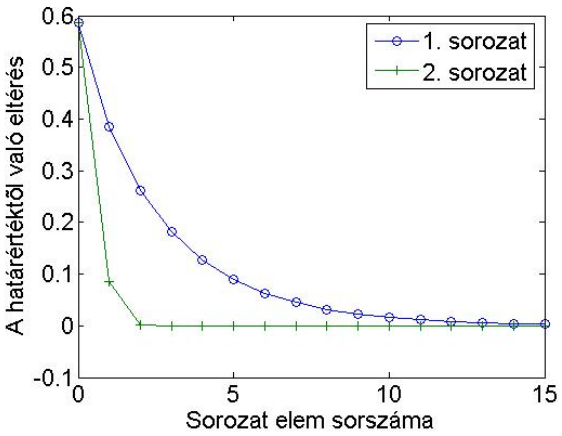
értéket (ha $e^{(k-1)} \neq 0, 1$) a k. lépéshez tartozó logaritmikus relatív csökkenésnek nevezzük. Mivel a határérték közelében a hibaértékek abszolútértékben kicsi számok, így az a hibasorozat csökken

gyorsabban, amelyre ez a hányados nagyobb. Az 1.3.2. ábrán a vizsgált két sorozatra ábrázoltuk a logaritmikus relatív csökkenést.

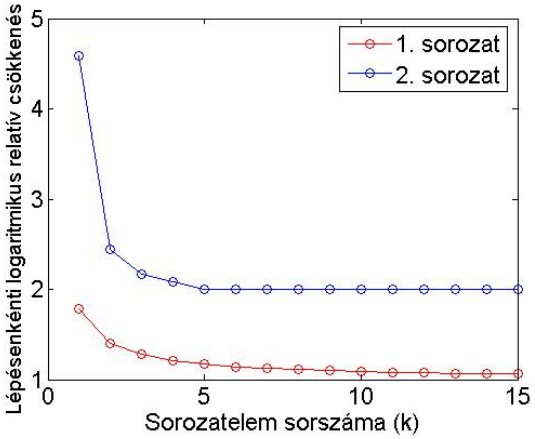
Az első sorozat esetén a logaritmikus relatív csökkenés 1 közelében van, míg a másodiknál 2 közelében. Ezek a számok tehát alkalmasak lehetnek a konvergencia jellemzésére.

Most vizsgáljuk meg általánosan a konvergenciasebesség kérdését normált térbeli sorozatokra! Vizsgáljunk olyan sorozatokat, melyek monoton módon tartanak a határértékhez (azaz egyik lépésben sem növekedhet a hiba abszolút értéke), és egyik sorozatelem sem egyezik meg a határértékkel (a fenti két példában ilyen sorozatokat mutattunk, és a gyakorlatban is tipikusan ilyen sorozatokat adnak az iterációs eljárások)!

Legyen tehát $\{x^{(k)}\}$ egy tetszőleges normált térbeli, $x^\star$-hoz monoton módon tartó konvergens sorozat. A k. sorozatelem hibáját jelölje $e^{(k)} := x^{(k)} - x^\star$ (a monotonitás miatt tehát $\|e^{(k)}\| \leq \|e^{(k-1)}\|$).



1.3.1. ábra: A vizsgált két sorozat elemeinek a határértéktől való távolsága.



1.3.2. ábra: A vizsgált két sorozat hibájának logaritmikus relatív csökkenése.

1.3.1. definíció.
Azt mondjuk, hogy az $\{x^{(k)}\}$ $x^\star$-hoz monoton módon konvergáló sorozat konvergenciarendje pontosan $p \geq 1$, ha a

$$\lim_{k\to\infty} \frac{\ln \|e^{(k)}\|}{\ln \|e^{(k-1)}\|}$$

véges határérték létezik és értéke p.

1.3.2. megjegyzés. A fenti definíció alapján mondhatjuk, hogy a bevezető feladatban az első sorozat elsőrendben, a második másodrendben konvergens. ⋄

Ha $p = 1$, akkor lineáris konvergenciáról, ha $1 < p < 2$, akkor szuperlineáris konvergenciáról, $p = 2$ esetén pedig másodrendű konvergenciáról beszélünk.

Hogyan lehet a konvergenciarendet meghatározni? Vizsgáljunk meg két speciális esetet!

1.3.3. tétel.
Ha egy $\{x_k\}$ sorozatra és egy $x^\star$ elemre az igaz, hogy

$$\|e^{(k)}\| = C_k \|e^{(k-1)}\|$$

valamilyen $0 < \underline{C} \leq C_k \leq \overline{C} < 1$ konstansokkal, akkor $x_k \to x^\star$ monoton módon és elsőrendben.

Bizonyítás. Az $\|e^{(k)}\| = C_k\|e^{(k-1)}\| \leq \overline{C}\|e^{(k-1)}\|$ becslés miatt $\|e^{(k)}\| \leq \overline{C}^k\|e^{(0)}\|$. Mivel $\overline{C} < 1$, ezekből következik, hogy a sorozat monoton módon fog az $x^\star$ elemhez tartani. Az $\|e^{(k)}\| = C_k\|e^{(k-1)}\|$ egyenlőség logaritmusát véve, majd osztva az $\ln\|e^{(k-1)}\| < 0$ értékkel (feltéve, hogy k elég nagy ahhoz, hogy a hiba már kisebb legyen, mint 1), azt kapjuk, hogy a logaritmikus relatív csökkenés

$$\frac{\ln \|e^{(k)}\|}{\ln \|e^{(k-1)}\|} = \frac{\ln C_k}{\ln \|e^{(k-1)}\|} + 1 \to 1$$

a C_k konstansokra vonatkozó $0 < \underline{C} \leq C_k \leq \overline{C} < 1$ feltétel miatt és amiatt, mert $\ln\|e^{(k-1)}\| \to -\infty$. Azaz a konvergenciarend valóban 1. ■

1.3.4. megjegyzés. Az $\|e^{(k)}\| \leq \overline{C}\|e^{(k-1)}\|$ becslésből az is látszik, hogy $M \approx -\ln 10/\ln \overline{C}$ lépésszám után csökken egy nagyságrendet a sorozat elemeinek határértéktől mért távolsága, hiszen $\overline{C}^M = 1/10$. Ez a gyakorlatban azt jelenti, hogy ha pl. $\overline{C} = 1/2$, akkor $M = -\ln 10/\ln(1/2) \approx 3.3219$, azaz kicsivel több, mint három lépésenként számíthatunk egy nagyságrendnyi hibacsökkenésre. ⋄

1.3.5. tétel.
Ha egy $\{x^{(k)}\}$ sorozatra és egy $x^\star$ elemre az igaz, hogy

$$\|e^{(k)}\| = C_k \|e^{(k-1)}\|^p \tag{1.3.1}$$

valamilyen $0 < \underline{C} \leq C_k \leq \overline{C} < \infty$ és $p > 1$ konstansokkal, és $\overline{C}^{1/(p-1)}\|e^{(0)}\| < 1$, akkor $x_k \to x^\star$ monoton módon, és a sorozat konvergenciarendje p.

Bizonyítás. Vezessük be az $\varepsilon^{(k)} = \overline{C}^{1/(p-1)} e^{(k)}$ jelölést. Ezzel a jelöléssel

$$\|\varepsilon^{(k)}\| = \overline{C}^{1/(p-1)} \|e^{(k)}\| \leq \overline{C}^{1/(p-1)} \overline{C} \|e^{(k-1)}\|^p = (\overline{C}^{1/(p-1)} \|e^{(k-1)}\|)^p = \|\varepsilon^{(k-1)}\|^p.$$

Tehát

$$\|\varepsilon^{(k)}\| \leq \|\varepsilon^{(k-1)}\|^p, \tag{1.3.2}$$

azaz

$$\|\varepsilon^{(k)}\| \leq \|\varepsilon^{(0)}\|^{p^k}.$$

Az $\|\varepsilon^{(0)}\| = \overline{C}^{1/(p-1)} \|e^{(0)}\| < 1$ feltétel miatt a fenti egyenlőségből következik, hogy $\|\varepsilon^{(k)}\| \to 0$ $(k \to \infty)$ monoton módon. Mivel

$$e^{(k)} = \frac{\varepsilon^{(k)}}{\overline{C}^{1/(p-1)}},$$

ezért $\|e^{(k)}\| \to 0$, azaz a sorozat valóban $x^\star$-hoz tart monoton módon.

A konvergenciarend igazolásához vegyük az (1.3.1) egyenlőség logaritmusát, majd osszunk $\ln \|e^{(k-1)}\|$-val. Ekkor

$$\frac{\ln \|e^{(k)}\|}{\ln \|e^{(k-1)}\|} = \frac{\ln C_k}{\ln \|e^{(k-1)}\|} + p \to p,$$

mivel $0 < \underline{C} \leq C_k \leq \overline{C}$ minden $k = 0, 1, \ldots$ esetén és $\ln \|e^{(k-1)}\| \to -\infty$, ha $k \to \infty$. Ezt akartuk megmutatni. ■

1.3.6. megjegyzés. Az (1.3.2) becslésből látható, hogy a közelítés pontosságának nagyságrendje minden lépésben kb. p-szereződik. Pl. ha $p = 2$ és egy adott közelítés hibája 10^{-3}, akkor a következő közelítésé már kb. 10^{-6}-os, a rákövetkezőé pedig kb. 10^{-12}-es lesz. Ez a lineáris konvergenciával összevetve nagyon gyors konvergenciát jelent. ⋄

1.3.7. megjegyzés. Könnyen látható, hogy az 1.3.5. tételben szereplő $\overline{C}^{1/(p-1)} \|e^{(0)}\| < 1$ feltétel azt jelenti, hogy az (1.3.1) egyenlőségből csak akkor következik a konvergencia, ha a sorozat nulladik eleme elegendően közel van a határértékhez. Az elsőrendű konvergenciához az 1.3.3. tételben nem kellett ezt a feltételt garantálni. ⋄

1.3.2. Függvények konvergenciavizsgálata

Térjünk át a függvények konvergenciavizsgálatára. Numerikus szempontból azok a valós-valós nemnegatív függvények érdekesek, melyek nullában nullához ill. végtelenben végtelenhez tartanak. Most ezek jellemzésével fogunk foglalkozni.

Jelentsen α 0-t vagy ∞-t. Tegyük fel, hogy $g : \mathbb{R} \to \mathbb{R}$ és $f : \mathbb{R} \to \mathbb{R}$ olyan függvények, melyek értelmezési tartományai metszetének α torlódási pontja, és mindkét függvény α-hoz tart α-ban. A két függvény konvergenciájának kapcsolatát fejezi ki az alább definiált ordó[6] jelölés.

[6] Az ordó jelölés Edmund Landau (Edmund Georg Hermann Landau (1877 Berlin – 1938) német matematikus nevéhez fűződik. Bővebb életrajz: http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Landau.html

1.3.8. definíció.
Ha azt írjuk, hogy $g(x) = \mathcal{O}(f(x))$ $(x \to \alpha)$ (ejtsd: "ordó ef"), akkor ezen a g és f függvények alábbi viszonyát értjük: Vannak olyan $\varepsilon > 0$ és $C > 0$ konstansok, mellyekkel $|g(x)| \leq C|f(x)|$ minden olyan közös értelmezési tartománybeli elemre, melyek α ε sugarú környezetébe esnek. A nulla ε sugarú környezetén a szokott módon a $(-\varepsilon, \varepsilon)$ intervallumot, a végtelen ε sugarú környezetén az $(1/\varepsilon, \infty)$ intervallumot értjük.

Ha a szövegösszefüggésből világos, hogy mely α pontban nézzük a határértéket, akkor ennek jelölését el szoktuk hagyni.

A definíció alapján írhatjuk a $g(x) = 2x^2 - 4x + 2$ függvény esetén, hogy $g(x) = \mathcal{O}(x^2)$ $(x \to \infty)$, hiszen $C = 2$ és $\varepsilon = 1$ megfelelő választás. Természetesen $g(x) = \mathcal{O}(x^3)$ is igaz, de a gyakorlatban törekszünk a legkisebb lehetséges hatványkitevő megadására. Nyilvánvalóan $g(x) \neq \mathcal{O}(x)$.

A $g(x) = 4x - 2x^2$ függvényről írhatjuk, hogy $g(x) = \mathcal{O}(x)$ $(x \to 0)$, hiszen $C = 4$ és $\varepsilon = 1$ megfelelő választás. Ha $-1 < x < 1$, akkor $4x - 2x^2 \leq 4x$. Nyilvánvalóan $g(x) = \mathcal{O}(1)$ is igaz, de a gyakorlatban törekszünk a legnagyobb lehetséges x-hatvány megadására. Nyilvánvalóan $g(x) \neq \mathcal{O}(x^2)$

1.3.9. példa. Az ordó jelölés alkalmazására további példaként tekintsük az e^x exponenciális függvény Taylor-sorfejtéssel való közelítését az $x_0 = 0$ pontban. Ha a másodrendű Taylor-polinommal közelítünk, akkor

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{e^\xi}{3!}x^3,$$

ahol az utolsó ún. Lagrange-féle maradéktagban ξ megfelelő x és 0 közé eső (x-től függő) konstans. Mivel

$$\frac{e^\xi}{3!}x^3 \leq \frac{e}{3!}x^3,$$

ha $0 \leq x \leq 1$, ezért a maradéktag helyett írhatjuk, hogy $\mathcal{O}(x^3)$. Tehát a Taylor-polinommal való közelítés

$$e^x = 1 + x + \frac{x^2}{2!} + \mathcal{O}(x^3)$$

alakú lesz. Hasonló felírás minden olyan esetben megtehető, amikor a Lagrange-féle maradéktagban szereplő derivált korlátos a nulla egy környezetében. ⋄

Az ordó jelölést gyakran alkalmazzuk olyan esetekben, amikor egy közelítés hibáját adjuk meg egy paraméter függvényében.

1.3.10. definíció.
Azt mondjuk, hogy a $v(h)$ h pozitív valós paramétertől függő közelítése egy $v \in (V, \|\cdot\|)$ elemnek (legalább) $r \geq 1$-edrendű közelítés, ha $\|v(h) - v\| = \mathcal{O}(h^r)$ $(h \to 0)$.

Az fenti definícióból és az ordó jelölés definíciójából következik, hogy ha $\|v(h) - v\| = \mathcal{O}(h^r)$, akkor van olyan $K > 0$ konstans, hogy $\|v(h) - v\| \leq Kh^r$ minden elegendően kis abszolútértékű h paraméterre. A definícióból az is következik, hogy ha $v(h)$ legalább elsőrendű közelítés, akkor $\lim_{h\to 0} v(h) = v$, másrészt ha pl. felezzük a h paraméter értékét, akkor kb. 2^r-ed részére csökken a $\|v(h) - v\|$ hiba.

1.3.11. példa. Közelítsük egy kétszer folytonosan differenciálható függvény esetén a derivált értékét egy x_0 pontban az

$$\frac{f(x_0+h)-f(x_0)}{h}$$

hányadossal. A Taylor-tétel miatt

$$f(x_0+h)=f(x_0)+hf'(x_0)+\frac{h^2}{2}f''(\xi)$$

alakban írható, ahol ξ megfelelő x_0 és x_0+h közé eső konstans. Tehát

$$\frac{f(x_0+h)-f(x_0)}{h}=\frac{f(x_0)+hf'(x_0)+h^2f''(\xi)/2-f(x_0)}{h}=f'(x_0)+hf''(\xi)/2.$$

Figyelembe véve, hogy f kétszer folytonosan differenciálható, azaz f'' véges, zárt intervallumon korlátos, az ordó jelölést használva írhatjuk, hogy

$$\frac{f(x_0+h)-f(x_0)}{h}=f'(x_0)+\mathcal{O}(h).$$

Tehát az adott hányados elsőrendű közelítése az első deriváltnak az x_0 pontban. ⋄

A definíció alapján annak igazolásához, hogy egy közelítés r-edrendű elég megmutatnunk, hogy $\|v(h)-v\|=\mathcal{O}(h^r)$. Gyakran azonban pontosan is ismerjük az $\mathcal{O}(h^r)$ hibatagot (az 1.3.11. példában pl. $hf''(\xi)/2$), ami lehetőséget ad magasabbrendű közelítések megadására. Ezt az eljárást *Richardson-extrapolációnak* nevezzük. Általánosan a Richardson-extrapoláció az alábbi módon működik. Tegyük fel, hogy $v(h)$ v-nek p-edrendű közelítése, és a hibát felírhatjuk $v(h)-v=g(h)h^r$ alakban, ahol $g:\mathbb{R}\to(V,\|\cdot\|)$ h-nak folytonosan differenciálható függvénye. Ekkor, ha $h/2$ értékkel is kiszámítjuk a közelítést, akkor $v(h/2)-v=g(h/2)h^r/2^r$ adódik. A két közelítést ezek után a

$$\begin{aligned}\frac{2^rv(h/2)-v(h)}{2^r-1}&=v+(g(h/2)-g(h))\frac{h^r}{2^r-1}\\&=v-\frac{g(h/2)-g(h)}{-h/2}\frac{h^{r+1}}{2(2^r-1)}=v+\mathcal{O}(h^{p+1})\end{aligned}\tag{1.3.3}$$

módon súlyozva eggyel magasabbrendű közelítését kapjuk a v elemnek.

1.3.12. példa. Tekintsük az előző példánkat, melyben a deriváltat az

$$\frac{f(x_0+h)-f(x_0)}{h}=f'(x_0)+\mathcal{O}(h)$$

módon közelítettük. Ez a felírás természetesen mutatja az elsőrendű konvergenciát, de amint láttuk, az $\mathcal{O}(h)$-val jelölt hiba értéke pontosan is ismert, nevezetesen a hiba $hf''(\xi)/2$, azaz

$$\frac{f(x_0+h)-f(x_0)}{h}=f'(x_0)+hf''(\xi)/2.\tag{1.3.4}$$

Írjuk fel ezt a közelítést felezve a h paramétert

$$\frac{f(x_0+h/2)-f(x_0)}{h/2}=f'(x_0)+hf''(\tilde{\xi})/4,\tag{1.3.5}$$

ahol most $\tilde{\xi}$ x_0 és $x_0 + h/2$ közé eső megfelelő konstans. Mivel f'' folytonos függvény, ezért ha h elegendően kicsi, akkor $f''(\xi)$ és $f''(\tilde{\xi})$ is elegendően közel lesz egymáshoz (mondhatjuk ezt annak ellenére, hogy ξ és $\tilde{\xi}$ értéke nem ismert) és $f''(x_0)$-hoz is. A h első hatványait tartalmazó tagok kiejtése érdekében vonjuk ki az (1.3.5) egyenlet kétszereséből az (1.3.4) egyenletet. Ekkor ha f háromszor is folytonosan deriválható, akkor azt kapjuk, hogy

$$\begin{aligned}
&\frac{2(f(x_0+h/2)-f(x_0))}{h/2} - \frac{f(x_0+h)-f(x_0)}{h} \\
=&f'(x_0) + h(f''(\tilde{\xi}) - f''(\xi))/2 = f'(x_0) + \frac{h}{2}\frac{f''(\tilde{\xi})-f''(\xi)}{\tilde{\xi}-\xi}(\tilde{\xi}-\xi) \\
=&f'(x_0) + \frac{h}{2}f'''(\xi^\star)(\tilde{\xi}-\xi) = f'(x_0) + \mathcal{O}(h^2),
\end{aligned}$$

ahol $\xi^\star$ ξ és $\bar{\xi}$ közé esik, azaz a közelítés rendje eggyel nagyobb lett. ◇

1.4. A MATLAB programcsomag

A MATLAB története az 1970-es évek közepén kezdődött. Cleve Moler, az Új-Mexikói Egyetem numerikus módszerek tanára felismerte, hogy ahelyett, hogy a hallgatók az egyes numerikus eljárások FORTRAN-ban való programozását csinálnák az órákon, az egyes numerikus eljárásokat előre megírt programokkal tesztelhetnék. Így a programírás helyett az algoritmusok vizsgálatára lehet koncentrálni. Cleve Moler elkészített egy programcsomagban pár függvényt a diákjai számára. 1984-ben egy Jack Little nevű villamomérnök vendégeskedett Molernél, akinek nagyon megtetszett a programcsomag. Ő is segített programokat írni, és segített a korábbi függvények C-nyelvre való átírásában. 1985-ben magalapították a MathWorks céget, amely mind a mai napig a MATLAB fejlesztője és forgalmazója.

A MATLAB függvények (m-fájlok) fő egysége a mátrix, minden eljárás mátrixokon alapul. Innét is kapta tulajdonképpen a nevét, ami a MATrix LABoratory szóösszetételből származik. A program parancsainak szintaxisa jól megjegyezhető és kényelmes, a függvények szerkezete egyszerű és könnyen áttekinthető. A korábban definiált függvények felhasználhatók újabb függvények definiálására is. Így a MATLAB-ban az eljárások sokkal gyorsabban megírhatók, mint más programozási nyelveken. Több olyan weblap található az interneten, ahova a MATLAB felhasználói töltöttek fel részletes leírással m-fájlokat
(pl. http://matlabdb.mathematik.uni-stuttgart.de/index.jsp).

Különböző alkalmazási területekhez külön eszköztárak (toolbox) készültek, így van pl. statisztikai, jelfeldolgozás, vagy parciális differenciálegyenletek eszköztár. Több területhez készült interaktív alkalmazás, ahol egyszerűen menüből állíthatók bizonyos eljárások paraméterei.

A MATLAB hátrányai között szokás említeni más matematikai programokkal szemben, hogy numerikusan hajtja végre a számításokat, ahol a numerikus számítások pontossága behatárolt (lásd részletesen a 2.5. fejezetet). A numerikus számítási mód azonban a legtöbb valódi alkalmazásokat tartalmazó feladat esetén az egyetlen lehetséges megoldási mód, mivel szimbolikusan nem lehet a megoldást előállítani. Szimbolikus számítások végezhetők a MATLAB-ban pl. a symbolic vagy a MAPLE for MATLAB (http://www.maplesoft.com/products/maplematlab/) eszköztár segítségével, vagy a speciálisan szimbolikus számításokra kifejlesztett MAPLE vagy MATHEMATICA (http://www.maplesoft.com/, http://www.wolfram.com/) programok segítségével. Másik hátrányként említik, hogy a függvények futási ideje lassabb, mintha azokat C-ben vagy más hasonló programozási nyelven írtuk volna. Ezt a hátrányt részben kompenzálja viszont az,

hogy a programok megírása sokkal gyorsabb és kényelmesebb eljárás, mint más nyelveken.

A programcsomag iránt mélyebben érdeklődő olvasóknak ajánljuk a magyar nyelven elérhető [33] könyvet vagy a MATLAB-ot forgalmazó cég egyik alapítója által írt és online is elérhető [25] könyvet. Nagyon hasznosak továbbá a MATLAB honlapján (www.mathworks.com) található linkek és információk is.

Ezen jegyzetnek nem célja a MATLAB programcsomag bemutatása és részletes ismertetése. Mégis, mivel a numerikus számítások legelterjedtebb és szinte nélkülözhetetlen eszköze a MATLAB, minden fejezet végén összegyűjtjük a fejezettel kapcsolatos MATLAB parancsokat. Mivel csak a numerikus eljárásokkal kapcsolatos parancsokra szeretnénk koncentrálni, így feltesszük, hogy az Olvasó ismeri már az alapvető MATLAB parancsokat. Ezek címszavakban a következők: sor- és oszlopvektor megadása, mátrixok megadása, műveletek mátrixokkal, műveletek elemenkénti elvégzése, hivatkozás mátrixok elemeire ill. almátrixaira, nevezetes mátrixok megadása (egységmátrix, nullmátrix, Toeplitz-mátrix, diagonális mátrix), a : jelölés, a for és a while ciklusok alkalmazása, az if elágazás, mátrixok inverzének, determinánsának és rangjának számítása.

A szemléltető ábrákat ill. az egyes eljárásokat bemutató függvényeket is MATLAB-ban készítettük. A függvényekhez tartozó m-fájlok elektronikusan is elérhetők, vagy egyszerűen e jegyzetből egy üres m-fájlba másolhatók.

1.5. A fejezettel kapcsolatos MATLAB parancsok

Most felsoroljuk azon MATLAB parancsokat, melyek a bevezető fejezetben szereplő fogalmakkal kapcsolatosak. Az egyszerűség kedvéért konkrét példákat mutatunk az alkalmazásra.

```
>> A=[1,2,3;4,5,6] % mátrixmegadás

A =

     1     2     3
     4     5     6

>> norm(A,2), norm(A,1), norm(A,inf) % A 2-es, 1-es és maximumnormák kiszámítása

ans =

   9.50803200069572


ans =

     9


ans =

    15

>> B=A*A'

B =
```

```
    14    32
    32    77

>> eig(B) % A B mátrix sajátértékeinek kiszámítása

ans =

   0.59732747374606
  90.40267252625394

>> [V,L]=eig(B)
% A B mátrix sajátvektorainak (V oszlopvektorai)
% ill. sajátértékeinek kiszámítása (L diagonális elemei)

V =

  -0.92236578007706   0.38631770311861
   0.38631770311861   0.92236578007706


L =

   0.59732747374606                   0
                  0  90.40267252625394

>> C=[1 2 3; 4 5 6; 7 8 9] % mátrixmegadás

C =

     1     2     3
     4     5     6
     7     8     9

>> tril(C) % A C mátrix alsó háromszög része

ans =

     1     0     0
     4     5     0
     7     8     9

>> triu(C) % A C mátrix felső háromszög része

ans =

     1     2     3
     0     5     6
     0     0     9

>> diag(C) % A C mátrix diagonálisát tartalmazó oszlopvektor

ans =
```

```
     1
     5
     9

>> diag(diag(C)) % A C mátrix diagonális mátrixa

ans =

     1     0     0
     0     5     0
     0     0     9
```

1.6. Feladatok

Normák

1.6.1. feladat. Azonosítsuk $\mathbb{R}^2$ elemeit a sík pontjaival! Adjuk meg a síkon azon pontok halmazát, melyek távolsága az origótól kisebb, mint egy! Használjuk az 1-es, 2-es és maximumnormákat!

1.6.2. feladat. Igazoljuk az 1.1.24. tételben szereplő formulát a maximumnorma képletére!

1.6.3. feladat. Igazoljuk az 1.1.27. tétel állításait!

1.6.4. feladat. Igazoljuk az 1-es, 2-es és maximumnormák ekvivalenciáját az ekvivalencia definíciójában szereplő c_1 és c_2 konstansok megkeresésével!

1.6.5. feladat. Tekinsük az $\|\mathbf{A}\| := \max_{i,j=1\ldots,n}\{|a_{ij}|\}$ mátrixnormát! Igazoljuk, hogy ez valóban norma. Mutassuk meg, hogy nem lehet vektornormából származtatni!

1.6.6. feladat. Igazoljuk, hogy indukált mátrixnorma esetén

$$\|\mathbf{A}\| = \max_{\|\mathbf{B}\|\leq 1}\{\|\mathbf{AB}\|\}.$$

1.6.7. feladat. Igazoljuk, hogy ha $\mathbf{A}$ nemszinguláris mátrix, akkor az $\|\overline{\mathbf{x}}\|_A := \|\mathbf{A}\overline{\mathbf{x}}\|$ hozzárendelés vektornorma bármilyen $\|\cdot\|$ vektornorma esetén!

1.6.8. feladat. Igazoljuk, hogy az 1.2.32. tételben szereplő $\|\cdot\|_{\mathbf{S},\mathbf{D}}$ norma valóban vektornorma!

1.6.9. feladat. Egy $\mathbf{A}$ mátrix Frobenius-normáját az alábbi képlettel számítjuk: $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2}$. Igazoljuk, hogy $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}^T\mathbf{A})$, ahol a $\mathrm{tr}(\cdot)$ jelölés az adott mátrix főátlóbeli elemeinek összegét jelenti (amely megegyezik a sajátértékek összegével is). Igazoljuk továbbá, hogy ha $\mathbf{A}$ és $\mathbf{B}$ ortogonálisan hasonlók, akkor Frobenius-normájuk megegyezik.

1.6.10. feladat. Legyen egy véges dimenziós V vektortérben $v_1, \ldots, v_n$ egy bázis. Tekintsük azt a hozzárendelést, amely egy tetszőleges $x = \sum_{i=1}^n \alpha_i v_i \in V$ vektorhoz ($\alpha_i \in \mathbb{K}$) a

$$\mu(x) = \sqrt{\sum_{i=1}^n |\alpha_i|^2}$$

értéket rendeli. Igazoljuk, hogy a μ függvény norma!

1.6.11. feladat. Tekintsük az f kétszer folytonosan deriválható függvényt az $[a,b]$ intervallumon! Definiáljuk egy adott n természetes szám esetén az $x_k = a + k(b-a)/n$ $(k = 0,\dots,n)$ osztópontokat, és legyen $\overline{\mathbf{x}}_n = [x_0,\dots,x_n]^T \in \mathbb{R}^{n+1}$ és $f(\overline{\mathbf{x}}_n) = [f(x_0),\dots,f(x_n)]^T \in \mathbb{R}^{n+1}$. Igazoljuk, hogy érvényes az alábbi konzisztenciatulajdonság:

$$\lim_{n\to\infty} \|f(\overline{\mathbf{x}}_n)\|_\infty \to \|f\|_{C[a,b]}!$$

(Útmutatás: Nyilvánvalóan a diszkrét norma mindig alulról becsüli a folytonosat. A másik becsléshez pedig használjuk ki, hogy a szakaszonként lineáris interpolációs függvény interpolációs hibája egy kétszer folytonosan deriválható függvény esetén felülről becsülhető az $[a,b]$ intervallumon az

$$\frac{M_2(b-a)^2}{8n^2}$$

kifejezéssel (6.2.5. tétel), ahol M_2 egy felső becslés $\|f''\|_{C[a,b]}$-re. Ezek után az állítás a rendőr-elvből következik.)

1.6.12. feladat. Tekintsük az előző példában adott f függvényt! Igazoljuk, hogy $\|f\|_{L^2[a,b]} := \sqrt{\int_a^b f^2(x)\,\mathrm{d}x}$ normát ad meg $C[a,b]$-ben. Igazoljuk, hogy $\|f(\overline{\mathbf{x}}_n)\|_2$ és $\|f\|_{L^2[a,b]}$ nem konzisztens az előző feladatban szereplő definíció értelmében, de az $\|f(\overline{\mathbf{x}}_n)\|_{l_2} := \sqrt{\|f(\overline{\mathbf{x}}_n)\|_2^2/n}$ norma (ezt is igazoljuk!) már konzisztens lesz $\|f\|_{L^2[a,b]}$-vel! (Útmutatás: Az első rész igazolásához tekintsük az $f(x) \equiv 1$ függvényt. A második rész igazolásához vegyük észre, hogy $(f^2(x_0) + \ldots + f^2(x_{n-1}))/n$ az $\int_a^b f^2(x)\,\mathrm{d}x$ integrál egy közelítő összege.)

1.6.13. feladat. Tegyük fel, hogy az $F : [a,b] \to [a,b]$, $F([a,b]) \subset [a,b]$ függvényre igaz, hogy valamilyen m pozitív egészre a $T := F^m = F \circ F \circ \ldots \circ F$ függvény kontrakció az $[a,b]$ intervallumon. Igazoljuk, hogy az F függvénynek pontosan egy fixpontja van!

1.6.14. feladat. Tekintsük az $F : [1,\infty) \to [1,\infty)$, $F(x) = x/2 + 1/x$ függvényt. Igazoljuk, hogy F kontrakció. Határozzuk meg a lehető legkisebb kontrakciós tényezőt! Adjuk meg F fixpontját!

1.6.15. feladat. Tegyük fel, hogy a Banach-féle fixponttétel feltételei közül a kontrakciós feltételt $(\exists\ 0 \le L < 1,\ \|F(x) - F(y)\| \le L\|x-y\|,\ \forall\ x,y \in H)$ kicseréljük az

$$\|F(x) - F(y)\| < \|x-y\|,\ \forall\ x,y \in H$$

feltételre. Igazoljuk, hogy ekkor F-nek maximum egy fixpontja lehet, de az is lehet, hogy nincs fixpont. Vizsgáljuk az $F : [1,\infty) \to [1,\infty)$, $F(x) = x + 1/x$ függvényt!

Nevezetes mátrixok

1.6.16. feladat. Igazoljuk, hogy ha $\mathbf{A} \in \mathbb{R}^{n\times n}$ ferdén szimmetrikus, akkor az

$$(\mathbf{E} + \mathbf{A})^{-1}(\mathbf{E} - \mathbf{A})$$

mátrix ($\mathbf{A}$ ún. Cayley-transzformáltja) ortogonális!

1.6.17. feladat. Igazoljuk, hogy ha egy Hessenberg-mátrix szimmetrikus, akkor tridiagonális!

1.6.18. feladat. Igazoljuk, hogy felső háromszögmátrixok szorzata és inverze (ha létezik) is felső háromszögmátrix!

1.6.19. feladat. Igazoljuk, hogy ha egy $\mathbf{T}$ felső háromszögmátrixra $\mathbf{T}^T\mathbf{T} = \mathbf{T}\mathbf{T}^T$, akkor $\mathbf{T}$ diagonális mátrix!

1.6.20. feladat. Legyen $\mathbf{Q} = \text{tridiag}\,[-1, 2, -1] \in \mathbb{R}^{3\times 3}$. Ez a mátrix M-mátrix. Adjunk meg olyan $\overline{\mathbf{g}} > 0$ vektort, mellyel $\mathbf{Q}\overline{\mathbf{g}} > 0$. Hogyan lehetne a $\overline{\mathbf{g}}$ vektort megadni, ha $\mathbf{Q}$ $n \times n$-es mátrix?

1.6.21. feladat. Igazoljuk, hogy az előző feladat 3×3-as mátrixa pozitív definit mátrix!

1.6.22. feladat. Legyen

$$\mathbf{C} = \begin{bmatrix} 1 & -0.1 & -0.2 \\ -0.1 & 1 & -0.1 \\ -0.2 & -0.1 & 1 \end{bmatrix}.$$

Igazoljuk, hogy $\mathbf{C}$ invertálható, és adjunk felső becslést az inverz mátrix 1-es normájára az inverz mátrix kiszámítása nélkül.

Sajátvektor, sajátérték

1.6.23. feladat. Igazoljuk, hogy ha egy szimmetrikus mátrix minden sajátértéke pozitív, akkor az pozitív definit mátrix. (Igazoljuk, hogy ha egy mátrix szimmetrikus, pozitív szemidefinit, és determinánsa zérustól különböző, akkor a mátrix pozitív definit! Igazoljuk, hogy ha egy szimmetrikus M-mátrixnak szigorúan domináns a főátlója, akkor a mátrix pozitív definit!)

1.6.24. feladat. Adjuk meg az alábbi mátrixok sajátvektorait és sajátértékeit! Ha lehetséges, akkor diagonalizáljuk őket!

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -8 & -12 & -6 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 3 & 2 & 4 \\ 1 & 4 & 4 \\ -1 & -2 & -2 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 5 & 1 & -1 \\ 1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

1.6.25. feladat. Igazoljuk, hogy ha $\mathbf{A} \in \mathbb{R}^{(2m+1)\times(2m+1)}$ olyan négyzetes mátrix, melyre $\det \mathbf{A} = 1$ és $\mathbf{A}$ ortogonális, akkor 1 sajátértéke $\mathbf{A}$-nak!

1.6.26. feladat. Határozzuk meg az $\mathbf{A} - \lambda\overline{\mathbf{v}}\overline{\mathbf{v}}^T$ mátrix sajátértékeit és sajátvektorait, ha tudjuk, hogy $\mathbf{A}$ egy szimmetrikus mátrix, melynek λ egy sajátértéke, és $\overline{\mathbf{v}}$ a hozzá tartozó sajátvektor!

Ellenőrző kérdések

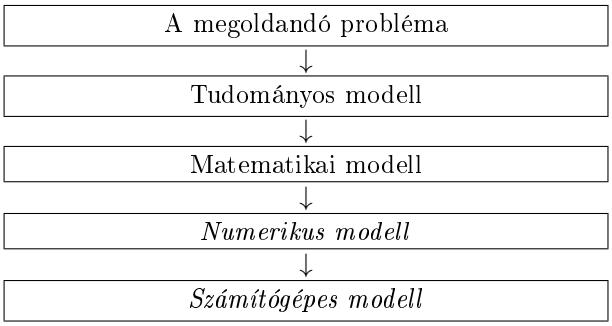
1. Melyek a nevezetes vektornormák?
2. Hogyan lehet vektornormából mátrixnormát létrehozni? Adjuk meg, hogy a nevezetes vektornormák esetén melyek ezek a mátrixnormák!
3. Milyen tulajdonságai vannak az indukált mátrixnormáknak?
4. Milyen kapcsolat van egy mátrix spektrálsugara és indukált normája között?
5. Ismertessük a Banach-féle fixponttételt!
6. Ismertessük a Gersgorin-tételeket!
7. Milyen mátrixokat hívunk M-mátrixnak? Hogyan ellenőrizhető ez a tulajdonság egy adott mátrix esetén?

2. Modellalkotás és hibaforrásai

Ebben a fejezetben bemutatjuk azt, hogy az egyes tudományágak által felvetett problémák megoldása során milyen út vezet a numerikus eljárások alkalmazásához. Megvizsgáljuk, hogy ezen folyamat alatt milyen hibák terhelhetik a végleges megoldást. Bevezetjük a kondíciószám fogalmát, amely azt méri, hogy egy feladat mennyire érzékeny a hibákra. Végül megismerjük a lebegőpontos számábrázolást.

2.1. Modellalkotás

A numerikus matematika a folytonos matematika problémáihoz konstruál megoldási eljárásokat, és elemzi azokat hatékonyságuk szempontjából. Ilyen eljárásokra általában azért van szükség, mert a folytonos problémát nem, vagy csak nagy nehézségek árán (sok idő vagy pénz) lehetne egzaktul megoldani. A folytonos matematikai problémák általában valamilyen más tudományág területéről származnak, pl. a fizika, kémia vagy a közgazdaságtan területéről, és míg eljutunk a feladat megoldásáig, általában többfajta egyszerűsítéssel, modellel kell élnünk. Ezeket a lépéseket szemlélteti a 2.1.1. ábra. Bár ebben a jegyzetben csak a numerikus és a számítógépes modellekkel



2.1.1. ábra: Egy probléma megoldásához vezető modellek.

fogunk foglalkozni, most egy példa erejéig bemutatjuk a modellalkotás többi lépését is.

Példaként tekintsük azt a fizikai feladatot, amikor egy inga lengésidejét szeretnénk meghatározni. Esetünkben ez a *megoldandó probléma*.

A feladat megoldásához természetesen élnünk kell bizonyos alapfeltevésekkel, pl. azzal, hogy az ingát úgy tekintjük, mint egy súlytalan kötélen lógó pontszerű testet, valamint hogy elhanyagolhatjuk a súrlódásból és a közegellenállásból származó veszteségeket. A feladatot így az energiamegmaradás törvényét használva oldhatjuk meg. Ez a *tudományos*, jelen esetben a fizikai modell.

Bevezetve az m jelölést a pontszerű test tömegére, az l jelölést a kötél hosszára és g-vel jelölve a gravitációs gyorsulást, az energiamegmaradás törvénye az

$$\frac{1}{2}ml^2(\phi'(t))^2 + mgl(1-\cos\phi(t)) = mgl(1-\cos\alpha)$$

alakban írható fel, ahol $\phi(t)$ a függőleges egyenessel bezárt szögét adja meg az ingának a t idő függvényében. Feltesszük, hogy a $t = 0$ időpillanatban ez a szög α volt ($\phi(0) = \alpha$). Az egyenlet átrendezéséből adódik, hogy

$$\phi'(t) = -\sqrt{\frac{2g}{l}}\sqrt{\cos\phi(t) - \cos\alpha}.$$

(Az elengedéstől a függőleges helyzetig a $\phi'(t)$ szögsebesség negatív). Innét szeretnénk a lengésidőt meghatározni. Ehhez osszuk el az egyenlet mindkét oldalát a jobb oldallal, és integráljuk mindkét oldalt a $[0, T/4]$ intervallumon, ahol T jelenti a keresett lengésidőt.

$$\int_0^{T/4} \frac{\phi'(t)}{-\sqrt{\frac{2g}{l}}\sqrt{\cos\phi(t) - \cos\alpha}}\,\mathrm{d}t = T/4.$$

Kétszer alkalmazva a helyettesítéses integrálás képletét (másodszor a $\sin\vartheta = \sin(\varphi/2)/\sin(\alpha/2)$ helyettesítéssel) a lengésidőre az alábbi kifejezést kapjuk:

$$\begin{aligned} T &= 2\sqrt{2}\sqrt{\frac{l}{g}}\int_0^{\alpha} \frac{1}{\sqrt{\cos\phi - \cos\alpha}}\,\mathrm{d}\phi \\ &= 4\sqrt{\frac{l}{g}}\int_0^{\pi/2} \frac{1}{\sqrt{1 - \sin^2(\alpha/2)\sin^2\vartheta}}\,\mathrm{d}\vartheta. \end{aligned}$$

Ez a *matematikai modell.*

Az integrál explicit alakban nem adható meg (elliptikus integrálról van szó), így az integrál értékét valamilyen numerikus integrálási eljárással kell közelítenünk (lásd a 8. fejezetet). Ez adja a *numerikus modellt.*

A numerikus modell által meghatározott számításokat számítógépen végezzük el. Ez adja a *számítógépes modellt.* Pl. $l = 1m$ és $g = 9.8m/s^2$ választás esetén $\alpha = 5°$-os kezdeti kitérés esetén $T = 2.008035541s$ adódik lengésidőnek, míg $\alpha = 90°$-ra $T = 2.369049722s$.

Szokásos eljárás az is, hogy a matematikai modellben szereplő integrált tovább egyszerűsítjük úgy, hogy az $1/\sqrt{1-x}$ függvényt az $x = 0$ pontban sorbafejtjük, a sorfejtést alkalmazzuk az $x = \sin^2(\alpha/2)\sin^2\vartheta$ választással, majd elvégezzük az integrálást. Ekkor azt kapjuk, hogy

$$T = 2\pi\sqrt{\frac{l}{g}}\left(1 + \frac{1}{4}\sin^2\frac{\alpha}{2} + \dots\right), \tag{2.1.1}$$

ahonnét a lengésidő becsülhető a

$$T = 2\pi\sqrt{\frac{l}{g}}$$

képlettel, amennyiben feltesszük az eredeti feladatban, hogy csak kis kitérésekre szeretnénk a lengésidőt megadni. A példában szereplő adatokkal a $T = 2.007089923s$ érték adódik, ami elég jó közelítése az $\alpha = 5°$-ra számított korábbi értéknek.

2.2. A modellalkotás hibaforrásai

A modellalkotási folyamat során számtalan helyen követhetünk el hibákat. Ezek természetes velejárói a modellalkotásnak. A célunk az, hogy a modellekben lévő hibákat megbecsüljük és kontrolálni tudjuk. Ha a modellalkotás végén nyert eredmény nem egyeztethető össze az eredeti feladattal, akkor a pontosabb megoldás érdekében a becslések alapján tudunk változtatni a modelleken.

Megjegyezzük, hogy a hiba jelenléte egy modellben nem feltétlenül rossz. Gondoljunk csak arra, hogy bizonyos esetekben éppen a hiba jelenléte eredményezi azt, hogy meg tudjuk oldani egzaktul a matematikai modell egyenletét (pl. az inga példájában a (2.1.1) sorfejtés csonkolása).

Vegyük most sorra a modellalkotás legfontosabb hibaforrásait (2.2.1. ábra)!

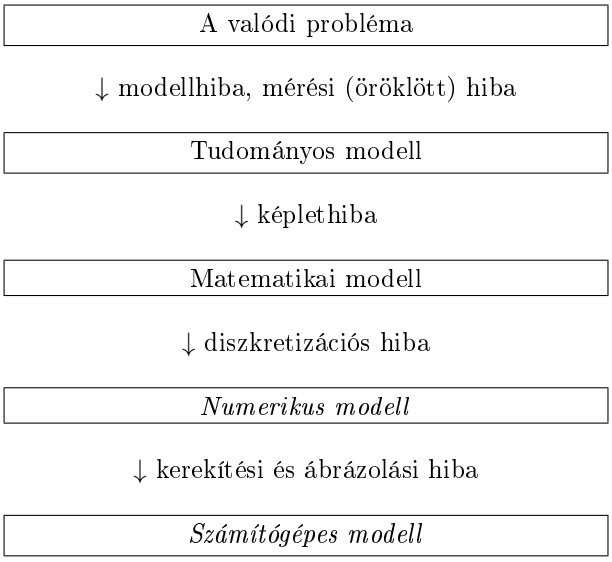
- Modellhiba: Általában a tudományos modellek nem tükrözik teljesen a valóságot. Az ebből származó hibát *modellhibának* nevezzük. Az ingás példában pl. a kötél tömege biztosan nem hanyagolható el, a test sem tekinthető pontszerűnek, és közegellenállás is van, stb. Ezek elhanyagolásából hiba kerül a modellbe.

- Mérési hiba: A valódi probléma vizsgálatához szükségünk van bizonyos paraméterek értékére. Ezeket általában mérnünk kell. Innét *mérési hiba* kerül a modellbe. Az ingás példában pl. meg kell mérnünk a kötél hosszát, és tudnunk kell a gravitációs gyorsulás értékét.

- Képlethiba: Amikor egy képletet kezelhetőségének érdekében úgy egyszerűsítünk, hogy bizonyos részeit elhagyjuk, vagy egyszerűbbel helyettesítjük, *képlethiba* keletkezik. Az ingás példánál képlethiba keletkezett, amikor a (2.1.1) végtelen sorból csak az első tagot tartottuk meg kis α kezdeti kitérés esetén.

- Diszkretizációs hiba: A numerikus eljárások során keletkező hiba a *diszkretizációs hiba*. Abból ered, hogy pl. a deriváltat differenciahányadossal, az integrált részletösszeggel és általában a folytonos függvényeket ún. rácsfüggvényekkel közelítjük. Az ingás példában az integrál közelítő kiszámításánál keletkezett diszkretizációs hiba.

- Kerekítési és ábrázolási hiba: Ha számítógéppel számítunk ki valamit, akkor a bevitt adatokat ábrázolja a számítógép a saját számrendszerében. Az ebből eredő hiba az *ábrázolási hiba*. A számítások során kerekíteni fog a számítógép, azaz *kerekítési hibát* követ el. Az ingás példánál ilyen hibákat követtünk el a lengésidő számszerű kiszámítása során.



2.2.1. ábra: A modellalkotás hibaforrásai.

A teljes modell elkészítése során a modellhibán, a mérési hibán és a képlethibán nem nagyon tudunk csökkenteni. Ezekről el kell fogadnunk, hogy vannak (vagy más modellt, pontosabb mérőműszert vagy más képletet kell alkalmaznunk). Az ábrázolási és kerekítési hibákról is el kell fogadnunk, hogy vannak, de a műveletek ügyes szervezésével csökkenteni lehet ezeket. Ennek lehetőségeit a 2.5. fejezetben és az egyes numerikus módszerekkel foglalkozó fejezetekben fogjuk majd bemutatni. A legtöbbet a diszkretizációs hibával fogunk foglalkozni majd. Ez a hiba általában tetszőlegesen kicsivé tehető, de ennek az az ára, hogy a számítógépes modell megoldása sokkal több időt fog igénybe venni.

2.3. A hiba mérése

Az előző fejezetben láttuk, hogy a modellalkotási folyamat során milyen hibák kerülhetnek a modellbe. A hiba általában elkerülhetetlen a modellezés során, de annak nagysága általában becsülhető.

Legyen x egy normált tér tetszőleges eleme, melyet közelítünk az $\hat{x}$ elemmel, amely a mérés, képletcsonkolás, diszkretizáció, számábrázolás vagy kerekítés miatt eltér az x elemtől.

2.3.1. definíció.
Az $\|\hat{x} - x\|$ értéket, azaz a pontos x és a közelítő $\hat{x}$ elem távolságát az $\hat{x}$ közelítés (x elemhez viszonyított) *abszolút hibájának* nevezzük. Amennyiben $\|x\| \neq 0$, akkor az

$$\frac{\|\hat{x} - x\|}{\|x\|}$$

hányadost az $\hat{x}$ közelítés (x elemhez viszonyított) *relatív hibájának* nevezzük.

Nyilvánvaló, hogy a relatív hiba sokkal jobban kifejezi a hiba nagyságát, hiszen az abszolút hibát a pontos érték normájához viszonyítva adja meg. Az abszolút és relatív hibákat általában nem ismerjük, hiszen kiszámításukhoz szükségünk lenne a közelített x elemre, amit természetesen nem ismerünk. Emiatt megadunk egy ún. abszolút és relatív hibakorlátot (jelölésük általában Δx és δx), melyek felülről becsülik a hibákat az $\|\hat{x} - x\| \leq \Delta x$ és

$$\frac{\|\hat{x} - x\|}{\|x\|} \leq \delta x$$

módon. Megjegyezzük, hogy a hibakorlátokat is a rövidség kedvéért általában csak egyszerűen hibának hívjuk. Az 1.1.10. tételt alkalmazva az $\|\hat{x} - x\| \leq \Delta x$ összefüggésből a

$$|\|\hat{x}\| - \|x\|| \leq \Delta x$$

becslést nyerjük, és így írhatjuk, hogy

$$\|\hat{x}\| = (1 + \delta)\|x\|, \tag{2.3.1}$$

valamilyen $|\delta| \leq \delta x$ konstanssal.

Érdemes megvizsgálni azt, hogy a valós számokkal végzett alapműveletek során hogyan viselkednek a hibák. Az egyszerűség kedvéért azt vizsgáljuk, hogy pozitív valós számokkal elvégezve az alapműveleteket, mi történik a hibával. Legyen tehát δx és δy az x és y pozitív valós számok közelítésének egy-egy relatív hibakorlátja. Ekkor a (2.3.1) egyenlőség miatt $\hat{x} = (1 + \delta_x)x$ és $\hat{y} = (1+\delta_y)y$, ahol $|\delta_x| \leq \delta x$ és $|\delta_y| \leq \delta y$ megfelelő konstansok. Vezessük be még a $\delta = \max\{\delta x, \delta y\}$

jelölést, melyről feltesszük, hogy mindig egynél kisebb értéket vesz fel, az alábbi képletekben. Ekkor az összeadás abszolút hibája:

$$|(\hat{x}+\hat{y})-(x+y)| = |(\hat{x}-x)+(\hat{y}-y)| = |\delta_x x + \delta_y y| \leq \delta(x+y)$$

és relatív hibája

$$\frac{|(\hat{x}+\hat{y})-(x+y)|}{x+y} \leq \delta.$$

A kivonás abszolút hibája

$$|(\hat{x}-\hat{y})-(x-y)| = |(\hat{x}-x)-(\hat{y}-y)| = |\delta_x x - \delta_y y| \leq \delta(x+y)$$

és relatív hibája

$$\frac{|(\hat{x}-\hat{y})-(x-y)|}{|x-y|} \leq \delta\frac{x+y}{|x-y|}.$$

A szorzás esetén az abszolút hibára azt kapjuk, hogy

$$|(\hat{x}\hat{y})-(xy)| = |(1+\delta_x)(1+\delta_y)xy-(xy)| = |\delta_x xy + \delta_y xy + \delta_x\delta_y xy| \leq \delta(2+\delta)xy$$

és relatív hibája

$$\frac{|(\hat{x}\hat{y})-(xy)|}{xy} \leq \delta(2+\delta).$$

Az osztás esetén az abszolút hiba

$$|(\hat{x}/\hat{y})-(x/y)| = |(1+\delta_x)x/(1+\delta_y)y-(x/y)| = \left|\left(\frac{(1+\delta_x)}{(1+\delta_y)}-1\right)\frac{x}{y}\right| \leq \left(\frac{1+\delta}{1-\delta}-1\right)\frac{x}{y} = \frac{2\delta}{1-\delta}\frac{x}{y}$$

és relatív hibája

$$\frac{|(\hat{x}/\hat{y})-(x/y)|}{x/y} \leq \frac{2\delta}{1-\delta}.$$

A fenti hibabecslő képletek két dolog miatt tanulságosak. Az egyik a kivonás relatív hibája, amely két egymáshoz közeli szám kivonása esetén nagyon nagy lehet. Sokkal nagyobb, mint a két kivont szám relatív hibája külön-külön. Két szám osztásánál pedig az abszolút hiba lehet nagyon nagy, amennyiben az osztás nevezője sokkal kisebb, mint a számlálója. A többi esetben a hiba kicsi, ha $\hat{x}$ és $\hat{y}$ hibája is kicsi.

2.4. Feladatok kondicionáltsága

A matematikai modellek általában (differenciál)egyenlet(rendszer)ek. Ezek általánosan az

$$F(x,d) = 0 \tag{2.4.1}$$

alakban írhatók, ahol d ismert mennyiséget jelöl (ezek a feladat ún. bemeneti adatai), x az egyenlet ismeretlenje (kimeneti adat), F pedig egy megfelelő függvény. A továbbiakban feltesszük, hogy d és x is egy-egy normált tér eleme, és $\|.\|$ mindig az aktuális térbeli normát jelenti. Példaként gondolhatunk az $ax^2+bx+c=0$ másodfokú egyenletre, ahol a $d=[a,b,c]^T$ vektor lehetne a bemenő adat, az $x=[x_1,x_2]^T$ pedig a megoldásvektor, de gondolhatunk differenciálegyenletekre is, ahol d megadja a kezdeti- és/vagy peremfeltételt, x pedig a differenciálegyenlet peremfeltételeket kielégítő megoldása.

2.4.1. definíció.

Azt mondjuk, hogy a (2.4.1) feladat *korrekt kitűzésű*, ha $\exists \eta > 0$ úgy, hogy a feladatnak egyértelmű megoldása (jelölése $x_{d+\delta d}$) van minden olyan $d + \delta d$ esetén, melyre $\|\delta d\| \leq \eta$, és $\exists K(\eta, d) > 0$ úgy, hogy $\|x_{d+\delta d} - x_d\| \leq K(\eta, d)\|\delta d\|$ (a megoldás folytonosan függ d-től).

A fenti definíciót Hadamard[1] alkalmazta differenciálegyenletekre vonatkozó kezdeti- és peremértékfeladatok esetén [14]. Egyenesen azt állította, hogy csak korrekt kitűzésű feladatokkal érdemes foglalkozni, ami vissza is vetette egy időre más feladatok megoldásának vizsgálatát. Miért is kell nem korrekt kitűzésű feladatokkal is foglalkozni? Általában véve azért, mert ha a matematikai modellekbe mérési adatok kerülnek, akkor az már nem szolgáltat korrekt kitűzésű feladatot. Pl. egy differenciálegyenlet megoldása korrekt kitűzésű lehet, ha a peremfeltételt megadó függvényt ismerjük. Ez pedig a gyakorlatban a legritkább esetben fordul elő, hiszen a függvényértékeket csak mérni tudjuk, a méréseket pedig legfeljebb véges sok pontban tudjuk csak elvégezni. Így a feladatnak már nem feltétlenül létezik egyértelmű megoldása. Ebben a jegyzetben csak korrekt kitűzésű feladatokkal foglalkozunk. A matematikai fizika és az analízis nem korrekt kitűzésű feladatainak jó összefoglalása található a [22] monográfiában.

2.4.2. példa. Nem korrekt kitűzésű feladat pl. az

$$x - |\{a \in \mathbb{R} \,|\, a^2 + a + d/4 = 0\}| = 0$$

egyenlet megoldása, ahol x az $a^2 + a + d/4 = 0$ egyenlet megoldásainak számát adja meg a d paraméter függvényében. Könnyen látható, hogy $d < 1$ esetén $x = 2$, $d = 1$ esetén $x = 1$, és $d > 1$ esetén $x = 0$. Azaz a $d = 1$ pontban a feladat nem korrekt kitűzésű. A megoldás d paramétertől való folytonos függése nem teljesül. Ha a d paraméter valamilyen mérési adat lenne, és pontos értéke $d = 1$ lenne, akkor ha egy kicsit is pontatlanul mérünk, akkor lényegesen megváltozik a feladat megoldása. ◇

2.4.3. példa. Tekintsük az alábbi két egyenletrendszert!

$$\begin{array}{rcl} 5x_1 - 331x_2 &=& 3.5 \\ 6x_1 - 397x_2 &=& 5.2 \end{array} \qquad \begin{array}{rcl} 4.9x_1 - 331x_2 &=& 3.5 \\ 6x_1 - 397x_2 &=& 5.2 \end{array}$$

Látható, hogy a két egyenlet csak az első egyenlet x_1 együtthatójában különbözik. Ez az együttható a második egyenletrendszerben 2%-kal kisebb, mint az elsőben. Ez látszólag nem nagy különbség, de az első lineáris egyenletrendszer megoldása $x_1 = 331.7$, $x_2 = 5$, a másiké $x_1 = 8.1499$, $x_2 = 0.1101$, amik igencsak távol állnak egymástól. Ha az együttható mérési eredményből származik, akkor szinte semmit nem mondhatunk majd a megoldásról, hiszen kis mérési hiba is nagy változást eredményezhet a megoldásban. Legyen a d bemenő adat az első egyenlet x_1 együtthatója, és a kimenő adat az $x = [x_1, x_2]^T$ megoldásvektor. Könnyű látni, hogy ez a feladat a $d = 5$ pontban korrekt kitűzésű, tehát a megoldás folytonosan függ d-től. Akkor vajon mi okozza a megoldás nagymértékű változását? Az, hogy a folytonosság definíciójában szereplő $K(\eta, d)$ konstans nagyon nagy. ◇

[1] Jacques Salomon Hadamard (1865–1963), francia matematikus. Főbb kutatási területei a differenciálegyenletek elmélete és a függvénytan volt. Az Ő nevéhez fűződik a Cauchy–Hadamard-tétel a hatványsorok konvergenciasugaráról vagy az Hadamard-mátrixok vizsgálata. Fontos számelméleti eredménye a prímszámtétel, mely szerint az n-nél nem nagyobb prímszámok száma tart $n/\ln n$-hez, ha n tart végtelenhez. Hadamard bővebb életrajza megtalálható pl. a http://www.gap-system.org/~history/Mathematicians/Hadamard.html honlapon.

2.4.4. definíció.
A

$$\kappa(d) = \lim_{\delta d \to 0} \frac{\|x_{d+\delta d} - x_d\|/\|x_d\|}{\|\delta d\|/\|d\|}$$

értéket a (2.4.1) feladat *(relatív) kondíciószámának* nevezzük.

A kondíciószám tehát szemléletesen azt méri, hogy pl. a bemenő adat 1%-os megváltozása hány százalékos változást eredményez a megoldásban. Ha a kondíciószám kicsi, akkor jól kondicionált feladatról, ha pedig nagy, akkor rosszul kondicionált feladatról beszélünk. Hogy mit értünk kicsin vagy nagyon, az mindig az aktuális feladattól elvárt viselkedés függvénye. A kondíciószám, ahogy jeleztük is, a d paraméter függvénye. Amennyiben x_d vagy d nulla, a relatív kondíciószám nem értelmezhető.

2.4.5. definíció.
A

$$\kappa_{abs}(d) = \lim_{\delta d \to 0} \frac{\|x_{d+\delta d} - x_d\|}{\|\delta d\|}$$

értéket a (2.4.1) feladat *abszolút kondíciószámának* nevezzük.

Korrekt kitűzésű feladatok esetén d egy környezetében az $x_{d+\delta d}$ (egyértelmű) megoldás felírható $x_{d+\delta d} = G(d+\delta d)$ alakban, ahol a G függvényt *megoldófüggvénynek* hívjuk. Amennyiben a G függvény differenciálható d-ben, a relatív kondíciószámra a

$$\kappa(d) = \frac{\|G'(d)\| \cdot \|d\|}{\|G(d)\|},$$

míg az abszolút kondíciószámra a

$$\kappa_{abs}(d) = \|G'(d)\|$$

formulát nyerjük.

2.4.6. példa. A 2.4.3. példában szereplő egyenletrendszer esetén a kondíciószámot 2-es normában számolva $\kappa(5) \approx 1985$ adódik, ami azt jelenti, hogy a d paraméter 2%-os változása a megoldásvektor 2-es normáját 3970%-kal is megváltoztathatja. Így érthető, hogy a megoldás annyira nagyot változhatott. ◇

Természetesen a modellalkotás során kerülni kell a rosszul kondicionált feladatokat, vagy legalábbis jó tudnunk egy feladatról, hogy annak megoldása rosszul kondicionált. Ennek jelentősége onnét látszik, hogy egy rosszul kondicionált feladatban, ha a bemenő adatok mérési eredmények, akkor egy kis mérési pontatlanság is jelentősen megváltoztathatja a feladat megoldását.

2.5. Gépi számábrázolás és következményei

Ahhoz, hogy megértsük, hogy a kerekítési és ábrázolási hibák honnét származnak, ismernünk kell, hogy a számítógépek hogyan kezelik a valós számokat. Természetesen számtalan kezelési mód létezik. Mi ezek közül a MATLAB által használt dupla pontosságú lebegőpontos számokkal fogunk foglalkozni. A MATLAB lebegőpontos számrendszere a kettes számrendszeren alapszik. Ennek ellenére a konkrét számpéldákat tízes számrendszerben mutatjuk be, hiszen azok sokkal megszokottabbak, és a jelenségek hasonlóan érvényesek a kettes számrendszerben is.

2.5.1. példa. Felsorolunk néhány MATLAB által szolgáltatott eredményt.

- A MATLAB $\mathrm{tg}(\pi/2)$ értékére az $1.6331e+016$ értéket adja, ami nyilvánvalóan hibás, hiszen $\pi/2$-nél nincs is értelmezve a tangens függvény.
- $2^{-1074} = 4.94066e-324$, de $2^{-1074}/2 = 0$ és $2^{-1074} \cdot 1.2 = 4.94066e-324 = 2^{-1074}$, ami megintcsak hibás.
- 10^{310} =Inf.
- Összetettebb példaként tekintsük az alábbi eljárást π értékének meghatározására [15]: Jelentse y_k az egységkörbe írt szabályos 2^k-szög félkerületét. Ekkor nyilván $y_k \to \pi$, ha $k \to \infty$. Továbbá érvényes az

$$y_{k+1} = 2^{k+1}\sqrt{\frac{1}{2}\left(1 - \sqrt{1-(2^{-k}y_k)^2}\right)}$$

rekurzió, ahol $y_1 = 2$, $y_2 = 2\sqrt{2}$, ... Az iterációt számítógépen végrehajtva az alábbi számokat kapjuk: $y_{10} = 3.14158627$, $y_{12} = 3.14166137$, ..., $y_{19} = 3.70727600$, ... Azaz szemmel láthatóan a sorozat nem tart π-hez. Ebben a feladatban a matematikai vagy numerikus modell a π értékét szolgáltatja, de a számítógépes modell már nem.

◇

A számítógép két egysége játszik fontos szerepet a számítások elvégzésénél. Az egyik a processzor (CPU=central processing unit), a másik pedig a memória. A processzor végzi a műveleteket, a memóriában pedig a műveletekhez szükséges adatokat tároljuk. Innét rögtön láthatjuk, hogy pl. irracionális számokat vagy végtelen szakaszos tizedes törteket nem tudunk számítógépen tárolni. A memória méretének korlátozottságából következik, hogy csak véges sok racionális számot tudunk a memóriában elhelyezni. Megjegyezzük, hogy mivel a számítógépek a kettes számrendszert használják általában a tároláshoz, ezért pl. az 1/10-et sem tudják pontosan eltárolni, hiszen az

$$\frac{1}{10} = \frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^8} + \frac{1}{2^9} + \frac{1}{2^{12}} + \frac{1}{2^{13}} + \dots$$

egyenlőség miatt kettes számrendszerben az 1/10 végtelen szakaszos tizedes tört.

A valós számok ábrázolására a legelterjedtebb módszer az ún. lebegőpontos számábrázolás.[2] Ebben az esetben a valós számot a

$$\pm b^k \left(\frac{a_0}{b^0} + \frac{a_1}{b^1} + \frac{a_2}{b^2} + \dots + \frac{a_{p-1}}{b^{p-1}}\right) \equiv a_0.a_1a_2\dots a_{p-1} \times b^k \tag{2.5.1}$$

alakban írjuk fel (ha lehet), ahol b a számábrázolás alapja, p a szereplő számjegyek (mantissza) száma, és k a kitevő (karakterisztika). Az a_i $(i = 0, \dots, p-1)$ számjegyekről feltesszük, hogy azok az alapnál kisebb nemnegatív egész számok. Ha $a_0 \neq 0$, akkor azt mondjuk, hogy a felírt szám normálalakban van. Ha normálalakban írjuk fel a számokat, akkor az ábrázolásuk egyértelmű lesz. Természetesen, ha egy valós szám nem írható fel ilyen alakban, akkor a számot kerekítenünk kell. Megmutatható, hogy a kerekítés szokásos szabálya miatt (nevezetesen, hogy az 5-ös számjegyet felfelé kerekítjük) a számítások értéke bizonyos esetekben felfelé tolódhat, így a számítógépek

[2] A lebegőpontos számábrázolás bevezetése Konrad Zuse (1910-1995) német mérnök nevéhez köthető, aki több számítógépet is épített az éppen a második világháborúra készülődő Németországban. Eredményeire nem figyeltek akkoriban fel. Már 1941-ben létrehozott egy jelfogókkal működő, a lebegőpontos kettes számrendszeren alapuló, teljesen programozható számítógépet [37].

általában a párosra kerekítés (angolul: round to even) szabályát használják. Azaz pl. a 3.155-öt a szokásos módon 3.16-ra kerekítjük felfelé (a 6 páros), a 3.145-öt viszont 3.14-re lefelé kerekítjük (a 4 páros).

Könnyen látható, hogy a lebegőpontosan ábrázolható számok rendelkeznek az alábbi tulajdonságokkal:

- Csak véges sok racionális számot tudunk előállítani.
- A lehetséges számok nem alkotnak testet. Pl. nem asszociatív az összeadás). (pl.: $123.4 + 0.04 + 0.03 + 0.02 + 0.01$ kétféle sorrendben összeadva más-más eredményt ad a $p = 4$, $b = 10$, $k = -2, \ldots, 2$ lebegőpontos számrendszerben (jelölése: $F(4, -2, 2)$).
- Az előállítható számok korlátos halmazt alkotnak. Pl. $F(4, -2, 2)$-ben a legnagyobb előállítható szám a 999.9, a legkisebb pedig -999.9. Ha ezeknél nagyobb abszolút értékű lesz egy számítás eredménye, akkor túlcsordulásról beszélünk.
- A nulla körül "relatívan nagy űr van". Pl. $F(4, -2, 2)$-ben a legkisebb pozitív előállítható szám normál alakban 0.01, nem normál alakban 0.00001. Ha ezeknél kisebb pozitív szám lesz a számítás eredménye, akkor azt már nullaként ábrázolja a számítógép (alulcsordulás).
- Az 1 utáni legkisebb előállítható szám az ún. gépi epszilonnal (ε_g) nagyobb 1-nél. Az $F(4, -2, 2)$ rendszerben ez 0.001.

A MATLAB dupla pontosságú lebegőpontos számrendszere a következő struktúrájú: A számokat kettes számrendszerben (a (2.5.1) előállításban $b = 2$) állítja elő 64 bitnyi tárhelyet felhasználva. Egy biten tároljuk az előjelet ($0 = +, 1 = -$). 52 biten a mantisszát tároljuk, pontosabban annak a tizedespont utáni részét, hiszen normálalak esetén a tizedesponttól balra mindenképpen 1-esnek kell szerepelnie. 11 biten a karakterisztika tárolódik úgy, hogy a kitevőhöz hozzáadunk 1023-at, és az így nyert szám 2-es számrendszerbeli alakját tároljuk. Így a karakterisztika -1023-tól 1024-ig tárolható. A -1023-as karakterisztika a 0 (ha a mantissza is csupa nulla) ill. annak jelzésére szolgál, hogy az adott szám nincs normálalakban (ilyenkor az értéke $\pm 0.a_1 \ldots a_{52} \times 2^{-1022}$). A csupa 1-essel kódolt 1024-es karakterisztika speciális célokra foglalt. Ha a mantissza nem nulla, az azt jelenti, hogy a számítás művelete nem szám (jelölésben NaN = not a number). Ilyen pl. a 0/0 művelet eredménye. Ha a mantissza nulla, akkor az előjelbit szerint vagy $+\infty$-t, vagy $-\infty$-t jelöl.

A legnagyobb ábrázolható pozitív szám az

$$M = 1.\underbrace{111\ldots111}_{52\text{db}} \times 2^{1023} = 1.79769 \times 10^{308}$$

és a legkisebb ábrázolható pozitív szám

$$m = 0.\underbrace{000\ldots000}_{51\text{db}}1 \times 2^{-1022} = 4.94066 \times 10^{-324}.$$

Vegyük észre, hogy míg az első szám normálalakban van, addig a második nem. A normálalakban ábrázolható legkisebb pozitív szám az

$$1.\underbrace{000\ldots000}_{52\text{db}} \times 2^{-1022} = 2.22507 \times 10^{-308}.$$

Az 1-et az

$$1 \equiv 1.\underbrace{000\ldots000}_{52\text{db}} \times 2^{0}$$

módon ábrázolhatjuk. Az 1-nél nagyobb legkisebb ábrázolható szám az

$$1.\underbrace{000\ldots000}_{51\text{db}}1 \times 2^0,$$

ami $\varepsilon_g = 2^{-52}$-nel nagyobb mint az 1.

Jelentse $fl(x)$ egy $x \in \mathbb{R}$, $-M \leq x \leq M$ valós szám lebegőpontosan ábrázolt képét.

2.5.2. tétel.
Legyen $-M \leq x \leq M$. Ekkor

$$|fl(x) - x| \leq \begin{cases} \varepsilon_0, & \text{ha } |x| < \varepsilon_0, \\ \frac{\varepsilon_g |x|}{2}, & \text{ha } \varepsilon_0 \leq |x| \leq M. \end{cases}$$

Bizonyítás. A tétel első része triviális. Essen x az $x_i < x_j$ szomszédos lebegőpontos számok közé. Legyen x_i mantisszája p jegyű, és karakterisztikája k. Ekkor

$$|fl(x) - x| \leq \frac{x_j - x_i}{2} = \frac{b^{-p+1} b^k}{2} \leq \frac{\varepsilon_g |x|}{2}. \blacksquare$$

Az ε_g gépi epszilon felét *gépi pontosságnak* nevezzük. Jelölése $\mathtt{u}$. A gépi pontosság tehát a lebegőpontosan ábrázolt számok relatív hibakorlátja. A fenti tétel közvetlen következménye, hogy egy M-nél nem nagyobb abszolút értékű valós szám lebegőpontos képére igaz az $fl(x) = (1+\delta)x$ egyenlőség, ahol $|\delta| \leq \mathtt{u}$.

Egy számítógépes számítást leegyszerűsítve úgy képzelhetjük el, hogy a processzor megkapja azt a két számot, amikkel műveleteket kell végeznie, lebegőpontossá alakítja őket, ezekkel pontosan elvégzi a műveletet, majd pedig kerekíti az eredményt úgy, hogy lebegőpontosan ábrázolható legyen. Legyen $\diamond$ egy tetszőleges művelet valós számok között. Ekkor az x és y valós számokkal számítógépen a következő értéket kapjuk a művelet eredményére:

$$x \boxed{\diamond} y := fl(fl(x) \diamond fl(y)).$$

Nézzünk meg két alapműveletet (a kivonást és az osztást), hogy hogy viselkedik lebegőpontos számokkal végrehajtva! Először vizsgáljuk a kivonás relatív hibáját ($x, y > 0$).

$$\frac{|x \boxed{-} y - (x-y)|}{|x-y|} = \frac{|(x(1+\delta_x) - y(1+\delta_y))(1+\delta_-) - (x-y)|}{|x-y|}$$
$$\leq \frac{|(x\delta_x - y\delta_y)(1+\delta_-)|}{|x-y|} + |\delta_-| \leq \mathtt{u}(1+\mathtt{u})\frac{x+y}{|x-y|} + \mathtt{u},$$

ahol $|\delta_x|, |\delta_y|, |\delta_-| \leq \mathtt{u}$.

Ha x közel van y-hoz (azaz két egymáshoz közeli számot vonunk ki egymásból), akkor a különbség relatív hibája jóval nagyobb lehet, mint a gépi pontosság.

Két közeli szám kivonásánál abból is ered hiba, hogy a mantissza hossza véges. Ezt a jelenséget *kiegyszerűsödésnek* nevezzük. A jelenség szemléltetésére mutatunk be most két példát.

2.5.3. példa. Tekintsük az $F(10,-10,10)$ számrendszert. Ebben írjuk fel az alábbi gyököket:

$$\sqrt{9876} = 9.937806599 \times 10^1, \qquad \sqrt{9875} = 9.937303457 \times 10^1.$$

Most vonjuk ki egymásból a két számot.

$$\sqrt{9876} - \sqrt{9875} = 0.000503142 \times 10^1 = 5.03142 \underbrace{\mathbf{0000}}_{\text{értelmetlen}} \times 10^{-3}.$$

A megjelenő 4 darab nullának a számításokhoz nincs semmi köze, abból erednek, hogy a számjegyeket balra toltuk a mantisszában a normálalak létrehozásához.

A kiegyszerűsödést kiküszöbölhetjük a nevezetes szorzat alkalmazásával:

$$\sqrt{9876} - \sqrt{9875} = \frac{1}{\sqrt{9876} + \sqrt{9875}} = 5.031418679 \times 10^{-3}.$$

Jól mutatja a két számolás közti különbséget, hogy az első érték relatív hibája 2.6×10^{-7}, míg a másodiké 1.6×10^{-10}. ⋄

2.5.4. példa. A 2.5.1. példában is a kiegyszerűsödés okozta a gondot. Ha k növekszik, akkor $1 - \sqrt{1-(2^{-k}y_k)^2}$ második tagja egyre közelebb lesz 1-hez, így kiegyszerűsödés lép fel. A jelenség nyilván kiküszöbölhető a képlet megfelelő átírásával, nevezetesen az

$$y_{k+1} = y_k \sqrt{\frac{2}{1+\sqrt{1-(2^{-k}y_k)^2}}}$$

iterációban már nem lép fel kiegyszerűsödés. ⋄

Most vizsgáljuk az osztás hibáját! Legyenek x és y pozitív számok.

$$\left| x \boxed{/} y - x/y \right| = \left| \frac{x(1+\delta_x)}{y(1+\delta_y)}(1+\delta_/) - \frac{x}{y} \right|$$

$$= \frac{x}{y}\left| \frac{(1+\delta_x)}{(1+\delta_y)}(1+\delta_/) - 1 \right| \le \frac{x}{y}\left| \frac{(1+\mathtt{u})^2}{1-\mathtt{u}} - 1 \right| = \frac{x}{y}\left(3\mathtt{u} + \mathcal{O}(\mathtt{u}^2)\right),$$

ahol $|\delta_x|, |\delta_y|, |\delta_/| \le \mathtt{u}$. Ez mutatja, hogy az osztás abszolút hibája nagyon nagy lehet, ha a nevezőben szereplő szám jóval kisebb a számlálónál. Megjegyezzük, hogy a relatív hiba $3\mathtt{u}$ (azaz gépi pontosság) nagyságrendű.

Azt, hogy a lebegőpontos számokkal milyen módon kell számolni ill. hogy hogy kell kerekíteni őket, szabványok rögzítik. A legelterjedtebb ilyen szabvány az IEEE 745-ös (IEEE=Institute of Electrical and Electronics Engineers [16]). Az interneten több, a lebegőpontos számokat szemléltető alkalmazás érhető el. Ezek közül ajánlunk kettőt a [17, 38] honlapokon.

A számítógépen megírt programokat jól jellemzi az, hogy végrehajtásuk során hány lebegőpontos műveletet $(+,-,\cdot,/)$ hajtunk végre. Ezek megszámolásával összehasonlíthatunk két eljárást futási sebesség szempontjából. A műveletigényt általában *flop*-ban szoktuk mérni (*fl*oating point *op*eration).

2.6. A fejezettel kapcsolatos MATLAB parancsok

```
>> realmin % A legkisebb ábázolható pozitív szám.

ans =

    2.225073858507201e-308

>> realmax % A legnagyobb ábázolható pozitív szám.

ans =

    1.797693134862316e+308

>> eps % A gépi epsilon.

ans =

    2.220446049250313e-016
```

Végezzünk el egy kísérletet arra vonatkozóan, hogy a számítógép mennyi idő alatt hajt végre 10^8 lebegőpontos műveletet. A futási időt a `tic` (stopper indul) és `toc` (stopper megáll) parancsokkal mérhetjük. Vizsgáljuk az összeadást, a szorzást és az osztást.

```
>> x=sqrt(2); tic, for i=1:10^8 x=x+1; end, toc;
x=sqrt(2); tic, for i=1:10^8 x=x*1.00000001; end, toc;
x=sqrt(2); tic, for i=1:10^8 x=x/1.0000001; end, toc;

Elapsed time is 47.930000 seconds.
Elapsed time is 49.164000 seconds.
Elapsed time is 66.737000 seconds.
```

Látható tehát, hogy az összeadáshoz és a szorzáshoz szinte pontosan ugyanannyi idő kellett, míg az osztáshoz kb. 36%-kal több. Az adatok azt jelentik, hogy a használt számítógép kb. 2×10^6 műveletet (speciálisan szorzást) képes 1 másodperc alatt elvégezni.

2.7. Feladatok

Lebegőpontos számábrázolás

2.7.1. feladat. Vizsgáljuk meg, hogy korrekt kitűzésű-e az $x+dy=1,\ dx+y=0$ egyenletrendszer a d valós paraméter függvényében! Adjuk meg a kondíciószámot maximumnormában!

2.7.2. feladat. Vizsgáljuk meg az $x=-d+\sqrt{d^2-4}$ kifejezés kondicionáltságát a d változó függvényében! Milyen d értékek esetén lesz korrekt kitűzésű a feladat? Adjunk meg olyan d értéket, melyre a (relatív) kondíciószám 100-nál nagyobb!

2.7.3. feladat. Az $a = 0.001$ választás mellett $A = 1-1/(1-2a)$ értéke -0.002004008016. Határozzuk meg mi is A értékét egy tízes számrendszerű, hatjegyű mantisszás lebegőpontos számokat használó számítógépen! Javasoljunk numerikus szempontból jobb számolást A-ra, és végezzük el úgy is a számolásokat!

2.7.4. feladat. Az $a = 1000$ választás mellett $A = 1/(\sqrt{a+1}-\sqrt{a})$ értéke 63.26136064087. Határozzuk meg mi is A értékét egy tizes számrendszerű, hatjegyű mantisszás lebegőpontos számokat használó számítógépen! Javasoljunk numerikus szempontból jobb számolást A-ra, és végezzük el úgy is a számolásokat!

2.7.5. feladat. Egy 10-es számrendszeren alapuló számítógép a $\sin x$, $\cos x$, x^2 függvények értékeit pontosan számolja, majd az eredmények ábrázolásánál hatjegyű mantisszára kerekít. Határozzuk meg ezen a számítógépen az $f(x) = \cos^2 x - \sin^2 x$ függvény értékét az $x = 0.7854$ helyen! Mekkora a számított eredmény relatív hibája? Indokoljuk az eredményt! Javasoljunk jobb képletet az $f(x)$ érték kiszámítására!

2.7.6. feladat. Írjunk MATLAB programot az

$$y_{k+1} = 2^{k+1}\sqrt{\frac{1}{2}\left(1-\sqrt{1-(2^{-k}y_k)^2}\right)}$$

iteráció vizsgálatára! Hasonlítsuk össze az eredményt az

$$y_{k+1} = y_k\sqrt{\frac{2}{1+\sqrt{1-(2^{-k}y_k)^2}}}$$

iterációval! Magyarázzuk meg az eltérést!

2.7.7. feladat. Írjunk MATLAB programot az

$$e^x = \lim_{n\to\infty}\sum_{i=0}^{n}\frac{x^i}{i!}$$

sor összegének kiszámítására. Futtassuk negatív értékek esetén (pl. $x = -25$)! Mit tapasztalunk? Adjunk magyarázatot a jelenségre! Javasoljunk jobb módszert e^{-25} kiszámítására!

2.7.8. feladat. Legyen $F(p, k_{\min}, k_{\max})$ a tizes alapú lebegőpontos számok halmaza (p a mantissza hossza, $k_{\min}$ és $k_{\max}$ pedig a minimális és maximális karakterisztikát jelentik), ahol p a mantissza hossza, és $k_{\min}$ és $k_{\max}$ a karakterisztika legkisebb és legnagyobb értéke. Adjuk meg az $F(1,-2,2)$ rendszerben megadható számokat!

2.7.9. feladat. Adjuk meg az $F(1,-2,2)$ rendszerben az $1/3$, $1/900$, $20\cdot 200$, $(((2+0.1)+0.1)+\cdots+0.1)$ (10 összeadás), $(((0.1+0.1)+0.1)+\cdots+0.1)+2$ értékeket!

2.7.10. feladat. Adjunk meg olyan lebegőpontos számrendszert ($F(p, k_{min}, k_{max})$), melyben az alábbi számok ábrázolhatók!
a) 5,50,500,5000;
b) 5,5.5,5.55;
c) 5,0.5,0.05,0.005;
d) 5,55,555,5555.

2.7.11. feladat. Milyen lebegőpontos számrendszerben számolható kerekítés nélkül $2.2 \cdot 3.45$, $1/80$, $2\times 10^2 \cdot 7\times 10^2$?

Ellenőrző kérdések

1. Milyen tulajdonságai vannak egy korrekt kitűzésű feladatnak?
2. Mikor mondjuk egy korrekt kitűzésű feladatról, hogy rosszul kondicionált?
3. Hogy ábrázolja a MATLAB a valós számokat, és milyen következményei vannak ennek?
4. Soroljuk fel azokat a lebegőpontos műveleteket, melyek nagy hibával rendelkezhetnek!
5. Mekkora a MATLAB gépi pontossága?

3. Lineáris egyenletrendszerek megoldása

Ebben a fejezetben azzal foglalkozunk, hogy hogyan oldhatunk meg négyzetes, teljes rangú mátrixú lineáris egyenletrendszereket. Először megismerkedünk a lineáris egyenletrendszer megoldásának kondicionáltságával. Ezután ismertetjük a Gauss-módszert és a vele kapcsolatos mátrixfelbontásokat. Majd áttérünk az iterációs megoldásokra. Néhány egyszerű iterációs eljáráson kívül ismertetjük a gradiens és konjugált gradiens módszereket is.

3.1. Lineáris egyenletrendszerek megoldhatósága

A lineáris egyenletrendszerek általános alakja a következő. Tegyük fel, hogy adottak az a_{ij} és b_i számok ($i = 1, \ldots, m; j = 1, \ldots, n$). Keressük azokat az x_j számokat ($j = 1, \ldots, n$), melyekre

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m. \end{aligned}$$

Bevezetve az $\overline{\mathbf{a}}_j = [a_{1j}, \ldots, a_{mj}]^T$ jelölést ($j = 1, \ldots, n$) az egyenletrendszer ún. vektoros alakban is megfogalmazható: keressük azokat az x_j számokat ($j = 1, \ldots, n$), mellyekkel a

$$\overline{\mathbf{b}} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

vektor az

$$x_1\overline{\mathbf{a}}_1 + \cdots + x_n\overline{\mathbf{a}}_n = \overline{\mathbf{b}}$$

módon írható fel lineáris kombinációként. Az ún. mátrixos, az elméleti vizsgálódások során a leggyakoribb, alak úgy adható meg, hogy bevezetjük az

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$

ún. együtthatómátrixot. Keressük azt az $\overline{\mathbf{x}}$ n-elemű oszlopvektort, mellyel $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$. A lineáris algebrából jól ismert az alábbi tétel.

3.1.1. tétel.
Egy $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszer akkor és csak akkor megoldható, ha az $\mathbf{A}$ együtthatómátrix és a $\overline{\mathbf{b}}$ vektorral kibővített együtthatómátrix rangja megegyezik: $r(\mathbf{A}) = r(\mathbf{A}|\overline{\mathbf{b}})$. Ha az egyenletrendszer megoldható és $r(\mathbf{A}) < n$, akkor végtelen sok megoldás van, ha $r(\mathbf{A}) = n$, akkor egyértelmű a megoldás.

Ebben a fejezetben egészen a 3.7. alfejezetig feltesszük, hogy az egyenletrendszer mátrixa négyzetes, és hogy az egyenletrendszernek pontosan egy megoldása van csak. Ennek szükséges és elégséges feltétele, hogy az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix determinánsa nullától különbözzön. Azt is feltesszük, hogy az egyenletrendszer együtthatói valós számok. A tételek könnyen átfogalmazhatók komplex mátrixokra is.

3.2. Lineáris egyenletrendszerek kondicionáltsága

Láttunk már példát arra, hogy az együtthatók kis változására is sokat változhat egy lineáris egyenletrendszer megoldása. Ebben a fejezetben megvizsgáljuk, hogy ez a változás minek a következménye. Legyen tehát $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\overline{\mathbf{b}} \in \mathbb{R}^n$ és $\det(\mathbf{A}) \neq 0$. Legyen $\overline{\mathbf{x}}$ az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszer megoldása.

Tekintsük először azt az esetet, amikor megváltoztatjuk, más szóval perturbáljuk, az egyenletrendszer jobb oldali vektorát (bemenő adat) egy $\boldsymbol{\delta}\overline{\mathbf{b}}$ vektorral. Ekkor általában az egyenletrendszer megoldása (kimenő adat) is megváltozik, azaz $\overline{\mathbf{x}}$ helyett $\overline{\mathbf{x}} + \boldsymbol{\delta}\overline{\mathbf{x}}$ lesz a megoldás valamilyen megfelelő $\boldsymbol{\delta}\overline{\mathbf{x}}$ vektorral.

$$\mathbf{A}(\overline{\mathbf{x}} + \boldsymbol{\delta}\overline{\mathbf{x}}) = \overline{\mathbf{b}} + \boldsymbol{\delta}\overline{\mathbf{b}}$$

Jelentsen $\|.\|$ egy tetszőleges vektornormát és az általa indukált mátrixnormát. Ekkor a lineáris egyenletrendszer megoldásfüggvénye

$$\overline{\mathbf{x}} = G(\overline{\mathbf{b}}) = A^{-1}\overline{\mathbf{b}},$$

így a relatív kondíciószám

$$\kappa(\overline{\mathbf{b}}) = \frac{\|\mathbf{A}^{-1}\| \cdot \|\overline{\mathbf{b}}\|}{\|\mathbf{A}^{-1}\overline{\mathbf{b}}\|} \leq \frac{\|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\| \cdot \|\overline{\mathbf{x}}\|}{\|\overline{\mathbf{x}}\|} = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| =: \kappa(\mathbf{A}).$$

A relatív kondíciószám tehát felülről becsülhető a mátrix és inverze normájának szorzatával. Mivel ez a szorzat sokszor előfordul különböző becslésekben, ezért külön nevet adunk neki.

3.2.1. definíció.
Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ reguláris mátrix. Ekkor a $\kappa(\mathbf{A}) = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$ számot a *mátrix kondíciószámának* nevezzük.

A mátrixok kondíciószáma függ a használt mátrixnormától. Ezt néha jelezni szoktuk a kondíciószám indexeként, azaz pl. $\kappa_2(\mathbf{A})$ a 2-es normabeli kondíciószámot jelenti.

3.2.2. tétel.
Egy $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix indukált mátrixnormában vett kondíciószámára a következő tulajdonságok érvényesek:

- $\kappa(\mathbf{A}) \geq 1$,
- $\kappa(\mathbf{A}) = \kappa(\mathbf{A}^{-1})$,
- $\kappa(\alpha\mathbf{A}) = \kappa(\mathbf{A})$, $\alpha \neq 0$,
- Ortogonális mátrixra $\kappa_2(\mathbf{A}) = 1$,
- Szimmetrikus mátrixokra $\kappa(\mathbf{A}) \geq |\lambda_{\max}/\lambda_{\min}|$, továbbá $\kappa_2(\mathbf{A}) = |\lambda_{\max}/\lambda_{\min}|$, ahol $\lambda_{\max}$ és $\lambda_{\min}$ a legnagyobb és legkisebb abszolút értékű sajátértéket jelenti.

Bizonyítás. Az első állítás következik az indukált mátrixnormákra vonatkozó $1 = \|\mathbf{E}\| = \|\mathbf{A}\mathbf{A}^{-1}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$ becslésből. A második és harmadik állítás triviális. A negyedik állításban azt alkalmazhatjuk, hogy ortogonális mátrixok 2-es normája 1. A szimmetrikus mátrixokra vonatkozó ötödik állítás az 1.2.30. és 1.2.31. tételekből és abból következik, hogy reguláris mátrix inverzének sajátértékei az eredeti mátrix sajátértékeinek reciprokai. ■

3.2.3. megjegyzés. Természetesen az előző tételben a második és a harmadik tulajdonság nem csak indukált normára, hanem bármilyen mátrixnormára is teljesül. Tehát csak az első állításhoz és az ötödik állítás első részéhez használtuk fel, hogy a kondíciószámot indukált mátrixnormában mérjük. ⋄

3.2.4. példa. Nagyon nagy kondíciószámú mátrix pl. az ún. Hilbert-mátrix. Az $(n \times n)$-es Hilbert-mátrix elemeit $(\mathbf{H}_n)_{i,j} = 1/(i+j-1)$ módon definiáltak. Pl. a 6×6-os Hilbert mátrix alakja így

$$\mathbf{H}_6 = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 & 1/6 \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & 1/7 \\ 1/3 & 1/4 & 1/5 & 1/6 & 1/7 & 1/8 \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 & 1/9 \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 & 1/10 \\ 1/6 & 1/7 & 1/8 & 1/9 & 1/10 & 1/11 \end{bmatrix}$$

Erre a mátrixra $\kappa_2(\mathbf{H}_6) \approx 1.6 \times 10^7$, míg a (10×10)-es mátrixra $\kappa_2(\mathbf{H}_{10}) \approx 3.5 \times 10^{13}$. ⋄

Most térjünk át arra az esetre, amikor nemcsak a jobb oldali vektor, de maga az együttható-mátrix is változik. Először két kisebb tételt igazolunk.

3.2.5. tétel.
Legyen $\mathbf{S} = \mathbf{E} + \mathbf{R} \in \mathbb{R}^{n\times n}$, ahol $\|\mathbf{R}\| =: q < 1$ valamilyen indukált normában. Ekkor $\mathbf{S}$ reguláris, és

$$\|\mathbf{S}^{-1}\| \leq \frac{1}{1-q}.$$

Bizonyítás. Legyen $\overline{\mathbf{x}} \in \mathbb{R}^n$ egy tetszőleges vektor. Ekkor $\overline{\mathbf{x}} = \mathbf{S}\overline{\mathbf{x}} - \mathbf{R}\overline{\mathbf{x}}$, azaz $\|\overline{\mathbf{x}}\| \leq \|\mathbf{S}\overline{\mathbf{x}}\| + \|\mathbf{R}\overline{\mathbf{x}}\|$. Így

$$\|\mathbf{S}\overline{\mathbf{x}}\| \geq \|\overline{\mathbf{x}}\| - \|\mathbf{R}\overline{\mathbf{x}}\| \geq \|\overline{\mathbf{x}}\| - \|\mathbf{R}\| \cdot \|\overline{\mathbf{x}}\| = \|\overline{\mathbf{x}}\|(1 - \|\mathbf{R}\|) = \|\overline{\mathbf{x}}\|(1-q),$$

ami mutatja, hogy $\overline{\mathbf{x}} \neq \mathbf{0}$ esetén $\mathbf{S}\overline{\mathbf{x}} \neq \mathbf{0}$, azaz $\mathbf{S}$ reguláris. Emellett

$$\|\mathbf{S}^{-1}\| = \sup_{\overline{\mathbf{z}} \neq \mathbf{0}} \frac{\|\mathbf{S}^{-1}\overline{\mathbf{z}}\|}{\|\overline{\mathbf{z}}\|} \underbrace{=}_{\overline{\mathbf{z}} = \mathbf{S}\overline{\mathbf{x}}} \sup_{\overline{\mathbf{x}} \neq \mathbf{0}} \frac{\|\overline{\mathbf{x}}\|}{\|\mathbf{S}\overline{\mathbf{x}}\|} \leq \frac{1}{1-q}. \blacksquare$$

3.2.6. tétel.

Legyen $\mathbf{A} \in \mathbb{R}^{n \times n}$ reguláris mátrix, és tegyük fel, hogy $\|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\| < 1$ valamilyen indukált normában. Ekkor az $\mathbf{A} + \boldsymbol{\delta}\mathbf{A}$ mátrix is reguláris, és

$$\|(\mathbf{A} + \boldsymbol{\delta}\mathbf{A})^{-1}\| \leq \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\|}.$$

Bizonyítás. Mivel $\|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\| < 1$, így az előző tétel miatt $\mathbf{E} + \mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}$ reguláris. Vegyük észre, hogy

$$(\mathbf{A} + \boldsymbol{\delta}\mathbf{A})\left((\mathbf{E} + \mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A})^{-1}\mathbf{A}^{-1}\right) = (\mathbf{A} + \boldsymbol{\delta}\mathbf{A})\left(\mathbf{A}(\mathbf{E} + \mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A})\right)^{-1} = \mathbf{E}.$$

Ezért $\mathbf{A} + \boldsymbol{\delta}\mathbf{A}$ inverze az $(\mathbf{E} + \mathbf{A}^{-1}\boldsymbol{\delta})^{-1}\mathbf{A}^{-1}$ mátrix, és így

$$\|(\mathbf{A} + \boldsymbol{\delta}\mathbf{A})^{-1}\| \leq \|(\mathbf{E} + \mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A})^{-1}\| \cdot \|\mathbf{A}^{-1}\| \leq \frac{1}{1 - \|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\|} \cdot \|\mathbf{A}^{-1}\|. \blacksquare$$

3.2.7. tétel.

Tegyük fel, hogy az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszer helyett az $(\mathbf{A}+\boldsymbol{\delta}\mathbf{A})\overline{\mathbf{y}} = \overline{\mathbf{b}}+\boldsymbol{\delta}\overline{\mathbf{b}}$ perturbált egyenletrendszert oldjuk meg, és az együtthatómátrix perturbációjára teljesül a $\|\boldsymbol{\delta}\mathbf{A}\| < 1/\|\mathbf{A}^{-1}\|$ feltétel valamilyen indukált normában. Ekkor a perturbált egyenletrendszernek is egyértelmű megoldása van. Ezt a megoldást $\overline{\mathbf{y}} = \overline{\mathbf{x}} + \boldsymbol{\delta}\overline{\mathbf{x}}$ alakban írva érvényes az alábbi becslés:

$$\frac{\|\boldsymbol{\delta}\overline{\mathbf{x}}\|}{\|\overline{\mathbf{x}}\|} \leq \frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A})\|\boldsymbol{\delta}\mathbf{A}\|/\|\mathbf{A}\|} \cdot \left(\frac{\|\boldsymbol{\delta}\overline{\mathbf{b}}\|}{\|\overline{\mathbf{b}}\|} + \frac{\|\boldsymbol{\delta}\mathbf{A}\|}{\|\mathbf{A}\|}\right).$$

Bizonyítás. Mivel $\|\boldsymbol{\delta}\mathbf{A}\| < 1/\|\mathbf{A}^{-1}\|$, ezért $\|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\| < 1$. Így $\mathbf{A} + \boldsymbol{\delta}\mathbf{A}$ reguláris. Továbbá

$$\boldsymbol{\delta}\overline{\mathbf{x}} = (\mathbf{A} + \boldsymbol{\delta}\mathbf{A})^{-1}(\boldsymbol{\delta}\overline{\mathbf{b}} - \boldsymbol{\delta}\mathbf{A}\overline{\mathbf{x}}).$$

Alkalmazzuk az előző tételt.

$$\|\boldsymbol{\delta}\overline{\mathbf{x}}\| \leq \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\|}(\|\boldsymbol{\delta}\overline{\mathbf{b}}\| + \|\boldsymbol{\delta}\mathbf{A}\| \cdot \|\overline{\mathbf{x}}\|)$$

$$= \frac{\|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\|}{1 - \|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\|}\left(\frac{\|\boldsymbol{\delta}\overline{\mathbf{b}}\|}{\|\mathbf{A}\|} + \frac{\|\boldsymbol{\delta}\mathbf{A}\| \cdot \|\overline{\mathbf{x}}\|}{\|\mathbf{A}\|}\right).$$

Innét

$$\frac{\|\boldsymbol{\delta}\overline{\mathbf{x}}\|}{\|\overline{\mathbf{x}}\|} \leq \frac{\|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\|}{1 - \|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\|} \left(\frac{\|\boldsymbol{\delta}\overline{\mathbf{b}}\|}{\|\mathbf{A}\| \cdot \|\overline{\mathbf{x}}\|} + \frac{\|\boldsymbol{\delta}\mathbf{A}\|}{\|\mathbf{A}\|} \right)$$
$$\leq \frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A})\|\boldsymbol{\delta}\mathbf{A}\|/\|\mathbf{A}\|} \cdot \left(\frac{\|\boldsymbol{\delta}\overline{\mathbf{b}}\|}{\|\overline{\mathbf{b}}\|} + \frac{\|\boldsymbol{\delta}\mathbf{A}\|}{\|\mathbf{A}\|} \right). \blacksquare$$

Érdemes egy kicsit elidőzni a fenti tétel állításán. A benne szereplő képlet azt fejezi ki, hogy a megoldás relatív hibája függ az együtthatómátrix és a jobb oldal relatív hibájától. A képletben szorzótényezőként szerepel még viszont az együtthatómátrix kondíciószáma. Ez mutatja, hogy a megoldás relatív hibája úgy is nagy lehet, ha az együtthatómátrix és a jobb oldal hibája nem túl nagy, de az együtthatómátrix kondíciószáma nagy.

A gyakorlatban egy egyenletrendszer megoldása során az együtthatómátrix és a jobb oldali vektor több dolog miatt is perturbálódhat. Pl. amiatt, hogy az elemeik mérési adatok, így hibával terheltek. De még csak nem is kell hibával terheltnek lenniuk az adatoknak, hiszen magából a lebegőpontos ábrázolásból is származik hiba. Csak egy $\mathbf{A} + \boldsymbol{\delta}\mathbf{A}$ mátrixot és egy $\overline{\mathbf{b}} + \boldsymbol{\delta}\overline{\mathbf{b}}$ vektort tudunk ábrázolni az eredetiek helyett. Ezekre érvényesek a $\|\boldsymbol{\delta}\mathbf{A}\|_\infty \leq \mathtt{u}\|\mathbf{A}\|_\infty$ és $\|\boldsymbol{\delta}\overline{\mathbf{b}}\|_\infty \leq \mathtt{u}\|\overline{\mathbf{b}}\|_\infty$ becslések[1]. Tegyük fel, hogy $\kappa_\infty(\mathbf{A})\mathtt{u} \leq 1/2$. Ekkor még ha kerekítési hiba nélkül oldjuk is meg az egyenletrendszert, az $\hat{\overline{\mathbf{x}}}$ megoldásra érvényes az

$$\frac{\|\overline{\mathbf{x}} - \hat{\overline{\mathbf{x}}}\|_\infty}{\|\overline{\mathbf{x}}\|_\infty} \leq 4\mathtt{u}\kappa_\infty(\mathbf{A})$$

becslés. Ez a hiba nagyon nagy lehet, ha a kondíciószám nagy.

3.3. Gauss-módszer

A lineáris egyenletrendszerek megoldási módszereit két nagy csoportba oszthatjuk: ún. direkt és iterációs módszerek. Direkt módszerek esetén a megoldást véges sok aritmetikai művelettel állítjuk elő. Ha minden lépést pontosan számolunk, azaz a lépések nem terheltek a lebegőpontos számolásból származó hibákkal, akkor pontosan megkapjuk az egyenletrendszer megoldását. Ilyen megoldási módszer pl. a Cramer-szabály. Megjegyezzük persze, hogy ennek alkalmazása háromnál több egyenlet esetén egyáltalán nem praktikus. Iterációs módszerek esetén egy, a megoldáshoz tartó vektorsorozatot állítunk elő, és ennek egy megfelelő elemével közelítjük a megoldást.

Kisméretű lineáris egyenletrendszerek megoldását általában úgy végezzük, hogy valamelyik egyenletből kifejezzük valamelyik ismeretlent, és azt a többi egyenletbe helyettesítjük. Ezzel a módszerrel az ismeretlenek száma eggyel csökkenthető. Másik lehetőség, hogy valamelyik egyenlet egy ügyesen választott számszorosát kivonjuk egy másik egyenletből, kiejtve ezzel valamelyik ismeretlent. Felcserélhetünk egymással egyenleteket is, hiszen ezek sorrendjétől nem függ a megoldás. Felcserélhetjük a változókat is, csak az egyenletrendszer megoldásának felírásakor figyelembe kell majd vennünk ezt a cserét. Ezekkel a módszerekkel ismét csökkenthető az ismeretlenek száma. A felsorolt eljárások egyike sem változtatja meg az egyenletrendszer megoldását. Természetesen nagyméretű egyenletrendszerek megoldásához vagy egy lineáris egyenletrendszerek megoldására szolgáló program elkészítéséhez a fenti eljárásokat valamilyen szabálynak megfelelően kellene alkalmaznunk.

A legegyszerűbb, de máig nagyon sokszor használt, direkt lineáris egyenletrendszer megoldó módszer az ún. *Gauss-módszer*, amely lényegében a fent ismertetett eljárások szisztematikus leírása. A módszer két lépésből áll. Az első az ún. *eliminációs rész* (Gauss-elimináció), a második pedig a *visszahelyettesítés*. Az eliminációs lépéssel olyan alakra hozzuk az egyenletrendszert,

[1] $\mathtt{u}$ a gépi pontosságot jelöli.

melynek utolsó egyenletében csak az utolsó ismeretlen szerepel, az utolsó előttiben csak az utolsó kettő, stb. Ezen új egyenletrendszer együtthatómátrixa már egy felső háromszögmátrix lesz. Az eljárás után a megoldás az utolsó egyenlettől visszafelé egyszerű visszahelyettesítéssel nyerhető. Nézzük az eljárást most részletesen. Tegyük fel egyelőre, hogy az eljárás során sehol sem fordul elő nullával való osztás! Később majd megvizsgáljuk, hogy ha ez nem áll fenn, akkor hogy lehet a módszert módosítani.

Tegyük fel tehát, hogy az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszer megoldását keressük, ahol $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\overline{\mathbf{b}} \in \mathbb{R}^n$ és $\det(\mathbf{A}) \neq 0$. Az egyszerűség kedvéért csak az ismeretlenek együtthatóit kiírva az egyenletrendszer az alábbi alakú.

$$\left.\begin{array}{cccc} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{array}\right| \begin{array}{c} b_1 \\ b_2 \\ b_3 \\ \\ b_n \end{array}$$

Lássuk el az együtthatókat egy $^{(1)}$ felső indexszel, amely mutatja, hogy ez az elimináció során nyert első (azaz az eredeti) egyenletrendszer

$$\left.\begin{array}{cccc} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & \dots & a_{3n}^{(1)} \\ \vdots & & & \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{array}\right| \begin{array}{c} b_1^{(1)} \\ b_2^{(1)} \\ b_3^{(1)} \\ \\ b_n^{(1)} \end{array} .$$

Első lépésként az első egyenlet segítségével kiejtjük a többi egyenletből az első változót. Ezt úgy érjük el, hogy az első egyenlet egy számszorosát kivonjuk a megfelelő egyenletből. Legyen tehát $l_{21} = a_{21}^{(1)}/a_{11}^{(1)}, \dots, l_{n1} = a_{n1}^{(1)}/a_{11}^{(1)}$. Ekkor könnyű látni, hogy az i. egyenletből kivonva az első egyenlet l_{i1}-szeresét az i. egyenletből kiesik az első változó. Az l_{ij} alakú szorzókat úgy indexeljük, hogy l_{ij} az i. sor j. elemének kinullázásához használt szorzót jelentse. Így az

$$\left.\begin{array}{cccc} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} - l_{21}a_{12}^{(1)} & \dots & a_{2n}^{(1)} - l_{21}a_{1n}^{(1)} \\ 0 & a_{32}^{(1)} - l_{31}a_{12}^{(1)} & \dots & a_{3n}^{(1)} - l_{31}a_{1n}^{(1)} \\ \vdots & & & \\ 0 & a_{n2}^{(1)} - l_{n1}a_{12}^{(1)} & \dots & a_{nn}^{(1)} - l_{n1}a_{1n}^{(1)} \end{array}\right| \begin{array}{c} b_1^{(1)} \\ b_2^{(1)} - l_{21}b_1 \\ b_3^{(1)} - l_{31}b_1 \\ \\ b_n^{(1)} - l_{n1}b_1 \end{array} .$$

egyenletrendszerhez jutunk. A könnyebb átláthatóság kedvéért lássuk el az elemeket most $^{(2)}$ felső indexszel. Mivel az első sor elemei nem változnak, így ott megtartjuk az $^{(1)}$-es indexet.

$$\left.\begin{array}{cccc} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ 0 & a_{32}^{(2)} & \dots & a_{3n}^{(2)} \\ \vdots & & & \\ 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{array}\right| \begin{array}{c} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(2)} \\ \\ b_n^{(2)} \end{array} .$$

Most kiszámítva az $l_{32} = a_{32}^{(2)}/a_{22}^{(2)}, \ldots, l_{n2} = a_{n2}^{(2)}/a_{22}^{(2)}$ szorzókat, hasonlóan lenullázhatjuk a második oszlop főátló alatti elemeit is, majd átindexelés után kapjuk az új

$$\left.\begin{array}{cccc} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} \\ 0 & 0 & \ldots & a_{3n}^{(3)} \\ \vdots & & & \\ 0 & 0 & \ldots & a_{nn}^{(3)} \end{array}\right| \begin{array}{c} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \\ b_n^{(3)} \end{array}$$

egyenletrendszert. Ezt folytatjuk addig, míg az utolsó előtti főátlóbeli elem alatti egyetlen elem is le nem nullázódik. Ekkor a megfelelő indexelést elvégezve kapjuk az

$$\left.\begin{array}{cccc} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} \\ 0 & 0 & \ldots & a_{3n}^{(3)} \\ \vdots & & & \\ 0 & 0 & \ldots & a_{nn}^{(n)} \end{array}\right| \begin{array}{c} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \\ b_n^{(n)} \end{array}$$

egyenletrendszert. Ezzel az eliminációs rész végére értünk.

A fenti egyenletrendszer együtthatómátrixa egy felső háromszögmátrix. Megoldása egyszerű visszahelyettesítéssel történik. Az utolsó egyenletből nyerjük az x_n ismeretlen értékét az

$$x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}}$$

formulával. Ezután visszafelé haladva rendre az

$$\begin{aligned} x_{n-1} &= \frac{b_{n-1}^{(n-1)} - a_{n-1,n}^{(n-1)} x_n}{a_{n-1,n-1}^{(n-1)}}, \\ &\vdots \\ x_2 &= \frac{b_2^{(2)} - x_n a_{2n}^{(2)} - \cdots - x_3 a_{23}^{(2)}}{a_{22}^{(2)}}, \\ x_1 &= \frac{b_1^{(1)} - x_n a_{1n}^{(1)} - \cdots - x_2 a_{12}^{(1)}}{a_{11}^{(1)}} \end{aligned}$$

értékeket kapjuk.

Azokat a főátlóbeli elemeket, melyek segítségével lenullázzuk az alattuk lévő elemeket, *főelemeknek* nevezzük. Egy adott főelem alatti elemek kinullázását felírhatjuk mátrixszorzás segítségével is. Legyen $\bar{\mathbf{l}}_k = [0, \ldots, 0, l_{k+1,k}, \ldots, l_{n,k}]^T \in \mathbb{R}^n$ $(k = 1, \ldots, n-1)$. Ekkor a Gauss-elimináció k-adik lépése úgy változtatja meg a mátrixot, mintha az $\mathbf{L}_k := \mathbf{E} - \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T$ mátrixszal szoroztuk volna azt. Az $\mathbf{L}_k$ $(k = 1, \ldots, n-1)$ mátrixszal való szorzást *Gauss-transzformációnak* nevezzük. A teljes eliminációs eljárás során létrejövő felső háromszögmátrix $\mathbf{L}_{n-1} \ldots \mathbf{L}_2 \mathbf{L}_1 \mathbf{A}$ alakban írható. Ezt a mátrixot általában $\mathbf{U}$-val jelöljük, utalva arra, hogy ez egy felső (*u*pper) háromszögmátrix.

3.3.1. tétel.
A Gauss-transzformáció $\mathbf{L}_k = \mathbf{E} - \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T$ mátrixa invertálható, és $\mathbf{L}_k^{-1} = \mathbf{E} + \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T$, $k = 1, \ldots, n-1$.

Bizonyítás. Azt kell megmutatni, hogy $(\mathbf{E} - \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T)(\mathbf{E} + \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T) = \mathbf{E}$. Elvégezve a szorzást azt kapjuk, hogy

$$(\mathbf{E} - \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T)(\mathbf{E} + \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T) = \mathbf{E} - \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T + \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T + \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T = \mathbf{E} + \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T = \mathbf{E}.$$

Az utolsó lépés onnét következik, hogy az $\bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T$ mátrix első k sora csupa nulla elemből áll, és a k. oszloptól jobbra is csak nulla elemek szerepelnek. Így a mátrix négyzete nullmátrixot ad. ■

Most vizsgáljuk meg, hogy milyen mátrixok esetén hajtható végre a Gauss-módszer a fent ismertetett módon, azaz úgy, hogy közben egyik főelem sem lesz nulla. Ekkor ugyanis a nullával való osztás miatt nem tud végigfutni az eljárás.

3.3.2. tétel.
A Gauss-módszer pontosan akkor hajtható végre, ha az $\mathbf{A}$ mátrix egyik főminorja sem zérus, azaz $\det(\mathbf{A}(1:k, 1:k)) \neq 0$ $(k = 1, \ldots, n)$.

Bizonyítás. A Gauss-elimináció során az egyes sorokból kivonjuk más sorok számszorosait. Ez az eljárás nem változtatja meg a determinánst. Tehát

$$\begin{aligned}
\det(\mathbf{A}(1:1,1:1)) &= \det(\mathbf{A}^{(1)}(1:1,1:1)) = a_{11}^{(1)} \neq 0,\\
\det(\mathbf{A}(1:2,1:2)) &= \det(\mathbf{A}^{(2)}(1:2,1:2)) = a_{11}^{(1)} a_{22}^{(2)} \neq 0,\\
&\vdots\\
\det(\mathbf{A}(1:n,1:n)) &= \det(\mathbf{A}^{(n)}(1:n,1:n)) = a_{11}^{(1)} a_{22}^{(2)} \ldots a_{nn}^{(n)} \neq 0.
\end{aligned}$$

Az utolsó feltétel a visszahelyettesítés miatt kell, hiszen ez úgy kezdődik, hogy $a_{nn}^{(n)}$-el osztanunk kell az x_n ismeretlen kiszámításához. Ebből következik az állítás. ■.

3.3.3. tétel.
Ha az $\mathbf{A}$ mátrix

1. szigorúan domináns főátlójú,
2. szimmetrikus, pozitív definit mátrix,
3. M-mátrix,

akkor a Gauss-módszer végrehajtható az előző algoritmussal.

Bizonyítás. 1. Legyen $\mathbf{A}$ először szigorúan domináns főátlójú. Először megmutatjuk, hogy az eliminációs eljárás ez a tulajdonság megőrződik. Elegendő ezt az első ill. i-edik sorra megmutatni. Tegyük fel tehát, hogy az első és i-edik sorban domináns a főátló, azaz

$$|a_{11}| > |a_{12}| + |a_{13}| + \cdots + |a_{1n}| \quad / \cdot |a_{i1}|$$

és

$$|a_{ii}| > |a_{i1}| + \cdots + |a_{i,i-1}| + |a_{i,i+1}| + \cdots + |a_{in}| \quad / \cdot |a_{11}|.$$

Ezekből

$$|a_{11}a_{i1}| \geq |a_{12}a_{i1}| + |a_{13}a_{i1}| + \cdots + |a_{1n}a_{i1}|,$$

$$|a_{ii}a_{11}| > |a_{i1}a_{11}| + \cdots + |a_{i,i-1}a_{11}| + |a_{i,i+1}a_{11}| + \cdots + |a_{in}a_{11}|$$

és

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

Az új mátrix i-edik sorának i-edik eleme: $(a_{ii}a_{11} - a_{i1}a_{1i})/a_{11}$.

$$|a_{ii}a_{11} - a_{i1}a_{1i}| \geq ||a_{ii}a_{11}| - |a_{i1}a_{1i}||$$

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

Végül az $|a_{11}|$ értékkel való osztás mutatja a dominanciát.

2. Legyen most $\mathbf{A}$ szimmetrikus, pozitív definit mátrix. Szimmetrikus, pozitív definit mátrix determinánsa pozitív (sajátértékek szorzata). Legyen $\overline{\mathbf{x}}_k$ olyan nemnulla vektor, melynek a $k+1 : n$ elemei nullák. Ezzel $\overline{\mathbf{x}}_k^T \mathbf{A} \overline{\mathbf{x}}_k > 0$, ami mutatja, hogy a k-adik minormátrix pozitív definit, azaz determinánsa pozitív. Tehát nem nulla.

3. Legyen végül $\mathbf{A}$ M-mátrix. Ekkor van olyan $\overline{\mathbf{g}} > 0$ vektor, hogy $\mathbf{A}\overline{\mathbf{g}} > 0$. Ismét azt mutatjuk meg, hogy a Gauss-elimináció egy lépése után a mátrix továbbra is M-mátrix marad. Elég ezt belátni az első transzformációra. Az első transzformáció után a főátlón kívüli elemek továbbra sem lehetnek pozitívak. Továbbá $\mathbf{A}^{-1}\overline{\mathbf{g}} > 0$ egy pozitív vektor, mellyel $\mathbf{L}_1\mathbf{A}\mathbf{A}^{-1}\overline{\mathbf{g}} = \mathbf{L}_1\overline{\mathbf{g}} > 0$, hiszen $\mathbf{L}_1$ nemnegatív mátrix csupa egyessel a főátlójában. ■

Vizsgáljuk meg a Gauss-módszer egyes lépéseinek műveletszámát! A műveletszámok ismerete a későbbiekben segít abban, hogy az adott problémát mindig a lehető leggyorsabban oldjuk meg.

Kezdjük az elimináció műveletigényével. Tegyük fel, hogy a k-adik főelemmel nullázunk. Ekkor ki kell számolnunk az $l_{k+1,k}, \ldots, l_{n,k}$ szorzókat ($n-k$ flop). A k-adik sor $k+1 : n+1$ elemét szoroznunk kell a kiszámított szorzókkal ($n-k+1$ flop), majd ki kell vonni azt az i-edik sorból ($i = k+1, \ldots, n$) (egyenként $n-k+1$ flop). Ez összesen $2(n-k+1)(n-k)$ flop. Ezután a műveletszámokat összeadjuk az összes főelemre. Így az elimináció műveletszáma

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

$$= \frac{2(n-1)n(2n-1)}{6} + \frac{3(n-1)n}{2} = \frac{4n^3 + 3n^2 - 7n}{6} = \frac{2}{3}n^3 + O(n^2) \text{ flop.}$$

Itt felhasználtuk az első n természetes szám összegére és négyzetösszegére vonatkozó összegképleteket. A visszahelyettesítés műveletigényéről könnyen láthatóan $1 + 3 + \cdots + 2n - 1 = n^2$ flop.

A Gauss-módszer teljes műveletigénye tehát

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

Vegyük észre, hogy nagy méretű egyenletrendszerek esetén a visszahelyettesítés műveletigénye elhanyagolható az eliminációéhoz képest.

Ingamódszer

A gyakorlatban sokszor olyan egyenletrendszereket kell megoldanunk, melyek együtthatómátrixa tridiagonális. Ilyen feladattal találkozunk pl. a spline-interpolációnál ill. a differenciálegyenletek numerikus megoldása esetén.

Legyen tehát a megoldandó lineáris egyenletrendszer

$$\begin{bmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & a_n & b_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n \end{bmatrix}$$

alakú, ahol az a_i, b_i, c_i, f_i értékek adottak, és az x_i értékeket keressük. A Gauss-eliminációt alkalmazva látjuk, hogy nem szükséges az eliminációt minden oszlopon teljesen végigfuttatni, hiszen a szubdiagonál alatt csak nulla elemek vannak, hanem elegendő csak a szubdiagonálbeli elemeket lenullázni. (Ezzel várható, hogy az n^3 nagyságrendű eliminációs műveletszám csak n nagyságrendű lesz.) Most az elimináció során a főelemmel leosztjuk a sorát is. Ez ugyan két osztást jelent soronként, de így minden sorban csak két előre nem ismert elem lesz a három helyett, továbbá a főátlóban 1-esek fognak állni, így a visszahelyettesítés során nem kell ezekkel az elemekkel osztani. Az első $i-2$ oszlop eliminációja után az egyenletrendszer az

$$\begin{bmatrix} 1 & -\alpha_2 & & & & & & \\ 0 & 1 & -\alpha_3 & & & & & \\ 0 & 0 & \ddots & \ddots & \ddots & & & \\ \vdots & \vdots & & 1 & -\alpha_i & & & \\ & & & a_i & b_i & c_i & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & \dots & & & & a_n & b_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} \beta_2 \\ \beta_3 \\ \vdots \\ \beta_i \\ f_i \\ \vdots \\ f_{n-1} \\ f_n \end{bmatrix}$$

alakú lesz, ahol α_i, β_i megfelelő konstansok. Nyilvánvalóan

$$\alpha_2 = -\frac{c_1}{b_1}, \quad \beta_2 = \frac{f_1}{b_1}$$

és az elimináció ill. a főelemmel való osztás miatt

$$\alpha_{i+1} = \frac{-c_i}{a_i\alpha_i + b_i}, \quad \beta_{i+1} = \frac{f_i - a_i\beta_i}{a_i\alpha_i + b_i}. \tag{3.3.1}$$

A fenti két képlet segítségével az α_i, β_i együtthatók meghatározhatók, sőt ha bevezetjük az $\alpha_1 = \beta_1 = 0$ kezdeti értékeket, akkor az is könnyen megjegyezhető, hogy honnét kell indítani a (3.3.1) rekurziót.

Az elimináció után tehát az

$$\begin{bmatrix} 1 & -\alpha_2 & & & \\ & 1 & -\alpha_3 & & \\ & & \ddots & \ddots & \\ & & & 1 & -\alpha_n \\ & & & & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} \beta_2 \\ \beta_3 \\ \vdots \\ \beta_n \\ (f_n - a_n\beta_n)/(b_n + \alpha_n a_n) \end{bmatrix}$$

egyenletrendszerhez jutunk. A visszahelyettesítésnél innét nyilvánvalóan

$$x_n = \frac{f_n - a_n\beta_n}{a_n\alpha_n + b_n},$$

majd a többi ismeretlen pedig az $x_{i-1} = \alpha_i x_i + \beta_i$ $(i = n : -1 : 2)$ képlettel nyerhető.

A most tridiagonális mátrixokra ismertetett lineáris egyenletrendszer megoldási módot ingamódszernek vagy Thomas[2]-algoritmusnak vagy egyszerűsített Gauss-módszernek nevezzük. Az algoritmus tehát a korábbi jelölésekkel az alábbi módon foglalható össze, melyből az is látszik, hogy az algoritmus $8n - 3$ flop műveletet igényel.

Inga-módszer
$\alpha_1 := 0$, $\beta_1 := 0$
for i:=1:n-1 **do**
  $\alpha_{i+1} := -c_i/(a_i\alpha_i + b_i)$
  $\beta_{i+1} := (f_i - a_i\beta_i)/(a_i\alpha_i + b_i)$
end for
$y_n := (f_n - a_n\beta_n)/(a_n\alpha_n + b_n)$
for i:=n:-1:2 **do**
  $y_{i-1} := \alpha_i y_i + \beta_i$
end for

Mivel az ingamódszer tulajdonképpen olyan Gauss-módszer, amely kihasználja, hogy a mátrix tridiagonális, így biztosan elakadás nélkül végigfut a 3.3.3. tételben felsorolt mátrixokra. Ezen belül, ha a mátrix szigorúan domináns főátlójú, akkor mivel ez a tulajdonság az elimináció során is megmarad, igaz lesz, hogy $|\alpha_i| < 1$ $(i = 2, \ldots, n)$.

3.4. LU-felbontás

Az előző fejezetben láttuk, hogy bizonyos feltételek mellett egy $\mathbf{A}$ mátrix Gauss-transzformációkkal felső háromszögmátrix alakra hozható. Ezt a tulajdonságot bővíti ki az alábbi tétel.

3.4.1. tétel. (LU-felbontás)
Tegyük fel, hogy az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrixra $\det(\mathbf{A}(1:k, 1:k)) \neq 0$ $(k = 1, \ldots, n-1)$, azaz a Gauss-elimináció végrehajtható vele. Ekkor létezik egy olyan $\mathbf{L}$ normált (főátlóban egyesek szerepelnek) alsó (lower) háromszögmátrix és egy $\mathbf{U}$ felső (upper) háromszögmátrix, melyekkel $\mathbf{A} = \mathbf{LU}$. Ha egy reguláris mátrixnak létezik LU-felbontása, akkor az LU-felbontása egyértelmű.

Bizonyítás. A Gauss-elimináció során a Gauss-transzformációk az $\mathbf{A}$ mátrixot felső háromszögmátrix alakúra hozzák. Legyen ez az $\mathbf{U}$ mátrix. Így tehát

$$\mathbf{L}_{n-1}\mathbf{L}_{n-2}\ldots\mathbf{L}_1\mathbf{A} = \mathbf{U}.$$

[2] Llewellyn Thomas (1903–1992) angol fizikus. A módszert többen is kitalálták a múlt század közepén. Thomas a [36] cikkben említi. Bruce [5] cikke az első, amelyik széles körhöz eljutó folyóiratban jelenteti meg a módszert. David Young hívja először Thomas-algoritmusnak. Magyarul szokás ingamódszernek is nevezni a módszert. Az elnevezés a szorzók és az ismeretlenek meghatározási sorrendjére utal.

Mivel $(\mathbf{E}-\bar{\mathbf{l}}_k\bar{\mathbf{e}}_k^T)^{-1}=\mathbf{E}+\bar{\mathbf{l}}_k\bar{\mathbf{e}}_k^T$, és $\bar{\mathbf{l}}_k\bar{\mathbf{e}}_k^T\bar{\mathbf{l}}_l\bar{\mathbf{e}}_l^T=\mathbf{0}$ ha $l>k$, az $\mathbf{A}$ mátrix az alábbi alakban írható

$$\mathbf{A}=\mathbf{L}_1^{-1}\dots\mathbf{L}_{n-2}^{-1}\mathbf{L}_{n-1}^{-1}\mathbf{U}=\left(\prod_{k=1}^{n-1}(\mathbf{E}+\bar{\mathbf{l}}_k\bar{\mathbf{e}}_k^T)\right)\mathbf{U}$$

$$=\underbrace{\left(\mathbf{E}+\sum_{k=1}^{n-1}\bar{\mathbf{l}}_k\bar{\mathbf{e}}_k^T\right)}_{=:\mathbf{L},\text{alsó normált háromszögmátrix}}\mathbf{U}=\mathbf{L}\mathbf{U}.$$

Az egyértelműség igazolásához tegyük fel, hogy van két különböző LU-felbontása is az $\mathbf{A}$ invertálható mátrixnak: $\mathbf{A}=\tilde{\mathbf{L}}\tilde{\mathbf{U}}=\mathbf{L}\mathbf{U}$. Ekkor

$$\tilde{\mathbf{L}}^{-1}\mathbf{L}=\tilde{\mathbf{U}}\mathbf{U}^{-1}=\mathbf{E},$$

mivel normált alsó háromszögmátrixok szorzata normált alsó háromszögmátrix, a felsőké felső háromszögmátrix. ■

3.4.2. következmény. Ha egy reguláris mátrixnak valamelyik főminorja nulla, akkor nincs LU-felbontása. ◇

3.4.3. megjegyzés. Az LU-felbontásban tehát $\mathbf{U}$ az elimináció során kialakuló felső háromszögmátrix, $\mathbf{L}$ pedig az l_{ij} szorzókból az

$$\mathbf{L}=\begin{bmatrix}1 & 0 & \dots & 0\\ l_{21} & 1 & \dots & 0\\ l_{31} & l_{32} & \dots & 0\\ \vdots & & & \\ l_{n1} & l_{n2} & \dots & 1\end{bmatrix}$$

módon készült mátrix. ◇

Az LU-felbontás alkalmazása

Az *LU*-felbontást a korábbiak alapján a következő esetekben érdemes kihasználni.

- Determináns. A mátrix determinánsa megkapható, ha az $\mathbf{U}$ mátrix főátlóbeli elemeit összeszorozzuk.
- Lineáris egyenletrendszer megoldása. Ha már kiszámoltuk egy mátrix LU-felbontását, akkor $\mathbf{L}$ és $\mathbf{U}$ tárolható $\mathbf{A}$ helyén a számítógép memóriájában, hiszen az $\mathbf{L}$ mátrix főátlójában álló egyeseket nem kell eltárolnunk. Ekkor az $\mathbf{A}\overline{\mathbf{x}}=\mathbf{L}(\mathbf{U}\overline{\mathbf{x}})=\overline{\mathbf{b}}$ egyenletrendszert már két háromszögmátrixú egyenletrendszer megoldásával meg tudjuk oldani. Így a műveletszám $2n^2$ flop lesz, ami nagyságrenddel kisebb, mint a Gauss-módszer $2n^3/3$ műveletigénye. Így tehát, ha több olyan lineáris egyenletrendszert kell megoldanunk, melyek csak a jobb oldali vektorban különböznek, akkor érdemes az első megoldása után eltárolni az LU-felbontás mátrixait, és a többi egyenletrendszert már a fenti módszerrel megoldani.
- Inverz mátrix. Tipikus példa a fenti esetre az, amikor egy mátrix inverzét kell meghatároznunk. Ekkor először meghatározhatjuk az LU-felbontást, ami $2n^3/3+\mathcal{O}(n^2)$ művelet, majd

pedig megoldunk n darab lineáris egyenletrendszert, melyek jobb oldalai az egységvektorok. Ez további $2n^3$ művelet, ami összesen $8n^3/3+\mathcal{O}(n^2)$ műveletet ad. Ha előre ismert az LU-felbontás, akkor csak $2n^3$ a műveletszám.

Ha nem ismert előre az LU-felbontás, akkor a fentinél jobb műveletszámot érhetünk el, nevezetesen $2n^3+\mathcal{O}(n^2)$-et, ha a Gauss-elimináció helyett az ún. Gauss–Jordan[3]-eliminációt alkalmazzuk, ami lineáris egyenetrendszer megoldásakor nem praktikus, de mátrixinverz számítás esetén igen (3.11.7. feladat). Lényege, hogy a főelemmel nemcsak a főátló alatt, hanem a főátló felett is nullázunk. Megjegyezzük még, hogy egy mátrix inverzét, annak nagy műveletigénye miatt, a gyakorlatban csak akkor számítjuk ki, ha explicit módon szükségünk van az inverz mátrix egyes elemeire.

3.5. Főelemkiválasztás, általános LU-felbontás, Cholesky-felbontás

3.5.1. Főelemkiválasztás

Eddig a Gauss-módszernél és az LU-felbontásnál feltettük, hogy az eljárás végrehajtható, azaz nem fordul elő olyan eset, hogy a főelem nulla lenne. Most nézzük meg azt az esetet, amikor ez a feltétel mégsem teljesül. Mit tehetünk ebben az esetben? Ha a k-adik oszlop eliminációjánál tartunk, és az $a_{kk}^{(k)}$ főelem nulla lenne, akkor a k-adik egyenletet cseréljük ki egy olyan nagyobb sorszámú $(k+1,\dots,n)$ egyenlettel, melyben a k-adik ismeretlenhez tartozó együttható nem nulla. Az ilyen csere biztosan végrehajtható, mert különben az együtthatómátrix k. oszlopa előállna az első $k-1$ oszlop lineáris kombinációjaként. Ez pedig nem fordulhat elő, mert $\det\mathbf{A}$-ról feltettük, hogy nem nulla.

Az egyenletcserének nem csak akkor van jelentősége, ha különben nem futna végig az eliminációs eljárás. Eddig nem vettük ugyanis figyelembe, hogy a Gauss-módszert általában számítógépek segítségével hajtjuk végre, azaz a számolásnál lebegőpontos számrendszert használunk. Milyen hatása van vajon ennek a számított megoldásra? Alkalmazva a 2.5.2. tétel eredményét a számábrázolás hibájáról, az egyenletrendszer együtthatómátrixának egy a_{ij} elemére $|fl(a_{ij})-a_{ij}|\le \mathtt{u}|a_{ij}|$, ahol $\mathtt{u}$ a gépi pontosság. Bevezetve egy $\mathbf{A}=[a_{ij}]\in\mathbb{R}^{n\times n}$ mátrix esetén az $|\mathbf{A}|\in\mathbb{R}^{n\times n}$ jelölést az $[|a_{ij}|]$ mátrixra, a fentiek szerint a

$$|fl(\mathbf{A})-\mathbf{A}|\le \mathtt{u}|\mathbf{A}|$$

becsléshez jutunk. Az alábbi, az LU-felbontás hibájára vonatkozó tételt bizonyítás nélkül közöljük [18].

3.5.1. tétel. (Golub, van Loan [18], 105. oldal)
Tegyük fel, hogy egy $\mathbf{A}\in\mathbb{R}^{n\times n}$ mátrix LU-felbontását készítjük el egy olyan számítógépen, amely $\mathtt{u}$ gépi pontosságú lebegőpontos számokat használ. Tegyük fel továbbá, hogy az elimináció során – mely alatt egyik főelem sem lett nulla – az $\hat{\mathbf{L}}$ és $\hat{\mathbf{U}}$ mátrixokhoz jutottunk, melyekkel $\hat{\mathbf{L}}\hat{\mathbf{U}}-\mathbf{A}=\boldsymbol{\delta}\mathbf{A}$. Ekkor érvényes a következő becslés:

$$|\boldsymbol{\delta}\mathbf{A}|\le 3(n-1)\mathtt{u}(|\mathbf{A}|+|\hat{\mathbf{L}}|\cdot|\hat{\mathbf{U}}|)+\mathcal{O}(\mathtt{u}^2).$$

A fenti becslésben szerepel az $|\hat{\mathbf{L}}|$ mátrix, így az egyik célunk az lehet, hogy az elimináció során ennek elemei a lehető legkisebbek legyenek. Arról, hogy ennek elérése során $\hat{\mathbf{U}}$ elemei sem

[3] Wilhelm Jordan (1842–1899) német geodéta. Nem tévesztendő össze (kiejtésben sem) a francia Camille Jordannal, akinek nevéhez pl. a Jordan-mérték elnevezés fűződik.

válnak túl naggyá, később lesz szó. Mivel $\hat{\mathbf{L}}$ elemeit úgy kapjuk, hogy a főelemmel osztjuk le az egyes oszlopok főátló alatti elemeit, így ezek az elemek 1-nél nem nagyobb abszolútértékűek, ha főelemnek az $\mathbf{A}^{(k)}(k+1:n,k)$ vektor legnagyobb abszolútértékű elemét választjuk.

Annak érdekében tehát, hogy minél pontosabb megoldást kapjunk, a sorcserét akkor is érdemes végrehajtani, ha a főelemek nem nullák. Azt az eljárást, amikor a k-adik lépésben a főelemet az $\mathbf{A}^{(k)}(k+1:n,k)$ oszlop legnagyobb abszolútértékű elemének választjuk, majd ezen elem sorát felcseréljük a k. sorral, részleges főelemkiválasztásnak hívjuk. A részleges főelemkiválasztáshoz összesen $(n^2-n)/2$ összehasonlítás szükséges.

Még jobban csökkenthető az $\hat{\mathbf{L}}$ mátrix elemeinek abszolútértéke, ha a legnagyobb abszolútértékű elemet a k-adik lépésben nem csak az oszlopban, hanem a teljes $\mathbf{A}^{(k)}(k:n,k:n)$ blokkban keressük. Ezután az így megtalált új főelem sorát és oszlopát cseréljük a k-adik sorral és oszloppal. Ezt az eljárást teljes főelemkiválasztásnak nevezzük. Ez $n(n+1)(2n+1)/6-1=n^3/3+O(n^2)$ összehasonlítással jár, ami – ha elfogadjuk, hogy egy összehasonlítás kb. ugyanannyi ideig tart, mint egy lebegőpontos művelet – már a Gauss-módszer nagyságrendjébe esik. Emiatt a gyakorlatban általában a részleges főelemkiválasztást alkalmazzák, a teljes főelemkiválasztást csak akkor érdemes használni, ha nagy pontosságú megoldásra van szükségünk.

3.5.2. példa. Tekintsük az alábbi feladatot, ahol tegyük fel, hogy a megoldás során mindig 4 számjegyre kerekítünk.

$$\begin{aligned} 0.003x_1 + 59.14x_2 &= 59.17 \\ 5.291x_1 - 6.13x_2 &= 46.78 \end{aligned}$$

A pontos megoldás $x_1 = 10.00$, $x_2 = 1.000$. Főelemkiválasztás nélkül a megoldás $x_1 = -10$, $x_2 = 1.001$ (a kiegyszerűsödés miatt hamis eredményt kapunk), részleges főelemkiválasztással a pontos eredményt kapjuk. $\diamond$

3.5.2. Általános LU-felbontás

Az LU-felbontás csak olyan $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrixokra végezhető el, melyeknek minden $1,2,\ldots,n-1$-ed rendű bal felső sarokdeterminánsa különbözik nullától. Az alábbi tétel azt mutatja, hogy minden $\mathbf{A}$ mátrix sorai átrendezhetők úgy, hogy az így keletkező mátrixnak legyen már LU-felbontása.

3.5.3. tétel. (LU-felbontás általános mátrixra)
Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy tetszőleges mátrix. Ekkor létezik egy olyan $\mathbf{L}$ alsó normált háromszögmátrix, melynek elemei egynél nem nagyobb abszolút értékűek, egy $\mathbf{U}$ felső háromszögmátrix, és egy $\mathbf{P}$ permutációs mátrix, melyekkel $\mathbf{PA} = \mathbf{LU}$.

Bizonyítás. Hajtsuk végre az alábbi eljárást az $\mathbf{A}$ mátrixszal. Az eljárás minden mátrixszal végrehajtható.

- Legyen $\mathbf{A}^{(1)} = \mathbf{A}$, és $k = 1$.
- Ha $\mathbf{A}^{(k)}(k:n,k) = \mathbf{0}$, akkor legyen $\mathbf{P}_k = \mathbf{L}_k = \mathbf{E}$, különben válasszuk ki a vektor nem nulla elemei közül a legnagyobb abszolútértékűt (legyen ez $a_{sk}^{(k)}$), és definiáljuk a

$$\mathbf{P}_k = [\bar{\mathbf{e}}_1,\ldots,\bar{\mathbf{e}}_{k-1},\bar{\mathbf{e}}_s,\bar{\mathbf{e}}_{k+1},\ldots,\bar{\mathbf{e}}_{s-1},\bar{\mathbf{e}}_k,\bar{\mathbf{e}}_{s+1},\ldots,\bar{\mathbf{e}}_n]^T$$

permutációs mátrixot, valamint legyen $\mathbf{L}_k$ a $\mathbf{P}_k\mathbf{A}^{(k)}$ mátrix k-adik oszlopához tartozó Gauss-transzformációs mátrix.

- Legyen $\mathbf{A}^{(k+1)} = \mathbf{L}_k\mathbf{P}_k\mathbf{A}^{(k)}$.
- Ezután legyen $k = k+1$, és járjunk el hasonlóan, míg $k = n$ nem lesz.

Végrehajtva az előző lépéseket, előáll egy

$$\mathbf{U} = \mathbf{L}_{n-1}\mathbf{P}_{n-1}\dots\mathbf{L}_1\mathbf{P}_1\mathbf{A}$$

felső háromszögmátrix. Mivel $\mathbf{P}_i^2 = \mathbf{E}$ (hiszen $\mathbf{P}_i$ az i-edik sort cseréli egy nagyobb sorszámú sorral, így ha kétszer alkalmazzuk, akkor az egységmátrixot (identitást) nyerjük), ezért

$$\bar{\mathbf{L}}_{n-1}\bar{\mathbf{L}}_{n-2}\dots\bar{\mathbf{L}}_1\mathbf{P}_{n-1}\mathbf{P}_{n-2}\dots\mathbf{P}_1\mathbf{A} = \mathbf{U},$$

ahol $\bar{\mathbf{L}}_{n-1} = \mathbf{L}_{n-1}$ és

$$\bar{\mathbf{L}}_k = \mathbf{P}_{n-1}\mathbf{P}_{n-2}\dots\mathbf{P}_{k+1}\mathbf{L}_k\mathbf{P}_{k+1}\dots\mathbf{P}_{n-2}\mathbf{P}_{n-1}.$$

Mivel $\mathbf{L}_k = \mathbf{E} - \bar{\mathbf{l}}_k\bar{\mathbf{e}}_k^T$, és $\mathbf{P}_k$ a k-adik sort vagy oszlopot cseréli egy nagyobb sorszámúra, ezért

$$\bar{\mathbf{L}}_k = \mathbf{E} - \underbrace{\mathbf{P}_{n-1}\mathbf{P}_{n-2}\dots\mathbf{P}_{k+1}\bar{\mathbf{l}}_k}_{=:\bar{\mathbf{l}}_k^\star,\ \bar{\mathbf{l}}_k(k+1:n)\text{ elemeinek cseréivel}}\quad\underbrace{\bar{\mathbf{e}}_k^T\mathbf{P}_{k+1}\dots\mathbf{P}_{n-2}\mathbf{P}_{n-1}}_{\bar{\mathbf{e}}_k^T}.$$

Az $\bar{\mathbf{L}}_k$ mátrixok $\mathbf{E} - \bar{\mathbf{l}}_k^\star\bar{\mathbf{e}}_k^T$ alakúak, így inverzei $\bar{\mathbf{L}}^{-1} = \mathbf{E} + \bar{\mathbf{l}}_k^\star\bar{\mathbf{e}}_k^T$ alakban írhatók, melyek szorzata normált alsó háromszögmátrix lesz. Legyen

$$\mathbf{L} = \bar{\mathbf{L}}_1^{-1}\dots\bar{\mathbf{L}}_{n-1}^{-1}.$$

A részleges főelemkiválasztásból következik, hogy $\mathbf{L}$-nek nincs 1-nél nagyobb abszolútértékű eleme (emiatt $\|\mathbf{L}\|\infty \le n$). Továbbá legyen

$$\mathbf{P} = \mathbf{P}_{n-1}\dots\mathbf{P}_1,$$

ami nyilván egy permutációs mátrix. Ezekkel a jelölésekkel tehát

$$\mathbf{PA} = \mathbf{LU}.$$

Ezt akartuk bizonyítani. ■

3.5.4. tétel. (Golub, Van Loan [18], 115. oldal)
Legyen adott egy $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix. Tegyük fel, hogy a lebegőpontos számokkal való részleges főelemkiválasztásos Gauss-eliminációs eljárás során a $\hat{\mathbf{P}}$, $\hat{\mathbf{L}}$ és $\hat{\mathbf{U}}$ mátrixokhoz jutottunk, melyekkel $\hat{\mathbf{P}}^T\hat{\mathbf{L}}\hat{\mathbf{U}} - \mathbf{A} = \boldsymbol{\delta}\mathbf{A}$. Ekkor érvényes a

$$|\boldsymbol{\delta}\mathbf{A}| \le 3(n-1)\mathtt{u}(|\mathbf{A}| + |\hat{\mathbf{P}}^T|\cdot|\hat{\mathbf{L}}|\cdot|\hat{\mathbf{U}}|) + \mathcal{O}(\mathtt{u}^2)$$

becslés. Tehát

$$\|\boldsymbol{\delta}\mathbf{A}\|_\infty \le 3(n-1)\mathtt{u}(\|\mathbf{A}\|_\infty + n\|\hat{\mathbf{U}}\|_\infty) + \mathcal{O}(\mathtt{u}^2).$$

Vegyük észre, hogy a fenti képletben szerepel az $\mathbf{U}$ mátrix maximumnormája. Vizsgáljuk meg, hogy ha részleges főelemkiválasztást csinálunk annak érdekében, hogy $\mathbf{L}$ elemei egynél kisebb abszolútértékűek legyenek, akkor vajon nem növekszik-e meg ezzel párhuzamosan $\|\mathbf{U}\|_\infty$. Vezessük be a

$$\rho = \max_{i,j,k}\frac{|\hat{a}_{ij}^{(k)}|}{\|\mathbf{A}\|_\infty}$$

ún. növekedési faktort, ami azt méri, hogy az elimináció során az eliminált $\mathbf{A}$ mátrix egyes lépésekbeli elemei mennyire nőhetnek meg $\|\mathbf{A}\|_\infty$-hoz képest. A növekedési faktorral a következő becslést kapjuk

$$\|\boldsymbol{\delta}\mathbf{A}\|_\infty \leq 3(n-1)\mathtt{u}(\|\mathbf{A}\|_\infty + n^2\rho\|\mathbf{A}\|_\infty) + \mathcal{O}(\mathtt{u}^2) \leq 6n^3\rho\|\mathbf{A}\|_\infty\mathtt{u} + \mathcal{O}(\mathtt{u}^2).$$

A gyakorlatban a növekedési faktor általában 10-nél nem szokott nagyobb lenni, így ez egy elfogadható becslést ad a hibamátrixra. Megjegyezzük azonban, hogy mesterségesen konstruált feladatokra lehet a növekedési faktor akár 2^{n-1} is (3.11.11. feladat).

3.5.3. Cholesky-felbontás

Láttuk, hogy bizonyos feltételek mellett egy $\mathbf{A}$ négyzetes mátrix $\mathbf{A} = \mathbf{LU}$ alakba írható. A gyakorlatban sokszor olyan egyenletrendszereket kell megoldanunk, melyek mátrixa szimmetrikus, pozitív definit. Ezt a tulajdonságot kihasználva kevesebb művelettel is megadhatunk egy az LU-felbontáshoz hasonló alakú felbontást.

3.5.5. tétel.
Tegyük fel, hogy az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix egyik főminorja sem zérus. Ekkor egyértelműen léteznek olyan $\mathbf{L}$ és $\mathbf{M}$ normált alsó háromszögmátrixok és egy $\mathbf{D}$ diagonális mátrix, melyekkel $\mathbf{A} = \mathbf{LDM}^T$.

Bizonyítás. A feltételek mellett elvégezhető az LU-felbontás, és $\mathbf{U}$ egyik diagonális eleme sem lesz nulla. Legyen $\mathbf{D} = \mathrm{diag}(\mathrm{diag}(\mathbf{U}))$, azaz $\mathbf{D}$ az $\mathbf{U}$ mátrix diagonálmátrixa. Ekkor az $\mathbf{M} = (\mathbf{D}^{-1}\mathbf{U})^T$ mátrix egy normált alsó háromszögmátrix. Továbbá $\mathbf{LD}(\mathbf{D}^{-1}\mathbf{U}) = \mathbf{LU} = \mathbf{A}$. Az egyértelműség az LU-felbontás egyértelműségéből következik. ■

Szimmetrikus $\mathbf{A}$ mátrix esetén a felbontás tovább egyszerűsödik.

3.5.6. tétel.
Szimmetrikus $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix esetén egyértelműen létezik egy $\mathbf{L}$ normált alsó háromszögmátrix és egy $\mathbf{D}$ diagonális mátrix, melyekkel $\mathbf{A} = \mathbf{LDL}^T$.

Bizonyítás. Az $\mathbf{M}^{-1}\mathbf{AM}^{-\top} = \mathbf{M}^{-1}\mathbf{LD}$ mátrix szimmetrikus és alsó háromszögmátrix, azaz diagonális. Mivel $\det(\mathbf{D}) \neq 0$, ezért $\mathbf{M}^{-1}\mathbf{L}$ is diagonális, de egyúttal normált alsó háromszögmátrix is. Azaz $\mathbf{M}^{-1}\mathbf{L} = \mathbf{E}$, tehát $\mathbf{M} = \mathbf{L}$. ■

Végül, ha a mátrix szimmetrikus, pozitív definit, akkor a felbontás tovább egyszerűsödik. Így jutunk el az ún. Cholesky[4]-felbontáshoz.

3.5.7. tétel. (Cholesky-felbontás)
Tegyük fel, hogy $\mathbf{A}$ egy szimmetrikus, pozitív definit mátrix. Ekkor létezik pontosan egy olyan pozitív diagonálisú $\mathbf{G}$ alsó háromszögmátrix, mellyel $\mathbf{A} = \mathbf{GG}^T$.

[4] Andre–Louis Cholesky, 1875 – 1918, francia katonatiszt, térképész. A normálegyenlet megoldására konstruált módszerét már nem tudta publikálni, mert az I. világháborúban 1918. augusztus 31-én halálos sebesülést kapott egy észak-franciaországi harcmezőn. Módszerét egy Benoit nevű katonatiszt társa publikálta [2]. A módszer publikálása után eléggé feledésbe merült, majd Jack Todd már tanította egy King's College-i (London) analízis kurzusán a II. világháború alatt. 1948-ban Fox, Huskey és Wilkinson elemzi a módszert, és Turing jelentet meg egy cikket a módszer stabilitásával kapcsolatban. Bővebb életrajz: http://www-history.mcs.st-andrews.ac.uk/Biographies/Cholesky.html.

Bizonyítás. Az előző tétel alapján az $\mathbf{A}$ mátrix egyértelműen felírható $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T$ alakban. A $\mathbf{D}$ diagonális mátrix pozitív főátlójú az $\mathbf{A}$ mátrix pozitív definitsége miatt. Legyen $\mathbf{G} = \mathbf{L} \cdot \mathrm{diag}(\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}})$, ami egy pozitív főátlójú alsó háromszögmátrix. Ekkor $\mathbf{G}\mathbf{G}^T = \mathbf{A}$. Ezt akartuk megmutatni. ■

Természetesen a Cholesky-felbontást nem az LU-felbontásból szokás meghatározni (emlékeztetőül ennek műveletigénye $2n^3/3 + \mathcal{O}(n^2)$). A $\mathbf{G}$ mátrix elemeit oszloponként balról jobbra haladva határozhatjuk meg az $\mathbf{A}$ mátrix elemeiből pl. az alábbi példában bemutatott módon.

3.5.8. példa. Példaként állítsuk elő az

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

mátrix Cholesky-felbontását! Olyan $\mathbf{G}$ mátrixot keresünk tehát, mellyel

$$\mathbf{G}\mathbf{G}^T = \begin{bmatrix} g_{11} & 0 & 0 \\ g_{21} & g_{22} & 0 \\ g_{31} & g_{32} & g_{33} \end{bmatrix} \begin{bmatrix} g_{11} & g_{21} & g_{31} \\ 0 & g_{22} & g_{32} \\ 0 & 0 & g_{33} \end{bmatrix} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}.$$

A fenti egyenlőségből egyszerre látszik, hogy a $\mathbf{G}$ mátrix első oszlopa úgy áll elő, hogy $\mathbf{A}$ első oszlopát elosztjuk első elemének gyökével. Ezek szerint tehát $g_{11} = a_{11}/\sqrt{a_{11}} = \sqrt{2}$, $g_{21} = a_{21}/\sqrt{a_{11}} = -1/\sqrt{2}$ és $g_{31} = a_{31}/\sqrt{a_{11}} = 0$.

Ezután az $\mathbf{A}(2:3,2)$ vektorból levonjuk a $[g_{21}, g_{31}]^T$ vektor g_{21}-szeresét, majd az így létrejövő vektort osztjuk ezen vektor első elemének gyökével. Ez lesz $\mathbf{G}(2:3,2)$ oszlopa. Tehát

$$\begin{bmatrix} 2 \\ -1 \end{bmatrix} - \begin{bmatrix} -1/\sqrt{2} \\ 0 \end{bmatrix} \frac{-1}{\sqrt{2}} = \begin{bmatrix} 3/2 \\ -1 \end{bmatrix},$$

majd ezt osztjuk $\sqrt{3/2}$-del. Így $g_{22} = \sqrt{3/2}$ és $g_{32} = -\sqrt{2/3}$.

A harmadik oszlophoz (speciálisan most annak egyetlen nem nulla eleméhez) az $\mathbf{A}(3,3)$ vektorból kell levonnunk a $\mathbf{G}(3,1:2)\mathbf{G}^T(3,1:2)$ mátrixot:

$$[2] - [0, -\sqrt{2/3}] \cdot [0, -\sqrt{2/3}]^T = [2 - 2/3] = [4/3],$$

majd ezt osztani a vektor első elemének gyökével: $2/\sqrt{3}$. Így $g_{33} = 2/\sqrt{3}$. A keresett $\mathbf{G}$ mátrix tehát a

$$\mathbf{G} = \begin{bmatrix} \sqrt{2} & 0 & 0 \\ -1/\sqrt{2} & \sqrt{3/2} & 0 \\ 0 & -\sqrt{2/3} & 2/\sqrt{3} \end{bmatrix}$$

mátrix lesz. ◇

A fenti példában bemutatott vektoros előállítást elemenként kiírva a

$$\begin{aligned} g_{ii} &= \sqrt{a_{ii} - \sum_{j=1}^{i-1} g_{ij}^2}, \ i = 1, \ldots, n, \\ g_{ki} &= \frac{1}{g_{ii}} \left(a_{ki} - \sum_{j=1}^{i-1} g_{ij} g_{kj} \right), \ k = i+1, \ldots, n, \ i = 1, \ldots, n \end{aligned} \tag{3.5.1}$$

képleteket nyerjük a $\mathbf{G}$ mátrix $\mathbf{A}$ elemeivel való elemenkénti előállítására. Ezzel a módszerrel a műveletigény $n^3/3 + \mathcal{O}(n^2)$ lesz, azaz kb. feleakkora, mint az LU-felbontásé (így valóban kihasználtuk a szimmetriát).

3.5.9. megjegyzés. Ha a mátrix nem szimmetrikus, pozitív definit mátrix, akkor az algoritmus valamelyik lépésben elakad. Ezt a jelenséget felhasználhatjuk arra, hogy megállapítsuk, hogy egy szimmetrikus mátrix pozitív definit-e. ⋄

3.6. Lineáris egyenletrendszerek klasszikus iterációs megoldása

Most áttérünk a lineáris egyenletrendszerek klasszikus iterációs megoldási módszereire. Továbbra is feltesszük, hogy az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszer együtthatómátrixa négyzetes, és determinánsa nullától különbözik. Ekkor az egyértelmű megoldást jelölje $\overline{\mathbf{x}}^\star$. Az iterációs módszerek általában olyan konvergens sorozatokat konstruálnak, melyek határértéke az egyenlet megoldása. Lineáris egyenletrendszerek esetén a lineáris iterációs eljárásokkal foglalkozunk, melyek alakja

$$\overline{\mathbf{x}}^{(k+1)} = \mathbf{B}\overline{\mathbf{x}}^{(k)} + \overline{\mathbf{f}}, \ k = 0, 1, \ldots, \tag{3.6.1}$$

ahol az $\{\overline{\mathbf{x}}^{(k)}\}$ sorozattól várjuk el, hogy tartson a megoldáshoz. Rögtön láthatjuk, hogy a mátrixszal való szorzás és a vektorösszeadás műveletigénye $2n^2$ flop. Azaz kb. $n/3$ iterációs lépést végezhetünk az iterációval ahhoz, hogy a Gauss-módszer műveletigényét ne haladjuk meg. Ez azt jelentené, hogy pl. egy 100×100-as egyenletrendszer megoldása során 33 lépésből elegendően közel kellene kerülnünk a megoldáshoz. Ennek egy tetszőlegesen választott kezdővektor esetén kicsi az esélye. Az iterációs módszereket általában ezért ún. *ritka mátrixokra* (azaz olyan mátrixokra, melyekben a nemnulla elemek száma $O(n)$ nagyságrendű) alkalmazzuk.[5] Megjegyezzük, hogy differenciálegyenletek numerikus megoldása során gyakran ilyen típusú mátrixokat kapunk, így ezekben az esetekben jól alkalmazható a módszer.

Az iterációs megoldások esetén a következő kérdések vetődnek fel.

- Hogyan válasszuk meg a $\mathbf{B}$ mátrixot és az $\overline{\mathbf{f}}$, $\overline{\mathbf{x}}^{(0)}$ vektorokat?
- Mikor konvergál a megoldáshoz a sorozat?
- Mekkora lesz a konvergencia sebessége?
- Honnét tudjuk, hogy mikor álljunk le az iterációval?

3.6.1. definíció.
Az $\overline{\mathbf{x}}^{(k+1)} = \mathbf{B}\overline{\mathbf{x}}^{(k)} + \overline{\mathbf{f}}$ iterációt az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszerrel *konzisztensnek* hívjuk, ha $\overline{\mathbf{x}}^\star = \mathbf{B}\overline{\mathbf{x}}^\star + \overline{\mathbf{f}}$. (Az $\overline{\mathbf{x}}^\star$ vektor az egyenletrendszer megoldása.)

Tekintsük az $\overline{\mathbf{F}} : \mathbb{R}^n \to \mathbb{R}^n$, $\overline{\mathbf{F}}(\overline{\mathbf{x}}) = \mathbf{B}\overline{\mathbf{x}} + \overline{\mathbf{f}}$ függvényt. Erre a függvényre valamilyen $\|.\|$ vektornormában és a neki megfelelő indukált mátrixnormában igaz, hogy

$$\|\overline{\mathbf{F}}(\overline{\mathbf{x}}_1) - \overline{\mathbf{F}}(\overline{\mathbf{x}}_2)\| = \|\mathbf{B}\overline{\mathbf{x}}_1 + \overline{\mathbf{f}} - (\mathbf{B}\overline{\mathbf{x}}_2 + \overline{\mathbf{f}})\| = \|\mathbf{B}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)\| \leq \|\mathbf{B}\|\|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\|$$

tetszőleges $\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2 \in \mathbb{R}^n$ vektorokra. Így tehát, ha $\|\mathbf{B}\| < 1$, akkor az $\overline{\mathbf{F}}$ leképezés kontrakció, és teljesülnek a Banach-féle fixponttétel feltételei. Ebben az esetben tehát a (3.6.1) iterációt

[5] Az iterációs módszereket főleg olyan ritka mátrixok esetén érdemes alkalmazni, melyekben a nemnulla elemek elhelyezkedése nem jól strukturált. Sávmátrixos egyenletrendszerek a Gauss-módszer egyszerű módosításával direkt módon is gyorsan megoldhatók (lásd pl. ingamódszer).

bármilyen vektorról indítva az $\overline{\mathbf{F}}$ leképezés fixpontjához fog tartani, ami a lineáris egyenletrendszerrel konzisztens iterációk esetén az egyenletrendszer megoldása. A normák ekvivalenciája miatt természetesen nem csak az adott vektornormában, hanem tetszőleges más normában is az egyenletrendszer megoldásához fog tartani az iteráció. Ezt a megállapítást az 1.2.32. tétellel összevetve láthatjuk, hogy konzisztens iterációk esetén a $\varrho(\mathbf{B}) < 1$ feltétel szükséges és elégséges feltétele annak, hogy a (3.6.1) iteráció tetszőleges kezdővektor esetén az egyenletrendszer megoldásához tartson. Ez az állítás közvetlenül is igazolható. Legyen $\overline{\mathbf{e}}^{(k)} = \overline{\mathbf{x}}^{(k)} - \overline{\mathbf{x}}^{\star}$ az ún. hibavektor. Ekkor a konvergencia azt jelenti, hogy $\overline{\mathbf{e}}^{(k)} \to \mathbf{0}$ $(k \to \infty)$, azaz $\|\overline{\mathbf{e}}^{(k)}\| \to 0$ valamilyen normában.

3.6.2. tétel.

Egy, az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszerrel konzisztens lineáris iteráció pontosan akkor tart az egyenletrendszer megoldásához tetszőleges kezdővektor esetén, ha $\varrho(\mathbf{B}) < 1$.

Bizonyítás. Az

$$\overline{\mathbf{e}}^{(k+1)} = \overline{\mathbf{x}}^{(k+1)} - \overline{\mathbf{x}}^{\star} = \mathbf{B}\overline{\mathbf{x}}^{(k)} + \overline{\mathbf{f}} - (\mathbf{B}\overline{\mathbf{x}}^{\star} + \overline{\mathbf{f}}) = \mathbf{B}\overline{\mathbf{e}}^{(k)}$$

egyenlőség miatt $\overline{\mathbf{e}}^{(k)} = \mathbf{B}^k\overline{\mathbf{e}}^{(0)}$. A konvergenciához az kell, hogy $\mathbf{B}^k$ nullmátrixhoz tartson, aminek szükséges és elégséges feltétele, hogy $\varrho(\mathbf{B}) < 1$ legyen. ■

A bizonyításból nyilvánvaló, hogy a konvergencia annál gyorsabb, minél kisebb a $\mathbf{B}$ mátrix konvergenciasugara. Így a konvergenciára és a konvergencia sebességére vonatkozó kérdésekre már meg is találtuk a választ. Foglalkozzunk most a (3.6.1) iteráció előállításával.

Tegyük fel, hogy az $\mathbf{A}$ együtthatómátrixot előállítottuk $\mathbf{A} = \mathbf{S} - \mathbf{T}$ alakban, ahol $\mathbf{S}$ reguláris mátrix. Ekkor a $\overline{\mathbf{b}} = \mathbf{A}\overline{\mathbf{x}} = (\mathbf{S}-\mathbf{T})\overline{\mathbf{x}}$ egyenlőséget $\mathbf{S}$ inverzével balról szorozva, majd $\overline{\mathbf{x}}$-et kifejezve $\overline{\mathbf{x}} = \mathbf{S}^{-1}\mathbf{T}\overline{\mathbf{x}} + \mathbf{S}^{-1}\overline{\mathbf{b}}$. Ezen egyenlőség miatt az

$$\overline{\mathbf{x}}^{(k+1)} = \underbrace{\mathbf{S}^{-1}\mathbf{T}}_{=\mathbf{B}}\overline{\mathbf{x}}^{(k)} + \underbrace{\mathbf{S}^{-1}\overline{\mathbf{b}}}_{=\overline{\mathbf{f}}}$$

iteráció konzisztens az egyenletrendszerrel.

Az $\mathbf{S}$ mátrixot prekondicionálási mátrixnak hívjuk. Ez a mátrix határozza meg, hogy mennyire nehéz vagy könnyű az iteráció végrehajtása. Megválasztását két, egymással ellentétes követelmény határozza meg. Egyrészt könnyen invertálhatónak kell lennie, hiszen az iterációhoz szükség van a mátrix inverzére, másrészt a $\mathbf{B} = \mathbf{S}^{-1}\mathbf{T} = \mathbf{S}^{-1}(\mathbf{S} - \mathbf{A}) = \mathbf{E} - \mathbf{S}^{-1}\mathbf{A}$ egyenlőség miatt jó lenne, ha $\mathbf{S}$ "közel lenne $\mathbf{A}$-hoz", hiszen akkor a $\mathbf{B}$ mátrix spektrálsugara jóval kisebb lehetne, mint 1, ami gyors konvergenciát eredményezne.

Nézzünk meg két speciális esetet $\mathbf{S}$ megválasztására! Legyen először $\mathbf{S} = \mathbf{A}$. Ekkor az iteráció $\overline{\mathbf{x}}^{(k+1)} = (\mathbf{E} - \mathbf{S}^{-1}\mathbf{A})\overline{\mathbf{x}}^{(k)} + \mathbf{S}^{-1}\overline{\mathbf{b}} = \mathbf{0}\overline{\mathbf{x}}^{(k)} + \mathbf{A}^{-1}\overline{\mathbf{b}}$ alakú lesz, ami egy lépésben konvergál a megoldáshoz bármely kezdővektor esetén. Ebben az esetben $\mathbf{S}$ ugyan közel van $\mathbf{A}$-hoz (hiszen megegyezik vele), de inverzének meghatározása egyenértékű az egyenletrendszer direkt megoldásával, így az eljárás nem ad semmi előnyt a direkt módszerekhez képest. A másik esetben legyen $\mathbf{S}$ az egységmátrix. Ekkor az iteráció $\overline{\mathbf{x}}^{(k+1)} = (\mathbf{E} - \mathbf{A})\overline{\mathbf{x}}^{(k)} + \overline{\mathbf{b}}$ alakú lesz. Az $\mathbf{S}$ mátrixot ebben az esetben nem kell invertálni, de $\mathbf{S}$-nek semmi köze sincs az $\mathbf{A}$ mátrixhoz, így konvergenciatulajdonságai nem lesznek nagyon jók.

Most ismertetünk néhány sokszor alkalmazott lehetőséget az $\mathbf{S}$ mátrix megválasztására. Mindegyik esetben az $\mathbf{A}$ mátrixot $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$ alakba[6] írjuk fel, ahol $\mathbf{D}$ a diagonális elemek, $\mathbf{L}$ a diagonális alatti elemek -1-szereseinek, míg $\mathbf{U}$ a diagonális feletti elemek -1-szereseinek mátrixa. Feltesszük, hogy $\mathbf{D}$ főátlójának (azaz $\mathbf{A}$ főátlójának) egyik eleme sem nulla. Ez mindig elérhető az egyenletek megfelelő átrendezésével.

[6] Felhívjuk a figyelmet arra, hogy az $\mathbf{L}$ és $\mathbf{U}$ mátrixok nem egyeznek meg az LU-felbontás hasonlóan jelölt mátrixaival. A jelölés csak arra utal, hogy alsó- ill. felső háromszögmátrixokról van szó.

3.6.1. Jacobi-iteráció

Az $\mathbf{S} = \mathbf{D}$ és $\mathbf{T} = \mathbf{U} + \mathbf{L}$ választással konstruált

$$\overline{\mathbf{x}}^{(k+1)} = \underbrace{\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})}_{:=\mathbf{B}_J}\overline{\mathbf{x}}^{(k)} + \mathbf{D}^{-1}\overline{\mathbf{b}} \tag{3.6.2}$$

iterációt ($\overline{\mathbf{x}}^{(0)}$ tetszőleges kezdővektor) *Jacobi*[7]*-iterációnak* nevezzük. A Jacobi-iteráció iterációs mátrixát a továbbiakban $\mathbf{B}_J$ jelöli majd. Az iterációt vektorkomponensenként kiírva az

$$x_i^{(k+1)} = -\frac{1}{a_{ii}}\left(\sum_{j=1,\neq i}^{n} a_{ij}x_j^{(k)} - b_i\right), \quad i = 1,\ldots,n$$

iterációt nyerjük.

3.6.2. Gauss–Seidel-iteráció

Az $\mathbf{S} = \mathbf{D} - \mathbf{L}$ és $\mathbf{T} = \mathbf{U}$ választású

$$\overline{\mathbf{x}}^{(k+1)} = \underbrace{(\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}}_{\mathbf{B}_{GS}}\overline{\mathbf{x}}^{(k)} + (\mathbf{D}-\mathbf{L})^{-1}\overline{\mathbf{b}}$$

iterációt ($\overline{\mathbf{x}}^{(0)}$ tetszőleges kezdővektor) Gauss–Seidel[8]-iterációnak nevezzük. Iterációs mátrixát a továbbiakban $\mathbf{B}_{GS}$ jelöli majd. Hogy lássuk az iteráció Jacobi-iterációval való kapcsolatát, alakítsuk át az iterációs képletet. Szorozzunk először balról a $(\mathbf{D}-\mathbf{L})$ mátrixszal, majd adjunk hozzá mindkét oldalhoz $\mathbf{L}\overline{\mathbf{x}}^{(k+1)}$-et, és végül szorozzunk $\mathbf{D}$ inverzével. A fenti ekvivalens átalakítások után az

$$\overline{\mathbf{x}}^{(k+1)} = \mathbf{D}^{-1}(\mathbf{L}\overline{\mathbf{x}}^{(k+1)} + \mathbf{U}\overline{\mathbf{x}}^{(k)}) + \mathbf{D}^{-1}\overline{\mathbf{b}}$$

iterációt nyerjük. Látható, hogy a Jacobi-iterációhoz képest csak annyi a különbség, hogy a $\mathbf{D}^{-1}\mathbf{L}\overline{\mathbf{x}}^{(k)}$ tag helyett a Gauss–Seidel-iterációban $\mathbf{D}^{-1}\mathbf{L}\overline{\mathbf{x}}^{(k+1)}$ szerepel. Látszólag ez az iteráció nem explicit, hiszen a jobb oldalon is szerepel az $\overline{\mathbf{x}}^{(k+1)}$ vektor, de mivel $\mathbf{D}^{-1}\mathbf{L}$ szigorú (a főátlóban nullák szerepelnek) alsó háromszögmátrix, így $\overline{\mathbf{x}}^{(k+1)}$ mégis egyszerűen meghatározható. Az $\overline{\mathbf{x}}^{(k+1)}$ vektor első elemének meghatározásához nincs szükség az $\overline{\mathbf{x}}^{(k+1)}$ vektorra. A második elem meghatározásához pedig csak a korábban meghatározott első elemre van szükség, stb. Még jobban látszik a kapcsolat a két iteráció között, ha a Gauss–Seidel-iteráció utóbbi alakját komponensenként is kiírjuk

$$x_i^{(k+1)} = -\frac{1}{a_{ii}}\left(\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} - b_i\right) \tag{3.6.3}$$

($i = 1,\ldots,n$). Ha összevetjük ezt a Jacobi-iteráció képletével, akkor látható, hogy a Jacobi-iteráció az $\overline{\mathbf{x}}^{(k+1)}$ vektor komponenseinek meghatározásához csak az $\overline{\mathbf{x}}^{(k)}$ vektort használja, míg a Gauss–Seidel-módszer az $\overline{\mathbf{x}}^{(k+1)}$ vektor i-edik elemének meghatározásához felhasználja a vektor korábban kiszámolt $j = 1,\ldots,i-1$ indexű elemeit ($\overline{\mathbf{x}}^{(k)}$ megfelelő elemei helyett).

[7] Carl Gustav Jacob Jacobi (1804-1851, német).
[8] Philipp Ludwig von Seidel (1821-1896, német).

3.6.3. példa. Példaként végezzünk el egy-egy lépést a Jacobi- és a Gauss–Seidel-iterációkkal az $\overline{\mathbf{x}}^{(0)} = [1,1,1]^T$ vektorról indulva a

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

egyenletrendszerre!

Mindkét esetben a koordinátánkénti változatát fogjuk alkalmazni az iterációknak.

A Jacobi-iteráció esetén az i-edik egyenletből kifejezzük az i-edik változót $(i = 1,2,3)$

$$\begin{aligned} x_1 &= x_2/2 + 1/2 \\ x_2 &= x_1/2 + x_3/2 + 1 \\ x_3 &= x_2/2 + 3/2 \end{aligned}$$

és így készítjük el az iterációt

$$\begin{aligned} x_1^{(k+1)} &= x_2^{(k)}/2 + 1/2 \\ x_2^{(k+1)} &= x_1^{(k)}/2 + x_3^{(k)}/2 + 1 \\ x_3^{(k+1)} &= x_2^{(k)}/2 + 3/2. \end{aligned}$$

Így $\overline{\mathbf{x}}^{(1)}$ vektorként az $[1,2,2]^T$ vektort nyerjük.

A Gauss–Seidel-módszernél felhasználjuk az új vektor már kiszámolt komponenseit, így ez az

$$\begin{aligned} x_1^{(k+1)} &= x_2^{(k)}/2 + 1/2 \\ x_2^{(k+1)} &= x_1^{(k+1)}/2 + x_3^{(k)}/2 + 1 \\ x_3^{(k+1)} &= x_2^{(k+1)}/2 + 3/2 \end{aligned}$$

iterációt adja. Az $[1,1,1]^T$ kezdővektorral tehát az első egyenletből azt kapjuk, hogy $x_1^{(1)} = 1$, a második egyenletből $x_2^{(1)} = 2$ és a harmadikból (és ez eltér a Jacobi-módszer esetén kapottól) $x_3^{(1)} = 5/2$.

Látható, hogy

$$\mathbf{B}_J = \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{bmatrix}$$

és $\varrho(\mathbf{B}_J) = 1/\sqrt{2} \approx 0.7071$, míg

$$\mathbf{B}_{GS} = \begin{bmatrix} 0 & 1/2 & 0 \\ 0 & 1/4 & 1/2 \\ 0 & 1/8 & 1/4 \end{bmatrix}$$

és a spektrálsugara $\varrho(\mathbf{B}_{GS}) = 1/2$. Tehát mindkét módszer az egyenletrendszer megoldásához tartó vektorsorozatot eredményez, és mivel a második esetben kisebb a spektrálsugár, ezért erre a feladatra a Gauss–Seidel-módszerrel előállított sorozat konvergál gyorsabban. ⋄

Első gondolatra, és talán az előző példa is azt sugallja, úgy tűnhet, hogy a Gauss–Seidel-módszer jobb, mint a Jacobi, hiszen minden lépésben felhasználjuk az új iterációs vektor már kiszámolt komponenseit. A következő példa azt mutatja, hogy ez nincs így.

3.6.4. példa. Legyen egy lineáris algebrai egyenletrendszer együtthatómátrixa

$$\mathbf{A} = \begin{bmatrix} 1 & 1/2 & 1 \\ 1/2 & 1 & 1 \\ -2 & 2 & 1 \end{bmatrix}.$$

Ekkor

$$\mathbf{B}_J = \begin{bmatrix} 0 & -1/2 & -1 \\ -1/2 & 0 & -1 \\ 2 & -2 & 0 \end{bmatrix}, \quad \mathbf{B}_{GS} = \begin{bmatrix} 0 & -1/2 & -1 \\ 0 & 1/4 & -1/2 \\ 0 & -3/2 & -1 \end{bmatrix}.$$

Így $\varrho(\mathbf{B}_J) = 1/2 < 1$ és $\varrho(\mathbf{B}_{GS}) = |-3/8 - \sqrt{73}/8| \approx 1.443 > 1$. Ezért a fenti $\mathbf{A}$ mátrixú egyenletrendszerekre a Jacobi-módszer konvergens a Gauss–Seidel-módszer pedig nem. ⋄

3.6.3. Relaxációs módszerek

Az előző két iterációs módszer esetén az iterációs mátrix spektrálsugara adott érték, csak az $\mathbf{S}$ prekondicionálási mátrix megválasztásától függ. Ha tehát a spektrálsugár egynél nem kisebb, vagy kisebb ugyan, de közel van egyhez, akkor a módszer nem vagy nagyon lassan konvergál. Ekkor azt tehetjük, hogy változtatunk az egyenletrendszeren (pl. átrendezzük a sorait), vagy másik $\mathbf{S}$ mátrixot választunk, vagy pedig az adott $\mathbf{S}$ prekondicionálási mátrix mellett bevezetünk egy paramétert az iterációba, melyet majd úgy választunk meg, hogy az iteráció konvergens, sőt gyorsan konvergáló legyen. Az első két lehetőséggel később foglalkozunk. Most tekintsük a harmadik eljárást részletesebben a Jacobi-módszer példáján.

A $(k+1)$-edik iterációs vektor i-edik eleme triviálisan felírható

$$x_i^{(k+1)} = x_i^{(k)} + (x_i^{(k+1)} - x_i^{(k)})$$

alakban. Vezessünk be most egy pozitív ω paramétert, és definiáljuk az alábbi iterációt

$$x_i^{(k+1)} = x_i^{(k)} + \omega(x_{i,J}^{(k+1)} - x_i^{(k)}). \tag{3.6.4}$$

Itt $x_{i,J}^{(k+1)}$ azt az értéket jelöli, amit a Jacobi-módszer adna a $(k+1)$-edik iterációs vektor i-edik elemére, ha azt az $\overline{\mathbf{x}}^{(k)}$ vektor elemiből számítanánk.

A fenti képlettel definiált iterációt relaxált Jacobi-módszernek (rövidítve JOR-módszernek (Jacobi overrelaxation)) hívjuk. (A relaxálás – angolul relaxation – szó itt arra utal, hogy a Jacobi-módszer képletét nem vesszük szigorúan, hanem módosítunk egy kicsit rajta.) Ez a módosítás úgy fogható fel, hogy ω értékétől függően vagy nagyobbat, vagy kisebbet változtatunk $x_i^{(k)}$ értékén, mint amit a Jacobi-módszer változtatna rajta. (Természetesen az $\omega = 1$ eset visszaadja a Jacobi-iterációt.) Ha $\omega > 1$, akkor túlrelaxálásról, ha $0 < \omega < 1$, akkor alulrelaxálásról beszélünk. Az ω paramétert relaxációs paraméternek nevezzük.

Komponensenként kiírva tehát a Jacobi-iteráció relaxált változata az alábbi alakú lesz.

$$x_i^{(k+1)} = x_i^{(k)} + \omega \left(-\frac{1}{a_{ii}} \left[\sum_{j=1,\neq i}^{n} a_{ij} x_j^{(k)} - b_i \right] - x_i^{(k)} \right)$$
$$= (1-\omega) x_i^{(k)} - \frac{\omega}{a_{ii}} \left[\sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(k)} - b_i \right], \ i = 1, \ldots, n.$$

Írjuk fel mátrixos alakban a JOR iterációt! Ehhez helyettesítsük a (3.6.4) képletbe a Jacobi-módszer által adott $\overline{\mathbf{x}}^{(k+1)}$ vektor képletét az iteráció (3.6.2) alakját felhasználva.

$$\overline{\mathbf{x}}^{(k+1)} = \overline{\mathbf{x}}^{(k)} + \omega(\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\overline{\mathbf{x}}^{(k)} + \mathbf{D}^{-1}\overline{\mathbf{b}} - \overline{\mathbf{x}}^{(k)}).$$

Ebből kapjuk, hogy

$$\overline{\mathbf{x}}^{(k+1)} = \underbrace{((1-\omega)\mathbf{E} + \omega\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U}))}_{\mathbf{B}_{J(\omega)}} \overline{\mathbf{x}}^{(k)} + \omega\mathbf{D}^{-1}\overline{\mathbf{b}}.$$

Az iterációs mátrix tehát

$$\mathbf{B}_{J(\omega)} := (1-\omega)\mathbf{E} + \omega\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U}) = \omega\mathbf{B}_J + (1-\omega)\mathbf{E}. \tag{3.6.5}$$

A JOR iteráció az $\mathbf{A} = \mathbf{S} - \mathbf{T}$ felbontásban az

$$\mathbf{S} = (1/\omega)\mathbf{D}, \quad \mathbf{T} = \frac{1-\omega}{\omega}\mathbf{D} + \mathbf{L} + \mathbf{U} \tag{3.6.6}$$

választásnak felel meg, hiszen

$$\mathbf{S}^{-1}\mathbf{T} = \omega\mathbf{D}^{-1}\left(\frac{1-\omega}{\omega}\mathbf{D} + \mathbf{L} + \mathbf{U}\right) = (1-\omega)\mathbf{E} + \omega\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U}) = \mathbf{B}_{J(\omega)}$$

és $\mathbf{S}^{-1}\overline{\mathbf{b}} = \omega\mathbf{D}^{-1}\overline{\mathbf{b}}$.

Vegyük észre, hogy minden ω választás esetén az egyenletrendszerrel konzisztens iterációt kapunk. Mivel ω szabadon választható paraméter, így lehetőséget ad arra, hogy úgy válasszuk meg, hogy az iteráció konvergens legyen, vagy hogy a konvergencia a lehető leggyorsabb legyen.

Most térjünk át a Gauss–Seidel-iteráció relaxációjára. Ezt a módszert SOR (successive over-relaxation) fogja rövidíteni a későbbiekben, és iterációs mátrixát $\mathbf{B}_{GS(\omega)}$ jelöli majd. Induljunk ki most a Gauss–Seidel-iteráció (3.6.3) koordinátánkénti alakjából, és alkalmazzuk a relaxáció (3.6.4) képletét az $x_{i,J}^{(k+1)}$ értéket a Gauss–Seidel-módszer által adott $x_{i,GS}^{(k+1)}$ értékre cserélve. Az $x_{i,GS}^{(k+1)}$ értéket a k-adik iterációs vektor elemeiből és a (relaxációval nyert) $(k+1)$-edik iterációs vektor már kiszámolt elemeiből számítjuk a Gauss–Seidel-iteráció képletével. Így kapjuk a SOR módszer koordinátánkénti alakját

$$x_i^{(k+1)} = x_i^{(k)} + \omega \left(-\frac{1}{a_{ii}} \left(\sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} - b_i \right) - x_i^{(k)} \right)$$
$$= (1-\omega) x_i^{(k)} - \frac{\omega}{a_{ii}} \left(\sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} - b_i \right).$$

Mátrixos alakra átírva az

$$\overline{\mathbf{x}}^{(k+1)} = \underbrace{(\mathbf{D}-\omega\mathbf{L})^{-1}((1-\omega)\mathbf{D}+\omega\mathbf{U})}_{\mathbf{B}_{GS(\omega)}}\overline{\mathbf{x}}^{(k)} + \omega(\mathbf{D}-\omega\mathbf{L})^{-1}\overline{\mathbf{b}}$$

iterációt nyerjük, tehát

$$\mathbf{B}_{GS(\omega)} = (\mathbf{D}-\omega\mathbf{L})^{-1}((1-\omega)\mathbf{D}+\omega\mathbf{U}).$$

A SOR módszerhez eljuthatunk úgy is, hogy az $\mathbf{A}=\mathbf{S}-\mathbf{T}$ felbontásban az

$$\mathbf{S} = \frac{1}{\omega}\mathbf{D}-\mathbf{L}, \quad \mathbf{T} = \frac{1-\omega}{\omega}\mathbf{D}+\mathbf{U} \tag{3.6.7}$$

mátrixokat választjuk.

Tetszőleges ω esetén a SOR módszer is konzisztens az egyenletrendszerrel és $\omega = 1$ esetén visszaadja a Gauss–Seidel-módszert.

3.6.4. Iterációs módszerek konvergenciája

Korábban már láttuk, hogy egy, a lineáris egyenletrendszerrel konzisztens iterációs módszer pontosan akkor konvergens, amikor iterációs mátrixának spektrálsugara kisebb egynél. Most vizsgáljuk meg azt a kérdést, hogy hogyan lehet ezt biztosítani egy adott egyenletrendszerre, ill. hogy az előző fejezetben ismertetett iterációk esetén mikor lehetünk biztosak a konvergenciában.

Térjünk vissza a konzisztens iterációk $\overline{\mathbf{x}}^{(k+1)} = \mathbf{S}^{-1}\mathbf{T}\overline{\mathbf{x}}^{(k)} + \mathbf{S}^{-1}\overline{\mathbf{f}}$ alakjához, ahol $\mathbf{A}=\mathbf{S}-\mathbf{T}$. Ebben az esetben tehát $\mathbf{S}^{-1}\mathbf{T}$ az iterációs mátrix, tehát ennek spektrálsugaráról kell biztosítani, hogy 1-nél kisebb legyen.

3.6.5. definíció.

Az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix $\mathbf{A}=\mathbf{S}-\mathbf{T}$ felbontását *reguláris felbontásnak* hívjuk, ha $\mathbf{S}$ reguláris, $\mathbf{S}^{-1} \geq \mathbf{0}$ és $\mathbf{T} \geq \mathbf{0}$.

3.6.6. tétel.

Ha az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrixnak, amely reguláris és $\mathbf{A}^{-1} \geq \mathbf{0}$, $\mathbf{A}=\mathbf{S}-\mathbf{T}$ egy reguláris felbontása, akkor $\varrho(\mathbf{S}^{-1}\mathbf{T}) < 1$.

Bizonyítás. A $\mathbf{B} = \mathbf{S}^{-1}\mathbf{T} \geq \mathbf{0}$ mátrix a feltételek szerint jól definiált és nemnegatív mátrix. Ekkor tetszőleges pozitív egész k-ra

$$\mathbf{0} \leq \left(\sum_{i=0}^{k}\mathbf{B}^i\right)\mathbf{S}^{-1} = \left(\sum_{i=0}^{k}\mathbf{B}^i\right)\underbrace{(\mathbf{E}-\mathbf{B})\mathbf{A}^{-1}}_{\mathbf{S}^{-1}} = (\mathbf{E}-\underbrace{\mathbf{B}^{k+1}}_{\geq\mathbf{0}})\underbrace{\mathbf{A}^{-1}}_{\geq\mathbf{0}} \leq \mathbf{A}^{-1}.$$

Azaz a $\sum_{i=0}^{\infty}\mathbf{B}^i$ nemnegatív tagú sor részletösszegei korlátosak, így a sor konvergens, aminek szükséges és elégséges feltétele $\varrho(\mathbf{B}) < 1$ (1.2.34. tétel). ■

3.6.7. tétel.

Legyen $\mathbf{A}$ egy olyan négyzetes mátrix, melynek a főátlóján kívül nincs pozitív eleme. Ekkor pontosan akkor van az $\mathbf{A}$ mátrixnak olyan $\mathbf{A}=\mathbf{S}-\mathbf{T}$ reguláris felbontása, melyre $\varrho(\mathbf{S}^{-1}\mathbf{T}) < 1$, ha $\mathbf{A}$ M-mátrix.

Bizonyítás. Legyen először $\mathbf{A}$ M-mátrix. Ekkor megmutatjuk, hogy a Jacobi-iteráció felbontása reguláris. A korábbi jelölésekkel legyen $\mathbf{S} = \mathbf{D} > \mathbf{0}$ (mivel $\mathbf{A}$ M-mátrix, így főátlójában pozitív elemek állnak) és $\mathbf{T} = \mathbf{L} + \mathbf{U} \geq \mathbf{0}$. Ez egy reguláris felbontás, továbbá mivel $\mathbf{A}^{-1} \geq \mathbf{0}$, ezért az előző tétel miatt $\varrho(\mathbf{S}^{-1}\mathbf{T}) < 1$.

A másik irány igazolásához csak azt kell megmutatnunk, hogy a mátrixnak van inverze, és az egy nemnegatív mátrix. Az

$$\mathbf{A}^{-1} = (\mathbf{S}-\mathbf{T})^{-1} = (\mathbf{S}(\mathbf{E}-\mathbf{S}^{-1}\mathbf{T}))^{-1} = (\mathbf{E} - \underbrace{\mathbf{S}^{-1}\mathbf{T}}_{\varrho(\mathbf{S}^{-1}\mathbf{T})<1})^{-1}\mathbf{S}^{-1} = \sum_{k=0}^{\infty}(\mathbf{S}^{-1}\mathbf{T})^k\mathbf{S}^{-1} \geq \mathbf{0}$$

egyenlőségből és becslésből látszik, hogy $\mathbf{A}$ invertálható, hiszen előállítottuk invertálható mátrixok segítségével az inverzét. Az inverz mátrixot előállító sor tagjai nemnegatívok, így az inverz nemnegatív mátrix. ■

3.6.8. tétel.

Ha az egyenletrendszer együtthatómátrixa M-mátrix, akkor a Jacobi, a Gauss–Seidel-iterációk és ezek relaxált változatai $\omega \in (0,1)$ mellett konvergálnak az egyenletrendszer megoldásához tetszőleges kezdeti vektor esetén.

Bizonyítás. Ha $\mathbf{A}$ M-mátrix, akkor $\mathbf{A}^{-1} \geq \mathbf{0}$. A JOR iterációra a (3.6.6) képletben szereplő $\mathbf{S}$ és $\mathbf{T}$ mátrixok $\omega \in (0,1]$ esetén reguláris felbontását adják $\mathbf{A}$-nak. Így az előző tétel szerint az iteráció konvergens lesz.

A SOR módszer esetén (3.6.7) szintén reguláris felbontást ad, ha $\omega \in (0,1]$.

Az $\omega = 1$ eset felel meg a Jacobi és Gauss–Seidel-módszereknek. ■

3.6.9. tétel.

Szigorúan domináns főátlójú együtthatómátrixok esetén a Jacobi- és a Gauss–Seidel-iterációk minden kezdővektor esetén az egyenletrendszer megoldásához konvergálnak.

Bizonyítás. Mivel szigorúan domináns főátló esetén a főátlóbeli elem abszolút értéke nagyobb, mint a sorbeli többi elem abszolútérték-összege, így a $q = \|\mathbf{D}^{-1}\mathbf{L}\|_\infty$ és $s = \|\mathbf{D}^{-1}\mathbf{U}\|_\infty$ jelölésekkel $q, s \geq 0$ és $q + s < 1$. Ezek után a Jacobi-iterációra az állítás az alábbi becslésből következik.

$$\varrho(\mathbf{B}_J) \leq \|\mathbf{B}_J\|_\infty = \|\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\|_\infty = q + s < 1.$$

A Gauss–Seidel-iterációra a $\mathbf{B}_{GS}$ iterációs mátrixot felírjuk

$$\mathbf{B}_{GS} = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{U} = (\mathbf{E}-\mathbf{D}^{-1}\mathbf{L})^{-1}\mathbf{D}^{-1}\mathbf{U}$$

alakban. Így tehát

$$\varrho(\mathbf{B}_{GS}) \leq \|\mathbf{B}_{GS}\|_\infty = \|(\mathbf{E}-\mathbf{D}^{-1}\mathbf{L})^{-1}\mathbf{D}^{-1}\mathbf{U}\|_\infty \leq \frac{1}{1-q}s = \frac{s}{1-q} < \frac{1-q}{1-q} = 1,$$

ahol az első tényező becslésénél a 3.2.5. tételt alkalmaztuk. ■

3.6.10. tétel.

Ha az $\mathbf{A}$ együtthatómátrix szimmetrikus és pozitív definit, akkor a Gauss-Seidel-iteráció konvergál az egyenletrendszer megoldásához tetszőleges kezdeti vektor esetén.

Bizonyítás. Mivel $\mathbf{A}$ szimmetrikus, így $\mathbf{U} = \mathbf{L}^T$. Tehát $\mathbf{A}$ az $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{L}^T$ alakban írható. Így $\mathbf{B}_{GS} = (\mathbf{D} - \mathbf{L})^{-1}\mathbf{L}^T$. Ez a mátrix már nem feltétlenül szimmetrikus, így lehetnek képzetes sajátvektorai és sajátértékei. Legyen pl. λ egy sajátértéke, és $\overline{\mathbf{v}}$ egy hozzá tartozó sajátvektor, azaz

$$(\mathbf{D} - \mathbf{L})^{-1}\mathbf{L}^T\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}.$$

A $\mathbf{D} - \mathbf{L}$ mátrixszal balról szorozva

$$\mathbf{L}^T\overline{\mathbf{v}} = \lambda(\mathbf{D} - \mathbf{L})\overline{\mathbf{v}}, \tag{3.6.8}$$

majd a sajátvektor konjugált transzponáltjával szintén balról szorozva kapjuk, hogy

$$\overline{\mathbf{v}}^H\mathbf{L}^T\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}^H(\mathbf{D} - \mathbf{L})\overline{\mathbf{v}}.$$

Az egyenlőség minkét oldalát a $\overline{\mathbf{v}}^H(\mathbf{D} - \mathbf{L})\overline{\mathbf{v}}$ kifejezésből kivonva

$$\overline{\mathbf{v}}^H\mathbf{A}\overline{\mathbf{v}} = (1 - \lambda)\overline{\mathbf{v}}^H(\mathbf{D} - \mathbf{L})\overline{\mathbf{v}}$$

adódik. A jobb oldalt tovább alakítva kapjuk, hogy

$$(1 - \lambda)\overline{\mathbf{v}}^H(\mathbf{D} - \mathbf{L})\overline{\mathbf{v}} = (1 - \lambda)(\overline{\mathbf{v}}^H\mathbf{D}\overline{\mathbf{v}} - \overline{\mathbf{v}}^H\mathbf{L}\overline{\mathbf{v}}) = (1 - \lambda)(\overline{\mathbf{v}}^H\mathbf{D}\overline{\mathbf{v}} - \overline{\lambda}\overline{\mathbf{v}}^H(\mathbf{D} - \mathbf{L}^T)\overline{\mathbf{v}}),$$

ahol felhasználtuk (3.6.8) konjugált transzponáltját. Mivel a bal oldalon egy valós (1×1)-es mátrix áll, melynek transzponált konjugáltja önmaga, így kapjuk az

$$(1 - \overline{\lambda})\overline{\mathbf{v}}^H(\mathbf{D} - \mathbf{L}^T)\overline{\mathbf{v}} = (1 - \lambda)(\overline{\mathbf{v}}^H\mathbf{D}\overline{\mathbf{v}} - \overline{\lambda}\overline{\mathbf{v}}^H(\mathbf{D} - \mathbf{L}^T)\overline{\mathbf{v}})$$

egyenlőséget. Egyszerűsítés után nyerjük az

$$(1 - |\lambda|^2)\overline{\mathbf{v}}^H(\mathbf{D} - \mathbf{L}^T)\overline{\mathbf{v}} = (1 - \lambda)\overline{\mathbf{v}}^H\mathbf{D}\overline{\mathbf{v}}$$

egyenlőséget, majd $(1 - \overline{\lambda})$-val szorozva mindkét oldalt

$$(1 - |\lambda|^2)\overline{\mathbf{v}}^H\mathbf{A}\overline{\mathbf{v}} = |1 - \lambda|^2\overline{\mathbf{v}}^H\mathbf{D}\overline{\mathbf{v}}$$

adódik. Mivel egy szimmetrikus pozitív definit mátrix diagonálisában pozitív elemek szerepelnek, így $\overline{\mathbf{v}}^H\mathbf{D}\overline{\mathbf{v}}$ pozitív. A $\mathbf{B}_{GS}$ mátrix sajátértéke nem lehet 1, mert ellenkező esetben a (3.6.8) összefüggésből $(\mathbf{D} - \mathbf{L} - \mathbf{L}^T)\overline{\mathbf{v}} = \mathbf{A}\overline{\mathbf{v}} = \mathbf{0}$ lenne, azaz $\mathbf{A}$ szinguláris lenne. Emiatt az egyenlőség jobb oldalán pozitív szám áll, így a bal oldal is pozitív, azaz $1 - |\lambda|^2 > 0$. Tehát $|\lambda| < 1$. ■

Térjünk át a SOR iteráció vizsgálatára. Először egy szükséges feltételt igazolunk a konvergenciára.

3.6.11. tétel. (Kahan [19], 1958)
A SOR módszer esetén

$$\varrho(\mathbf{B}_{GS(\omega)}) \geq |1 - \omega|,$$

azaz a konvergencia szükséges feltétele $\omega \in (0, 2)$.

Bizonyítás. Igazak az alábbi egyenlőségek:

$$\prod_{i=1}^{n} |\lambda_i| = |\det(\mathbf{B}_{GS(\omega)})| =$$

$$= |\det((\mathbf{D}-\omega\mathbf{L})^{-1})| \cdot |\det((1-\omega)\mathbf{D}+\omega\mathbf{U})| = |1-\omega|^n.$$

Tehát

$$\varrho(\mathbf{B}_{GS(\omega)}) = \max_{i=1,\dots,n} |\lambda_i| \underbrace{\geq}_{\text{Számt. mért. köz.}}$$

$$\geq \left(\prod_{i=1}^{n} |\lambda_i|\right)^{1/n} = |1-\omega|. \blacksquare$$

A következő tétel, melyet bizonyítás nélkül közlünk, azt mondja, hogy szimmetrikus pozitív definit mátrixokra a Kahan-tétel feltétele elégséges is a konvergenciához.

3.6.12. tétel. (Ostrowski [26], 1954; Reich [27], 1949)
Ha $\mathbf{A}$ szimmetrikus, pozitív definit mátrix, és $\omega \in (0,2)$, akkor

$$\varrho(\mathbf{B}_{GS(\omega)}) < 1,$$

azaz a SOR iteráció konvergens lesz.

Már említettük, hogy a differenciálegyenletek numerikus megoldása során a leggyakrabban előforduló mátrixtípusok a szimmetrikus pozitív definit mátrixok, a diagonálisan domináns mátrixok és az M-mátrixok. Az előző tételek alapján elmondhatjuk, hogy szigorúan domináns főátlójú mátrixok esetén a Jacobi- és a Gauss–Seidel-iteráció is konvergens lesz. Néha az egyenletek sorrendjének megváltoztatásával könnyen el is érhető, hogy a mátrix szigorúan domináns főátlóval rendelkezzen. Szimmetrikus pozitív definit mátrixok esetén a SOR módszer konvergens lesz minden $\omega \in (0,2)$ választás esetén (és csak ezekre). M-mátrixok esetén a SOR és a JOR módszerek is konvergálnak $\omega \in (0,1]$ esetén, de a Stein–Rosenberg-tétel miatt a SOR módszer a célravezetőbb (gyorsabban konvergál).

3.6.5. Leállási feltételek

Már csak annak tisztázása maradt hátra, hogy mikor hagyjunk abba egy iterációs eljárást. Honnét tudjuk, hogy milyen messze vagyunk az ismeretlen megoldástól az iteráció során? Most felsorolunk néhány lehetséges szabályt, ún. leállási feltételt az iterációhoz.

- Ha $\|\mathbf{B}\| < 1$ valamilyen normában, akkor a Banach-féle fixponttétel miatt

$$\|\overline{\mathbf{x}} - \overline{\mathbf{x}}^{(j)}\| \leq \frac{\|\mathbf{B}\|^j}{1-\|\mathbf{B}\|}\|\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(0)}\|.$$

  A $\|\mathbf{B}\|$ értékből és az első iteráció eredményéből megmondhatjuk, hogy hány iterációra van szükség az adott normabeli pontosság eléréséhez.

- Vizsgáljuk az egymás utáni iterációk eredményeit, ha $\|\overline{\mathbf{x}}^{(k+1)} - \overline{\mathbf{x}}^{(k)}\|$ elegendően kicsi, akkor leállítjuk az iterációt.

- Kiszámítjuk az ún. maradékvektorokat: $\overline{\mathbf{r}}^{(k)} = \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}^{(k)}$. Ha $\|\overline{\mathbf{r}}^{(k+1)} - \overline{\mathbf{r}}^{(k)}\|/\|\overline{\mathbf{r}}^{(0)}\|$ elegendően kicsi, akkor leállítjuk az iterációt.

- Megadunk egy $k_{\max}$ értéket, ahol mindenképpen abbahagyjuk az iterációt.

Ezek közül akár többet is alkalmazhatunk. Pl. az utolsó feltételt minden programba érdemes beépíteni, hogy gondoskodjunk a tényleges leállásról.

3.7. Variációs módszerek

Ebben a fejezetben a lineáris egyenletrendszerek megoldásának egy újabb iterációs megoldási lehetőségét mutatjuk be szimmetrikus, pozitív definit mátrixokra. Az alapötlet az, hogy megadunk egy többváltozós függvényt, melynek abszolút minimumhelye az egyenletrendszer megoldása. Ezt a minimumhelyet keressük meg egy megfelelő iterációs eljárással.

Legyen tehát $\mathbf{A} \in \mathbb{R}^{n\times n}$ szimmetrikus, pozitív definit mátrix, és tekintsük a

$$\phi(\overline{\mathbf{x}}) = \frac{1}{2}\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{x}}^T\overline{\mathbf{b}}$$

n-változós függvényt. Az összeszorzott mátrixok mérete szerint a jobb oldali kifejezés valóban egy valós számot (egy (1×1)-es mátrixot) rendel minden vektorhoz. Először megmutatjuk, hogy ennek a függvénynek egyetlen abszolút minimumhelye van az $\mathbb{R}^n$ halmazon és ez pontosan az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszer megoldása.

3.7.1. tétel.
A $\phi(\overline{\mathbf{x}})$ függvénynek abszolút minimuma van az $\overline{\mathbf{x}}^\star = \mathbf{A}^{-1}\overline{\mathbf{b}}$ pontban. A minimum értéke $-\overline{\mathbf{b}}^T\mathbf{A}^{-1}\overline{\mathbf{b}}/2$.

Bizonyítás. Legyen $\overline{\mathbf{x}} = \overline{\mathbf{x}}^\star + \Delta\overline{\mathbf{x}}$ egy tetszőleges vektor. Ekkor a $\Delta\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}}^\star = \Delta\overline{\mathbf{x}}^T\overline{\mathbf{b}}$ egyenlőség miatt

$$\phi(\overline{\mathbf{x}}^\star + \Delta\overline{\mathbf{x}}) = \phi(\overline{\mathbf{x}}^\star) + \frac{1}{2}\Delta\overline{\mathbf{x}}^T\mathbf{A}\Delta\overline{\mathbf{x}}. \tag{3.7.1}$$

Ezután $\mathbf{A}$ pozitív definitségéből következik az állítás. A minimum értéke egyszerű behelyettesítéssel adódik. ■.

A $\phi(\overline{\mathbf{x}})$ függvény mindegyik változója szerint parciálisan deriválható, így minden pontban értelmezve van a függvény gradiense. Most kiszámítjuk a gradiensfüggvényt. Határozzuk meg a $\phi(\overline{\mathbf{x}})$ függvény x_k változó szerinti parciális deriváltját $(k = 1,\dots,n)$. Írjuk ki ehhez a $\phi(\overline{\mathbf{x}})$ függvényt részletesen.

$$\phi(\overline{\mathbf{x}}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j - \sum_{j}^{n} b_jx_j.$$

Hagyjuk el az x_k változót nem tartalmazó tagokat, hiszen a deriválás során ezek úgyis eltűnnek, így egyszerűsödik a deriválandó kifejezés.

$$\frac{\partial\phi}{\partial x_k}(\overline{\mathbf{x}}) = \left(\frac{1}{2}\left(\sum_{k\neq i=1}^{n} a_{ik}x_ix_k + \sum_{k\neq j=1}^{n} a_{kj}x_kx_j + a_{kk}x_k^2\right) - b_kx_k\right)'_{x_k} = \sum_{j=1}^{n} a_{kj}x_j - b_k.$$

Azaz a gradiensfüggvény $\operatorname{grad}\phi(\overline{\mathbf{x}}) = \mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}$. Egy lineáris egyenletrendszer esetén az $\overline{\mathbf{r}} = \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}$ vektort maradékvektornak hívjuk. Jelölésére az angol residual szó kezdőbetűjét használjuk. A maradékvektor értéke az $\overline{\mathbf{x}}$ vektor függvénye, de ezt nem szoktuk a jelölésben feltüntetni. A szövegkörnyezetből mindig világos lesz, hogy a maradékvektort melyik $\overline{\mathbf{x}}$ vektorral képeztük. A maradékvektor megmutatja, hogy egy adott $\overline{\mathbf{x}}$ vektor esetén mekkora az eltérés az egyenletrendszer két oldala között. Nyilvánvalóan $\overline{\mathbf{x}} = \overline{\mathbf{x}}^\star$ esetén $\overline{\mathbf{r}} = \mathbf{0}$. Fontos észrevétel, hogy a gradiensvektor a maradékvektor (-1)-szerese.

Az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszer megoldása tehát a $\phi(\overline{\mathbf{x}})$ függvény minimumhelyének megkeresésével egyenértékű. Hogy jobban el lehessen képzelni, hogy milyen függvényt kell minimalizálni, írjuk a (3.7.1) képletben a $\Delta\overline{\mathbf{x}}$ vektor helyére az $\overline{\mathbf{x}} - \overline{\mathbf{x}}^\star$ vektort

$$\phi(\overline{\mathbf{x}}) = -\frac{1}{2}\overline{\mathbf{b}}^T\mathbf{A}^{-1}\overline{\mathbf{b}} + \frac{1}{2}(\overline{\mathbf{x}} - \overline{\mathbf{x}}^\star)^T\mathbf{A}(\overline{\mathbf{x}} - \overline{\mathbf{x}}^\star).$$

Innét látszik, hogy a függvény $c \geq -\overline{\mathbf{b}}^T\mathbf{A}^{-1}\overline{\mathbf{b}}/2$ értékhez tartozó szintvonala egy $\overline{\mathbf{x}}^\star$ centrumú hiperellipszoid. ($\mathbf{A} = \mathbf{E}$ esetén koncentrikus kör.)

3.7.2. példa. Tekintük a $2x_1 = 4$, $8x_2 = 8$ egyenletrendszert, melynek megoldása $x_1^\star = 2, x_2^\star = 1$. Ekkor ez a megoldás lesz a

$$\phi(\overline{\mathbf{x}}) = x_1^2 + 4x_2^2 - 4x_1 - 8x_2 = (x_1 - 2)^2 + 4(x_2 - 1)^2 - 8$$

kétváltozós függvény minimuma, ahogy ez az átalakításból látható is. A $c = 0$ értékhez tartozó szintvonal egyenlete

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

ami egy $\overline{\mathbf{x}}^\star$ közepű, $\sqrt{8}$ ill. $\sqrt{2}$ féltengelyű ellipszis egyenlete. ◇

Az előbbiek alapján a $\phi(\overline{\mathbf{x}})$ függvény minimumhelyének megkeresése tulajdonképpen egy olyan felület "legmélyebben" fekvő pontjának megkeresése, melynek szintvonalai koncentrikus hiperellipszoidok (két dimenzióban koncentrikus ellipszisek, lásd 3.7.1. ábra).

Először foglalkozzunk az iránymenti minimumok megkeresésével, azaz azzal az esettel, amikor nem az egész felületen, hanem csak egy adott pontból adott irányban szeretnénk megkeresni a legkisebb függvényértékét $\phi(\overline{\mathbf{x}})$-nek.

3.7.3. tétel.
Legyenek $\overline{\mathbf{x}}$ és $\overline{\mathbf{p}} \neq \mathbf{0}$ adott vektorok. A $g(\alpha) = \phi(\overline{\mathbf{x}} + \alpha\overline{\mathbf{p}})$ egyváltozós függvény egyértelmű minimumát az $\alpha = \overline{\mathbf{p}}^T\overline{\mathbf{r}}/(\overline{\mathbf{p}}^T\mathbf{A}\overline{\mathbf{p}})$ választás esetén veszi fel.

Bizonyítás. Mivel

$$g(\alpha) = \phi(\overline{\mathbf{x}} + \alpha\overline{\mathbf{p}}) = \phi(\overline{\mathbf{x}}) - \alpha\overline{\mathbf{p}}^T\overline{\mathbf{r}} + \frac{1}{2}\alpha^2\overline{\mathbf{p}}^T\mathbf{A}\overline{\mathbf{p}}$$
$$= \phi(\overline{\mathbf{x}}) + \alpha\left(\frac{1}{2}\alpha\overline{\mathbf{p}}^T\mathbf{A}\overline{\mathbf{p}} - \overline{\mathbf{p}}^T\overline{\mathbf{r}}\right)$$

α-ban másodfokú polinom, melynek főegyütthatója pozitív ($\mathbf{A}$ pozitív definitsége miatt), ezért a minimumát valóban az $\alpha = \overline{\mathbf{p}}^T\overline{\mathbf{r}}/(\overline{\mathbf{p}}^T\mathbf{A}\overline{\mathbf{p}})$ értéknél veszi fel. ■

Ezek után meg is adhatjuk a legegyszerűbb algoritmust a $\phi(\overline{\mathbf{x}})$ függvény minimumhelyének megkeresésére. Tegyük fel, hogy valamilyen módszerrel (ezt a későbbiekben pontosítani fogjuk) meghatároztuk az egyes lépéseknél alkalmazandó keresési irányokat: $\overline{\mathbf{p}}_1, \overline{\mathbf{p}}_2, \ldots$. Ekkor az alábbi módon járhatunk el. Választunk egy kiindulási közelítést, legyen ez $\overline{\mathbf{x}}_0 = \mathbf{0}$ (de választhatunk tetszőleges más vektort is). Az $\overline{\mathbf{x}}_0$ pontból a $\overline{\mathbf{p}}_1$ keresési irányban az előző tétellel meghatározzuk a $\phi(\overline{\mathbf{x}})$ legkisebb függvényértékét és a minimumhelyét. Ezt a minimumhelyet választjuk a következő $\overline{\mathbf{x}}_1$ közelítésnek. Ezután hasonlóan járunk el mindig a következő keresési irányt választva az adott pontból. Az algoritmus az alábbi módon foglalható össze.

Alap algoritmus: minimalizálás adott $\overline{\mathbf{p}}_1, \overline{\mathbf{p}}_2, \ldots \neq \mathbf{0}$ keresési irányokkal,
$\mathbf{A} \in \mathbb{R}^{n\times n}$ pozitív definit, $\overline{\mathbf{b}} \in \mathbb{R}^n$ adott.
$k := 0$, $\overline{\mathbf{r}}_0 := \overline{\mathbf{b}}$, $\overline{\mathbf{x}}_0 := \mathbf{0}$ (lehet más is)

while $\overline{\mathbf{r}}_k \neq \mathbf{0}$ **do**
  $k := k + 1$
  $\alpha_k := \overline{\mathbf{p}}_k^T \overline{\mathbf{r}}_{k-1} / (\overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{p}}_k)$
  $\overline{\mathbf{x}}_k := \overline{\mathbf{x}}_{k-1} + \alpha_k \overline{\mathbf{p}}_k$
  $\overline{\mathbf{r}}_k := \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_k$
end while

Az alapalgoritmus megismerése után már csak azt a kérdést kell megvizsgálnunk, hogy milyen módon válasszuk meg a keresési irányokat ahhoz, hogy a minimumhelyet a leghatékonyabban tudjuk meghatározni.

3.7.1. Gradiens-módszer

Ismert, hogy egy többváltozós függvény a gradiensvektorával ellentétes irányban csökken a leggyorsabban. Ehhez elég felidéznünk az iránymenti deriváltról tanultakat. Így kézenfekvőnek látszik mindig a gradiensvektor (-1)-szeresét (ami a korábban mondottak szerint éppen az adott pontbeli $\overline{\mathbf{r}} = \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}$ maradékvektor) választani keresési iránynak. Az így nyert iterációs módszert a $\phi(\overline{\mathbf{x}})$ függvény minimumhelyének megkeresésére gradiens-módszernek, másképpen a legmeredekebb ereszkedés (angolul steepest descent) módszerének hívjuk.

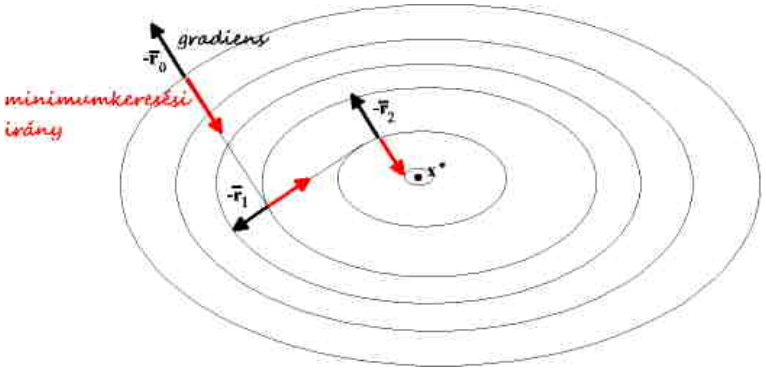
A gradiens-módszernél tehát a 3.7.3. tételt alkalmazva egy $\overline{\mathbf{x}}$ pontból, ahol $\overline{\mathbf{r}} = \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}$ a maradékvektor, az $\overline{\mathbf{r}}$ vektor irányába az $\overline{\mathbf{x}} + (\overline{\mathbf{r}}^T\overline{\mathbf{r}}/(\overline{\mathbf{r}}^T\mathbf{A}\overline{\mathbf{r}}))\overline{\mathbf{r}}$ pontba lépünk tovább. Innét pedig az itteni

$$\overline{\mathbf{b}} - \mathbf{A}\left(\overline{\mathbf{x}} + \frac{\overline{\mathbf{r}}^T\overline{\mathbf{r}}}{\overline{\mathbf{r}}^T\mathbf{A}\overline{\mathbf{r}}}\overline{\mathbf{r}}\right)$$

maradékvektor irányában keressük a következő iránymenti minimumot. Az

$$\overline{\mathbf{r}}^T\left(\overline{\mathbf{b}} - \mathbf{A}\left(\overline{\mathbf{x}} + \frac{\overline{\mathbf{r}}^T\overline{\mathbf{r}}}{\overline{\mathbf{r}}^T\mathbf{A}\overline{\mathbf{r}}}\overline{\mathbf{r}}\right)\right) = \overline{\mathbf{r}}^T\overline{\mathbf{r}} - \overline{\mathbf{r}}^T\mathbf{A}\frac{\overline{\mathbf{r}}^T\overline{\mathbf{r}}}{\overline{\mathbf{r}}^T\mathbf{A}\overline{\mathbf{r}}}\overline{\mathbf{r}} = \overline{\mathbf{r}}^T\overline{\mathbf{r}} - \frac{\overline{\mathbf{r}}^T\overline{\mathbf{r}}}{\overline{\mathbf{r}}^T\mathbf{A}\overline{\mathbf{r}}}\overline{\mathbf{r}}^T\mathbf{A}\overline{\mathbf{r}} = 0$$

egyenlőség miatt látható, hogy az iteráció során az egymás utáni keresési irányok merőlegesek egymásra. A gradiens-módszer során végzett lépéseket szemlélteti egy kétváltozós egyenletrendszer esetén a 3.7.1. ábra.



3.7.1. ábra:

A gradiens-módszer algoritmusa tehát a következő.

Gradiens-módszer, $\mathbf{A} \in \mathbb{R}^{n\times n}$ pozitív definit mátrix, $\overline{\mathbf{b}} \in \mathbb{R}^n$ adott vektor.

$k := 0$, $\overline{\mathbf{r}}_0 := \overline{\mathbf{b}}$, $\overline{\mathbf{x}}_0 := \mathbf{0}$ (lehet más is)
while $\overline{\mathbf{r}}_k \neq \mathbf{0}$ **do**
  $k := k + 1$
  $\alpha_k := \overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1} / (\overline{\mathbf{r}}_{k-1}^T \mathbf{A} \overline{\mathbf{r}}_{k-1})$
  $\overline{\mathbf{x}}_k := \overline{\mathbf{x}}_{k-1} + \alpha_k \overline{\mathbf{r}}_{k-1}$
  $\overline{\mathbf{r}}_k := \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_k$
end while

Vizsgáljuk meg a gradiens-módszer konvergenciáját. A következő tétel arra ad becslést, hogy a ϕ függvény értéke hogyan változik egy-egy lépés során.

3.7.4. tétel.
A gradiens-módszer során érvényes a

$$\frac{\phi(\overline{\mathbf{x}}_{k+1}) + (1/2)\overline{\mathbf{b}}^T \mathbf{A}^{-1} \overline{\mathbf{b}}}{\phi(\overline{\mathbf{x}}_k) + (1/2)\overline{\mathbf{b}}^T \mathbf{A}^{-1} \overline{\mathbf{b}}} \leq 1 - \frac{1}{\kappa_2(\mathbf{A})}$$

becslés ($k = 0, 1, \ldots$)

Bizonyítás. Egyszerű átalakításokkal kapjuk, hogy a bal oldali tört nevezője

$$\begin{aligned}
\phi(\overline{\mathbf{x}}_k) + (1/2)\overline{\mathbf{b}}^T \mathbf{A}^{-1} \overline{\mathbf{b}} &= \frac{1}{2}\overline{\mathbf{x}}_k^T \mathbf{A} \overline{\mathbf{x}}_k - \overline{\mathbf{x}}_k^T \overline{\mathbf{b}} + (1/2)\overline{\mathbf{b}}^T \mathbf{A}^{-1} \overline{\mathbf{b}} \\
&= \frac{1}{2}\overline{\mathbf{x}}_k^T (\mathbf{A}\overline{\mathbf{x}}_k - \overline{\mathbf{b}}) + \frac{1}{2} \underbrace{(\overline{\mathbf{b}}^T \mathbf{A}^{-1} - \overline{\mathbf{x}}_k^T)}_{(\overline{\mathbf{b}}^T - \overline{\mathbf{x}}_k^T \mathbf{A})\mathbf{A}^{-1}} \overline{\mathbf{b}} \\
&= -\frac{1}{2}\overline{\mathbf{x}}_k^T \overline{\mathbf{r}}_k + \frac{1}{2}\overline{\mathbf{r}}_k^T \mathbf{A}^{-1} \overline{\mathbf{b}} = \frac{1}{2}\overline{\mathbf{r}}_k^T \mathbf{A}^{-1}(\overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_k) = \frac{1}{2}\overline{\mathbf{r}}_k^T \mathbf{A}^{-1} \overline{\mathbf{r}}_k.
\end{aligned}$$

Továbbá a $(k+1)$-edik maradékvektorra

$$\begin{aligned}
\overline{\mathbf{r}}_{k+1} - \overline{\mathbf{r}}_k &= \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_{k+1} - (\overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_k) \\
&= -\mathbf{A}(\overline{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_k) = -\mathbf{A}\alpha_{k+1}\overline{\mathbf{r}}_k.
\end{aligned}$$

Így

$$\overline{\mathbf{r}}_{k+1} = (\mathbf{E} - \mathbf{A}\alpha_{k+1})\overline{\mathbf{r}}_k.$$

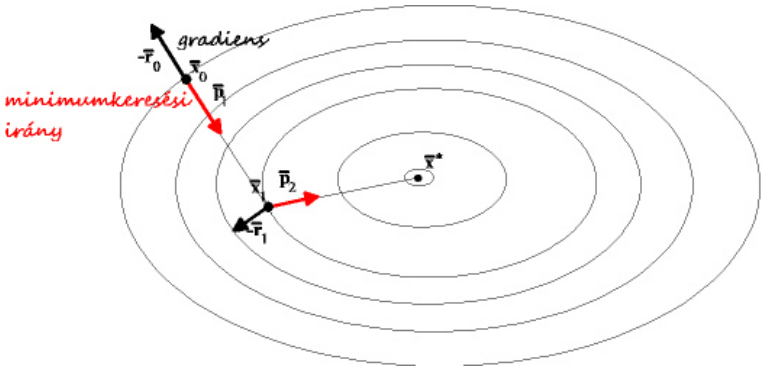
Tehát

$$\begin{aligned}
\frac{\phi(\overline{\mathbf{x}}_{k+1}) + (1/2)\overline{\mathbf{b}}^T \mathbf{A}^{-1} \overline{\mathbf{b}}}{\phi(\overline{\mathbf{x}}_k) + (1/2)\overline{\mathbf{b}}^T \mathbf{A}^{-1} \overline{\mathbf{b}}} &= \frac{\overline{\mathbf{r}}_{k+1}^T \mathbf{A}^{-1} \overline{\mathbf{r}}_{k+1}}{\overline{\mathbf{r}}_k^T \mathbf{A}^{-1} \overline{\mathbf{r}}_k} = \frac{\overline{\mathbf{r}}_k^T (\mathbf{E} - \alpha_{k+1}\mathbf{A})\mathbf{A}^{-1}(\mathbf{E} - \alpha_{k+1}\mathbf{A})\overline{\mathbf{r}}_k}{\overline{\mathbf{r}}_k^T \mathbf{A}^{-1} \overline{\mathbf{r}}_k} \\
&= 1 - \frac{2\alpha_{k+1}\overline{\mathbf{r}}_k^T \overline{\mathbf{r}}_k - \alpha_{k+1}^2 \overline{\mathbf{r}}_k^T \mathbf{A} \overline{\mathbf{r}}_k}{\overline{\mathbf{r}}_k^T \mathbf{A}^{-1} \overline{\mathbf{r}}_k} \\
&= 1 - \frac{(\overline{\mathbf{r}}_k^T \overline{\mathbf{r}}_k)^2}{\overline{\mathbf{r}}_k^T \mathbf{A} \overline{\mathbf{r}}_k \overline{\mathbf{r}}_k^T \mathbf{A}^{-1} \overline{\mathbf{r}}_k} \\
&\leq 1 - \frac{1}{\|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2} = 1 - \frac{1}{\kappa_2(\mathbf{A})}. \blacksquare
\end{aligned}$$

A ϕ függvény változását mutató előző tételből jól látható, hogy a függvény nagyon lassan csökken, azaz a módszer nagyon lassan konvergál abban az esetben, ha $\kappa_2(\mathbf{A})$ nagy érték.

3.7.2. Konjugált gradiens-módszer

Természetesen vetődik fel az a kérdés, hogy a keresési irányok másfajta megválasztásával nem lehetne-e gyorsítani valahogy a konvergenciát. Vizsgáljuk meg a kérdést egy kétváltozós lineáris egyenletrendszeren! Tekintsük a 3.7.2. ábrát. Itt az $\overline{\mathbf{x}}_0$ pontból indulva átlépünk az $\overline{\mathbf{x}}_1$ pontba. Ha itt olyan keresési irányt tudnánk választani, amely a minimumhely irányába mutat, akkor a következő lépésben már meg is találnánk a minimumot, hiszen abban a keresési irányban az iránymenti minimum egyben abszolút minimum is. De hogyan lehetséges így választani a keresési irányt, hiszen ehhez látszólag szükségünk lenne a keresett $\overline{\mathbf{x}}^\star$ minimumhelyre?



3.7.2. ábra:

Induljunk ki abból a korábban igazolt állításból, hogy az $\overline{\mathbf{x}}_0$ pontbeli keresési irány (a gradiensvektor (-1)-szerese) merőleges az $\overline{\mathbf{x}}_1$ pontbeli maradékvektorra.

$$0 = \overline{\mathbf{p}}_1^T \overline{\mathbf{r}}_1 = \overline{\mathbf{p}}_1^T (\overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_1) = \overline{\mathbf{p}}_1^T (\mathbf{A}\overline{\mathbf{x}}^\star - \mathbf{A}\overline{\mathbf{x}}_1) = \overline{\mathbf{p}}_1^T \mathbf{A}(\overline{\mathbf{x}}^\star - \overline{\mathbf{x}}_1).$$

Ebből a képletből látható, hogy az $\overline{\mathbf{x}}_1$ pontból az $\overline{\mathbf{x}}^\star$ megoldásba vezető vektor teljesíti a $\overline{\mathbf{p}}_1^T \mathbf{A}(\overline{\mathbf{x}}^\star - \overline{\mathbf{x}}_1) = 0$ feltételt. Az egyszerűbb megfogalmazás kedvéért vezessük be az alábbi fogalmat.

3.7.5. definíció.
Legyen adva egy $\mathbf{A} \in \mathbb{R}^{n\times n}$ szimmetrikus, pozitív definit mátrix. Azt mondjuk, hogy az $\overline{\mathbf{x}}$ és $\overline{\mathbf{y}}$ *vektorok* $\boldsymbol{A}$*-konjugáltak (vagy* $\boldsymbol{A}$*-ortogonálisak), ha* $\overline{\boldsymbol{x}}^T \boldsymbol{A}\overline{\boldsymbol{y}} = 0$.

Vegyük észre, hogy ha $\mathbf{A}$ az egységmátrix, akkor az $\mathbf{A}$-ortogonalitás pontosan a hagyományos skaláris szorzatbeli ortogonalitást jelenti.

A definíció segítségével tehát mondhatjuk, hogy a $\overline{\mathbf{p}}_2$ keresési irányt úgy kell megválasztanunk, hogy az legyen A-ortogonális a $\overline{\mathbf{p}}_1$ keresési irányra. Rögtön látjuk, hogy ez lehetséges az $\overline{\mathbf{x}}^\star$ megoldás ismerete nélkül is. Keressük

$$\overline{\mathbf{p}}_2 = \overline{\mathbf{r}}_1 - \beta_1 \overline{\mathbf{p}}_1$$

alakban a második keresési irányt. Az A-ortogonalitási feltételt felhasználva ekkor

$$\beta_1 = \frac{\overline{\mathbf{p}}_1^T \mathbf{A}\overline{\mathbf{r}}_1}{\overline{\mathbf{p}}_1^T \mathbf{A}\overline{\mathbf{p}}_1}.$$

Továbbá, hasonlóan a gradiens-módszer konvergenciasebességéről szóló 3.7.4. tétel bizonyításában szereplő átalakításokhoz

$$\overline{\mathbf{r}}_{k+1} - \overline{\mathbf{r}}_k = \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_{k+1} - (\overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_k)$$

$$= -\mathbf{A}(\overline{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_k) = -\mathbf{A}\alpha_{k+1}\overline{\mathbf{p}}_{k+1}.$$

Így

$$\overline{\mathbf{r}}_{k+1} = \overline{\mathbf{r}}_k - \mathbf{A}\alpha_{k+1}\overline{\mathbf{p}}_{k+1}.$$

Kétváltozós lineáris egyenletrendszerre tehát az alábbi algoritmus alkalmazható. Az algoritmus két lépésben megtalálja az abszolút minimumhelyet. Azaz egy tetszőleges kezdőpontból a maradékvektor irányában keres először iránymenti minimumot, majd ebben a pontban meghatároz egy vektort, mely A-konjugált az első keresési irányra. Ebben az irányban megkeresve az iránymenti minimumot, az abszolút minimumot kapjuk.

Konjugált gradiens-módszer kétismeretlenes lineáris egyenletrendszerre.
$\mathbf{A} \in \mathbb{R}^{2\times 2}$ pozitív definit mátrix, $\overline{\mathbf{b}} \in \mathbb{R}^2$ tetszőleges vektor.

$\overline{\mathbf{x}}_0 := \mathbf{0}$, $\overline{\mathbf{r}}_0 := \overline{\mathbf{b}}$
$\overline{\mathbf{p}}_1 := \overline{\mathbf{r}}_0$
$\alpha_1 := \overline{\mathbf{p}}_1^T\overline{\mathbf{r}}_0/(\overline{\mathbf{p}}_1^T\mathbf{A}\overline{\mathbf{p}}_1)$
$\overline{\mathbf{x}}_1 := \overline{\mathbf{x}}_0 + \alpha_1\overline{\mathbf{p}}_1$
$\overline{\mathbf{r}}_1 := \overline{\mathbf{r}}_0 - \alpha_1\mathbf{A}\overline{\mathbf{p}}_1$
$\beta_1 := \overline{\mathbf{p}}_1^T\mathbf{A}\overline{\mathbf{r}}_1/(\overline{\mathbf{p}}_1^T\mathbf{A}\overline{\mathbf{p}}_1)$
$\overline{\mathbf{p}}_2 = \overline{\mathbf{r}}_1 - \beta_1\overline{\mathbf{p}}_1$
$\alpha_2 := \overline{\mathbf{p}}_2^T\overline{\mathbf{r}}_1/(\overline{\mathbf{p}}_2^T\mathbf{A}\overline{\mathbf{p}}_2)$
$\overline{\mathbf{x}}_2 := \overline{\mathbf{x}}_1 + \alpha_2\overline{\mathbf{p}}_2$ ($= \overline{\mathbf{x}}^\star$ pontos megoldás)

Ezt a módszert a konjugált keresési irányok választása miatt konjugált gradiens-módszernek nevezzük. A módszert teljes egészében Magnus Hestenes és Eduard Stiefel adta meg először az 1950-es évek elején. Mivel két iterációs lépésben pontosan számolva pontosan megkapjuk a megoldást, ezért ez a módszer elméletileg direkt módszernek tekinthető.

Most rátérünk a konjugált gradiens-módszer többdimenziós algoritmusára, amely könnyen adódik a kétváltozóban megismert algoritmus általánosításaként. (Később majd még pontosítjuk ezt az algoritmust.)

Konjugált gradiens-módszer, első változat.
$\mathbf{A} \in \mathbb{R}^{n\times n}$ pozitív definit mátrix, $\overline{\mathbf{b}} \in \mathbb{R}^n$ adott.

$k := 0$, $\overline{\mathbf{r}}_0 := \overline{\mathbf{b}}$, $\overline{\mathbf{x}}_0 := \mathbf{0}$, $\overline{\mathbf{p}}_1 = \overline{\mathbf{r}}_0$
while $\overline{\mathbf{r}}_k \neq \mathbf{0}$ **do**
  $k := k + 1$
  $\alpha_k := \overline{\mathbf{p}}_k^T\overline{\mathbf{r}}_{k-1}/(\overline{\mathbf{p}}_k^T\mathbf{A}\overline{\mathbf{p}}_k)$
  $\overline{\mathbf{x}}_k := \overline{\mathbf{x}}_{k-1} + \alpha_k\overline{\mathbf{p}}_k$
  $\overline{\mathbf{r}}_k := \overline{\mathbf{r}}_{k-1} - \alpha_k\mathbf{A}\overline{\mathbf{p}}_k$
  $\beta_k := \overline{\mathbf{p}}_k^T\mathbf{A}\overline{\mathbf{r}}_k/(\overline{\mathbf{p}}_k^T\mathbf{A}\overline{\mathbf{p}}_k)$
  $\overline{\mathbf{p}}_{k+1} := \overline{\mathbf{r}}_k - \beta_k\overline{\mathbf{p}}_k$
end while

Természetesen az korántsem világos, hogy ez az algoritmus is hasonló jó tulajdonságokkal rendelkezik, mint a kétdimenziós esetben bemutatott algoritmus. Azt, hogy ez mégis így van, a következő tétel mutatja.

3.7.6. tétel.
Ha $\overline{\mathbf{r}}_{k-1} \neq 0$ egy adott k-ra (azaz nem ért véget az eljárás a $(k-1)$-edik lépésben), akkor

$$\overline{\mathbf{x}}_k \in \lin\{\overline{\mathbf{p}}_1, \dots, \overline{\mathbf{p}}_k\} = \lin\{\overline{\mathbf{r}}_0, \dots, \overline{\mathbf{r}}_{k-1}\} =: V_k,$$

valamint $k \geq 2$ esetén

$$\overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_j = 0, \ j = 0, \dots, k-2,$$

és

$$\overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{p}}_j = 0, \ j = 1, \dots, k-1.$$

Bizonyítás. A bizonyítás k-ra vonatkozó teljes indukcióval történik.
A $k = 1$ eset: $\overline{\mathbf{r}}_0 = \overline{\mathbf{p}}_1 \neq \mathbf{0}$ és $\alpha_1 \neq 0$, ezért $\overline{\mathbf{x}}_1 = \alpha_1 \overline{\mathbf{p}}_1 = \alpha_1 \overline{\mathbf{r}}_0$.
Bár a bizonyításhoz nem szükséges, talán hasznos a $k = 2$ eset részletes vizsgálata is: Ha tehát $\overline{\mathbf{r}}_0 \neq \mathbf{0}$ és $\overline{\mathbf{r}}_1 \neq \mathbf{0}$, akkor $\overline{\mathbf{x}}_2 = \overline{\mathbf{x}}_1 + \alpha_2 \overline{\mathbf{p}}_2 = \alpha_1 \overline{\mathbf{p}}_1 + \alpha_2 \overline{\mathbf{p}}_2 = \alpha_1 \overline{\mathbf{r}}_0 + \alpha_2 (\overline{\mathbf{r}}_1 - \beta_1 \overline{\mathbf{r}}_0)$, ami mutatja az állítás első részének helyességét. Továbbá

$$\begin{aligned}
\overline{\mathbf{p}}_2^T \mathbf{A} \overline{\mathbf{p}}_1 &= (\overline{\mathbf{r}}_1^T - \beta_1 \overline{\mathbf{p}}_1^T) \mathbf{A} \overline{\mathbf{p}}_1 = \overline{\mathbf{p}}_1^T \mathbf{A} \overline{\mathbf{r}}_1 - \beta_1 \overline{\mathbf{p}}_1^T \mathbf{A} \overline{\mathbf{p}}_1 = 0, \\
\overline{\mathbf{r}}_1^T \overline{\mathbf{r}}_0 &= (\overline{\mathbf{r}}_0^T - \alpha_1 \overline{\mathbf{p}}_1^T \mathbf{A}) \overline{\mathbf{r}}_0 = \overline{\mathbf{r}}_0^T \overline{\mathbf{r}}_0 - \alpha_1 \overline{\mathbf{p}}_1^T \mathbf{A} \overline{\mathbf{r}}_0 \\
&= \overline{\mathbf{p}}_1^T \overline{\mathbf{r}}_0 - \alpha_1 \overline{\mathbf{p}}_1^T \mathbf{A} \overline{\mathbf{p}}_1 = 0.
\end{aligned}$$

Tegyük fel most, hogy az állítás igaz $k = l$ esetén, azaz $\overline{\mathbf{x}}_l \in \{\overline{\mathbf{p}}_1, \dots, \overline{\mathbf{p}}_l\} = \{\overline{\mathbf{r}}_0, \dots, \overline{\mathbf{r}}_{l-1}\}$, $\overline{\mathbf{r}}_{l-1}^T \overline{\mathbf{r}}_j = 0$ $(j = 1, \dots, l-2)$, $\overline{\mathbf{p}}_l^T \mathbf{A} \overline{\mathbf{p}}_j = 0$ $(j = 1, \dots, l-1)$. Vizsgáljuk a $k = l+1$ esetet!

$$\begin{aligned}
\overline{\mathbf{x}}_{l+1} &= \overbrace{\underbrace{\overline{\mathbf{x}}_l}_{\in \lin(\overline{\mathbf{p}}_1, \dots, \overline{\mathbf{p}}_l)}}^{\in \lin(\overline{\mathbf{r}}_0, \dots, \overline{\mathbf{r}}_{l-1})} + \alpha_{l+1} \underbrace{\overline{\mathbf{p}}_{l+1}}_{= \overline{\mathbf{r}}_l - \beta_l \overline{\mathbf{p}}_l} \in \lin(\overline{\mathbf{p}}_1, \dots, \overline{\mathbf{p}}_{l+1}) \\
&= \lin(\overline{\mathbf{r}}_0, \dots, \overline{\mathbf{r}}_l).
\end{aligned}$$

Most négy különböző esetet kell végignéznünk j értéke szerint.

1. $j < l-1$ esetén
$$\overline{\mathbf{r}}_l^T \overline{\mathbf{r}}_j = (\overline{\mathbf{r}}_{l-1}^T - \alpha_l \overline{\mathbf{p}}_l^T \mathbf{A}) \overline{\mathbf{r}}_j = 0.$$

2. $j = l-1$ esetén, mivel
$$\overline{\mathbf{r}}_{l-1}^T \overline{\mathbf{r}}_{l-1} = \underbrace{\overline{\mathbf{p}}_l^T}_{\overline{\mathbf{r}}_{l-1}^T - \beta_{l-1} \overline{\mathbf{p}}_{l-1}^T} \overline{\mathbf{r}}_{l-1}$$
és
$$\overline{\mathbf{p}}_l^T \mathbf{A} \overline{\mathbf{r}}_{l-1} = \overline{\mathbf{p}}_l^T \mathbf{A} \underbrace{\overline{\mathbf{p}}_l}_{\overline{\mathbf{r}}_{l-1} - \beta_{l-1} \overline{\mathbf{p}}_{l-1}},$$
ezért
$$\overline{\mathbf{r}}_l^T \overline{\mathbf{r}}_{l-1} = (\overline{\mathbf{r}}_{l-1}^T - \alpha_l \overline{\mathbf{p}}_l^T \mathbf{A}) \overline{\mathbf{r}}_{l-1} = 0.$$

3. $j < l$ esetén
$$\overline{\mathbf{p}}_{l+1}^T \mathbf{A}\overline{\mathbf{p}}_j = (\overline{\mathbf{r}}_l^T - \beta_l \overline{\mathbf{p}}_l^T)\mathbf{A}\overline{\mathbf{p}}_j = 0.$$

4. $j = l$ esetén, mivel
$$\overline{\mathbf{r}}_l^T \mathbf{A}\overline{\mathbf{p}}_l = \overline{\mathbf{p}}_l^T \mathbf{A}\overline{\mathbf{r}}_l,$$
ezért
$$\overline{\mathbf{p}}_{l+1}^T \mathbf{A}\overline{\mathbf{p}}_l = (\overline{\mathbf{r}}_l^T - \beta_l \overline{\mathbf{p}}_l^T)\mathbf{A}\overline{\mathbf{p}}_l = 0.$$
Ezt akartuk megmutatni. ■

3.7.7. megjegyzés. A bizonyításbeli azonosságokat felhasználva a következő összefüggések nyerhetők:
$$\alpha_k = \frac{\overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1}}{\overline{\mathbf{p}}_k^T \mathbf{A}\overline{\mathbf{p}}_k},$$
továbbá
$$\overline{\mathbf{r}}_k^T \overline{\mathbf{r}}_k = \overline{\mathbf{r}}_k^T(\overline{\mathbf{r}}_{k-1} - \alpha_k \mathbf{A}\overline{\mathbf{p}}_k) = -\alpha_k \overline{\mathbf{r}}_k^T \mathbf{A}\overline{\mathbf{p}}_k,$$
és
$$\overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1} = \overline{\mathbf{r}}_{k-1}^T(\overline{\mathbf{r}}_k + \alpha_k \mathbf{A}\overline{\mathbf{p}}_k) = \alpha_k \overline{\mathbf{r}}_{k-1}^T \mathbf{A}\overline{\mathbf{p}}_k = \alpha_k \overline{\mathbf{p}}_k^T \mathbf{A}\overline{\mathbf{p}}_k,$$
és így
$$\beta_k = -\frac{\overline{\mathbf{r}}_k^T \overline{\mathbf{r}}_k}{\overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1}}.$$
◇

3.7.8. definíció.
Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy szimmetrikus, pozitív definit mátrix. Egy $\overline{\mathbf{x}} \in \mathbb{R}^n$ *vektor* ***A****-normáján* az $\|\overline{\mathbf{x}}\|_{\mathbf{A}} = \sqrt{\overline{\mathbf{x}}^T \mathbf{A}\overline{\mathbf{x}}}$ értéket értjük.

Vezessük be az $\overline{\mathbf{e}}^{(k)} = \overline{\mathbf{x}}^\star - \overline{\mathbf{x}}_k$ jelölést.

3.7.9. tétel.
Ha $\overline{\mathbf{r}}_{k-1} \neq \mathbf{0}$, akkor $\overline{\mathbf{x}}_k$ az egyetlen pont V_k-ban, melyre $\|\overline{\mathbf{e}}^{(k)}\|_{\mathbf{A}}$ minimális.
$$\|\overline{\mathbf{e}}^{(1)}\|_{\mathbf{A}} \geq \|\overline{\mathbf{e}}^{(2)}\|_{\mathbf{A}} \geq \cdots \geq \|\overline{\mathbf{e}}^{(k)}\|_{\mathbf{A}},$$
továbbá $\overline{\mathbf{e}}^{(k)} = \mathbf{0}$ valamilyen $k \leq n$ esetén.

Bizonyítás. Tekintsük az $\overline{\mathbf{x}}^\star - (\overline{\mathbf{x}}_k - \Delta\overline{\mathbf{x}}) = \overline{\mathbf{e}}^{(k)} + \Delta\overline{\mathbf{x}}$ vektorokat, ahol $\Delta\overline{\mathbf{x}}$ egy tetszőleges V_k-beli vektor.
$$\|\overline{\mathbf{e}}^{(k)} + \Delta\overline{\mathbf{x}}\|_{\mathbf{A}} = (\overline{\mathbf{e}}^{(k)} + \Delta\overline{\mathbf{x}})^T \mathbf{A}(\overline{\mathbf{e}}^{(k)} + \Delta\overline{\mathbf{x}})$$
$$= (\overline{\mathbf{e}}^{(k)})^T \mathbf{A}\overline{\mathbf{e}}^{(k)} + \Delta\overline{\mathbf{x}}^T \mathbf{A}\Delta\overline{\mathbf{x}} + \overbrace{2 \cdot \underbrace{(\overline{\mathbf{e}}^{(k)})^T \mathbf{A}}_{=(\mathbf{A}\overline{\mathbf{e}}^{(k)})^T = (\overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_k)^T = \overline{\mathbf{r}}_k^T} \cdot \underbrace{\Delta\overline{\mathbf{x}}}_{\in V_k}}^{=0}$$
$$= (\overline{\mathbf{e}}^{(k)})^T \mathbf{A}\overline{\mathbf{e}}^{(k)} + \Delta\overline{\mathbf{x}}^T \mathbf{A}\Delta\overline{\mathbf{x}}.$$
Látható tehát, hogy $\Delta\overline{\mathbf{x}} = \mathbf{0}$ esetén lesz a norma a legkisebb, azaz $\overline{\mathbf{x}}_k$ a megoldást legjobban közelítő vektor $\mathbf{A}$-normában a V_k térből.

A $V_1 \subset V_2 \subset \cdots \subset V_k$ tartalmazásokból következik a hibavektorok $\mathbf{A}$-normájának monoton csökkenése.

Ha nem ér véget az eljárás korábban, akkor $V_n = \mathbb{R}^n$, hiszen az $\overline{\mathbf{r}}_k$ vektorok ortogonálisak. Továbbá $\overline{\mathbf{x}}_n$ az $\mathbb{R}^n$-ben $\mathbf{A}$-normában legjobban közelítő vektor, azaz maga az $\overline{\mathbf{x}}^\star$ megoldás. ■

3.7.10. következmény. A tétel közvetlen következménye, hogy a konjugált gradiens-módszer legfeljebb n lépésből véget ér, azaz annyi lépésből, ahány ismeretlenje volt az egyenletrendszernek. Ahogy már korábban említettük, elméletileg (azaz számábrázolási és kerekítési hibák nélkül) a konjugált gradiens-módszer egy direkt módszer. A gyakorlatban viszont a hibával terhelt számítások miatt iterációs módszerként alkalmazzuk. ◇

3.7.11. következmény. Ha a k-adik lépésben leállítjuk az iterációt, akkor megkapjuk V_k-ból a megoldást $\mathbf{A}$-normában legjobban közelítő vektort. ◇

A konjugált gradiens-módszer konvergenciájára vonatkozóan megadunk most bizonyítás nélkül két tételt. A tételek azt mutatják, hogy kis kondíciószámú vagy kevés különböző sajátértékkel rendelkező mátrixok esetén a módszer gyorsan konvergál.

3.7.12. tétel. (Daniel [10], 1967)

Legyen $\mathbf{A}$ szimmetrikus, pozitív definit mátrix, melynek kondíciószáma $\kappa_2(\mathbf{A})$. Ekkor a konjugált gradiens-módszer hibavektorára az alábbi becslés érvényes

$$\|\overline{\mathbf{e}}^{(k)}\|_{\mathbf{A}} \leq 2 \left(\frac{\sqrt{\kappa_2(\mathbf{A})} - 1}{\sqrt{\kappa_2(\mathbf{A})} + 1} \right)^k \|\overline{\mathbf{e}}^{(0)}\|_{\mathbf{A}}.$$

3.7.13. tétel.

Ha az $\mathbf{A}$ mátrixnak s db különböző sajátértéke van, akkor a konjugált gradiens-módszer legfeljebb s lépésben megadja a megoldást.

A korábbi eredmények figyelembevételével az algoritmus az alábbi alakra egyszerűsödik:

Konjugált gradiens-módszer, javított algoritmus.
$\mathbf{A} \in \mathbb{R}^{n \times n}$ pozitív definit mátrix, $\overline{\mathbf{b}} \in \mathbb{R}^n$ adott vektor.

$k := 0$, $\overline{\mathbf{r}}_0 := \overline{\mathbf{b}}$, $\overline{\mathbf{x}}_0 := \mathbf{0}$, $\overline{\mathbf{p}}_1 = \overline{\mathbf{r}}_0$
while $\overline{\mathbf{r}}_k \neq \mathbf{0}$ **do**
  $k := k + 1$
  $\alpha_k := \overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1} / (\overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{p}}_k)$ $(2n - 1 + 2n^2 + n - 1$ flop$)$
  $\overline{\mathbf{x}}_k := \overline{\mathbf{x}}_{k-1} + \alpha_k \overline{\mathbf{p}}_k$ $(2n$ flop$)$
  $\overline{\mathbf{r}}_k := \overline{\mathbf{r}}_{k-1} - \alpha_k \mathbf{A} \overline{\mathbf{p}}_k$ $(2n$ flop$)$
  $\beta'_k := \overline{\mathbf{r}}_k^T \overline{\mathbf{r}}_k / (\overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1})$ $(2n - 1$ flop$)$
  $\overline{\mathbf{p}}_{k+1} := \overline{\mathbf{r}}_k + \beta'_k \overline{\mathbf{p}}_k$ $(2n$ flop$)$
end while

A konjugált gradiens-módszer műveletigénye $2n^2 + 11n - 3 = \mathcal{O}(2n^2)$ iterációnként. Így ha $n/3$-nál több lépés kellene, akkor a Gauss-módszer gyorsabb nála. Emiatt inkább ritka mátrixokra

alkalmazzák. Előnye a korábban ismertetett iterációs módszerekkel szemben, hogy nem kell hozzá optimális relaxációs paramétert keresni a gyors konvergencia eléréséhez.

3.7.14. megjegyzés. A konjugált gradiens-módszert gyakran használják prekondicionáltan. Ez azt jelenti, hogy ekvivalens átalakításokkal olyan alakra hozzák az egyenletrendszert, melynek együtthatómátrixának kondíciószáma jóval kisebb, mint az eredeti egyenletrendszer mátrixáé, ugyanakkor a megoldása közel azonos művelettel jár. A mi esetünkben lehet $\mathbf{C}$ szimmetrikus, pozitív definit mátrix, melyre $\mathbf{C}^2$ közelíti $\mathbf{A}$-t és $\mathbf{C}^2$ könnyen invertálható. Tekintsük ekkor a

$$(\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1})(\mathbf{C}\overline{\mathbf{x}}) = \mathbf{C}^{-1}\overline{\mathbf{b}}$$

egyenletrendszert. Ha ezt a konjugált gradiens-módszerrel oldjuk meg, akkor bár minden iterációs lépésben egy $\mathbf{C}^2$ együtthatómátrixú lineáris egyenletrendszert kell megoldanunk, az eljárás gyorsan konvergálhat, mert a mátrix jól kondicionált. ⋄

3.8. A QR-felbontás

Korábban már láttunk kétfajta mátrixfelbontást. Az egyik a Schur-felbontás volt, amely szerint minden mátrix unitér hasonlósági transzformációval felső háromszögmátrixba vihető. A másik az LU-felbontás, amely szerint bizonyos mátrixok felírhatók egy normált alsó és egy felső háromszögmátrix szorzataként. Mindkét felbontás hasznosnak bizonyult az alkalmazások során. Ebben a fejezetben egy újabb felbontással fogunk megismerkedni: a QR-felbontással, amely hasonlóan hasznos eszköz lesz a későbbiekben. A QR-felbontás egy ortogonális és egy felső háromszögmátrix szorzataként állítja elő a teljes oszloprangú mátrixokat. Egy mátrixot *teljes oszloprangúnak* hívunk, ha rangja megegyezik oszlopainak számával. Nyilvánvalóan ez csak úgy lehet, ha sorainak száma legalább annyi, mint oszlopaié.

Tulajdonképpen már találkoztunk is olyan eljárással, amely QR-felbontást ad. Az 1.1.30. tétel szerint a Gram–Schmidt ortogonalizációs eljárás során a lineárisan független $\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_n \in \mathbb{R}^m$ $(n \leq m)$ vektorokhoz megadhatók olyan ortonormált $\overline{\mathbf{q}}_1, \ldots, \overline{\mathbf{q}}_n$ vektorok, hogy $\operatorname{lin}(\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_l) = \operatorname{lin}(\overline{\mathbf{q}}_1, \ldots, \overline{\mathbf{q}}_l)$ $(l \leq n)$. Így tehát vannak olyan α_{ij} számok, melyekkel

$$\begin{aligned}
\overline{\mathbf{x}}_1 &= \alpha_{11}\overline{\mathbf{q}}_1 \\
\overline{\mathbf{x}}_2 &= \alpha_{12}\overline{\mathbf{q}}_1 + \alpha_{22}\overline{\mathbf{q}}_2 \\
&\vdots \\
\overline{\mathbf{x}}_n &= \alpha_{1n}\overline{\mathbf{q}}_1 + \ldots + \alpha_{nn}\overline{\mathbf{q}}_n.
\end{aligned}$$

Ha tehát a teljes rangú $\mathbf{A} \in \mathbb{R}^{m\times n}$ mátrix oszlopvektorai $\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_n$, akkor $\mathbf{A}$ felírható az

$$\mathbf{A} = [\overline{\mathbf{q}}_1, \ldots, \overline{\mathbf{q}}_n] \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ 0 & \alpha_{22} & \ldots & \alpha_{2n} \\ & & & \\ 0 & 0 & \ldots & \alpha_{nn} \end{bmatrix}$$

alakban. Ezt a felbontást redukált QR-felbontásnak hívjuk. Az első tényező egy ortonormált vektorokat tartalmazó $(m \times n)$-es mátrix, a második pedig egy $(n \times n)$-es felső háromszögmátrix. Ha most kiegészítjük a $\overline{\mathbf{q}}_1, \ldots, \overline{\mathbf{q}}_n$ vektorokat ortonormált bázissá a $\overline{\mathbf{q}}_{n+1}, \ldots, \overline{\mathbf{q}}_m$ vektorokkal,

akkor az

$$\mathbf{A} = [\overline{\mathbf{q}}_1, \ldots, \overline{\mathbf{q}}_n, \overline{\mathbf{q}}_{n+1}, \ldots, \overline{\mathbf{q}}_m] \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ 0 & \alpha_{22} & \ldots & \alpha_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \alpha_{nn} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

előállításban az első mátrix egy $m \times m$-es ortogonális mátrix, a második pedig egy $m \times n$ méretű felső háromszögmátrix. Ezzel elő is állt egy QR-felbontás. Vegyük észre, hogy a $\mathbf{Q}$ mátrix $n+1 : m$ oszlopai az $\mathbf{A}$ mátrix előállításában nem vesznek részt, hiszen ezek az $\mathbf{R}$ mátrix nulla sorainak elemeivel szorzódnak. Természetesen nem véletlen, hogy a Gram–Schmidt-eljárás során ritkán említik csak, hogy alkalmas QR-felbontás létrehozására. Ennek oka az, hogy a numerikus számítások során felhalmozódó kerekítési és ábrázolási hibák eléggé pontatlanná teszik a módszert. Most megismerünk két másik, sokkal gyakrabban alkalmazott eljárást QR-felbontás előállítására.

3.8.1. QR-felbontás Householder-tükrözésekkel

Legyen adott a $\overline{\mathbf{v}} \neq \mathbf{0}$ normálvektorával az $\mathbb{R}^n$ térben egy $(n-1)$-dimenziós hipersík. Legyen adott továbbá egy tetszőleges $\overline{\mathbf{x}}$ vektor. Adjuk meg $\overline{\mathbf{x}}$-nek a hipersíkra való tükörképét! Mivel az $\overline{\mathbf{x}}$ vektornak a hipersíkra merőleges komponense

$$\frac{1}{\|\overline{\mathbf{v}}\|_2}\overline{\mathbf{v}}\frac{\overline{\mathbf{v}}^T\overline{\mathbf{x}}}{\|\overline{\mathbf{v}}\|_2},$$

ezért a tükörkép úgy adódik, hogy $\overline{\mathbf{x}}$-ből kivonjuk ezen vektor kétszeresét. Tehát a tükörkép

$$\overline{\mathbf{x}} - 2\frac{1}{\|\overline{\mathbf{v}}\|_2}\overline{\mathbf{v}}\frac{\overline{\mathbf{v}}^T\overline{\mathbf{x}}}{\|\overline{\mathbf{v}}\|_2} = \overline{\mathbf{x}} - 2\frac{\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\overline{\mathbf{x}} = \left(\mathbf{E} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\right)\overline{\mathbf{x}}.$$

Ezek alapján egy adott $\overline{\mathbf{v}} \neq 0$ normálvektorú hipersíkra úgy tükrözhetjük a vektorokat, hogy azokat a

$$\mathbf{H} = \mathbf{E} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}$$

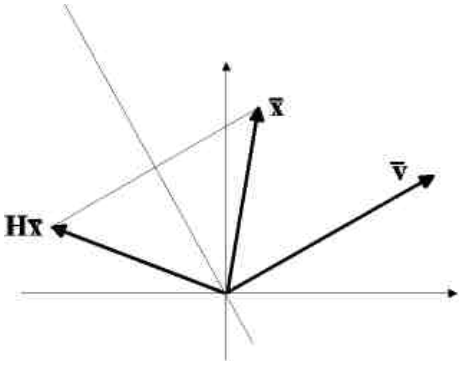
mátrixszal szorozzuk. A tükrözést a $\overline{\mathbf{v}}$ vektor egyértelműen meghatározza. A tükrözést a 3.8.1. ábra szemlélteti.

3.8.1. tétel.
Egy adott $\overline{\mathbf{v}} \neq \mathbf{0}$ vektor esetén a $\mathbf{H}$ mátrix szimmetrikus és ortogonális.

Bizonyítás. A szimmetria látszik, hiszen transzponálva visszakapjuk $\mathbf{H}$-t. Az ortogonalitáshoz azt mutatjuk meg, hogy a mátrix inverze saját maga, mert ez a szimmetria miatt elegendő az ortogonalitáshoz.

$$\left(\mathbf{E} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\right)\left(\mathbf{E} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\right) = \mathbf{E} - 4\frac{\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}} + 4\frac{\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\frac{\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}} = \mathbf{E}. \blacksquare$$

A tétel következménye tehát, hogy tetszőleges $\overline{\mathbf{x}} \in \mathbb{R}^n$ vektorra $\mathbf{H}(\mathbf{H})\overline{\mathbf{x}} = \mathbf{E}\overline{\mathbf{x}} = \overline{\mathbf{x}}$, ami valóban megfelel a tükrözési tulajdonságnak. Az is könnyen látszik, hogy a $\overline{\mathbf{v}}$ vektorra merőleges vektorok, azaz a hipersík vektorai helyben maradnak a szorzás során.



3.8.1. ábra:

Viszgáljuk meg most azt a kérdést, hogy milyen $\overline{\mathbf{v}}$ vektorral visz a tükrözés egy $\overline{\mathbf{x}}$ vektort egy $\overline{\mathbf{e}}_1$-gyel párhuzamos vektorba! Azaz szeretnénk az $\overline{\mathbf{x}}$ vektorhoz egy olyan $\overline{\mathbf{v}}$ vektort találni, hogy a vele konstruált $\mathbf{H}$ tükrözésre (sematikusan)

$$\mathbf{H}\overline{\mathbf{x}} = \mathbf{H}\begin{bmatrix} * \\ * \\ \vdots \\ * \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

teljesüljön. Mivel

$$\underbrace{\mathbf{H}\overline{\mathbf{x}}}_{\in \mathrm{lin}(\overline{\mathbf{e}}_1)} = \overline{\mathbf{x}} - \frac{2\overline{\mathbf{v}}^T\overline{\mathbf{x}}}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\overline{\mathbf{v}},$$

így $\overline{\mathbf{v}} \in \mathrm{lin}(\overline{\mathbf{x}}, \overline{\mathbf{e}}_1)$. Legyen $\overline{\mathbf{v}} = \overline{\mathbf{x}} + \alpha\overline{\mathbf{e}}_1$ valamilyen α konstanssal.

Ekkor

$$\begin{aligned}
\mathbf{H}\overline{\mathbf{x}} &= \overline{\mathbf{x}} - \frac{2(\overline{\mathbf{x}}^T + \alpha\overline{\mathbf{e}}_1^T)\overline{\mathbf{x}}}{(\overline{\mathbf{x}} + \alpha\overline{\mathbf{e}}_1)^T(\overline{\mathbf{x}} + \alpha\overline{\mathbf{e}}_1)}(\overline{\mathbf{x}} + \alpha\overline{\mathbf{e}}_1) \\
&= \overline{\mathbf{x}} - 2\frac{\overline{\mathbf{x}}^T\overline{\mathbf{x}} + \alpha x_1}{\overline{\mathbf{x}}^T\overline{\mathbf{x}} + 2\alpha x_1 + \alpha^2}\overline{\mathbf{x}} - \alpha\frac{2\overline{\mathbf{v}}^T\overline{\mathbf{x}}}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\overline{\mathbf{e}}_1 \\
&= \left(1 - 2\frac{\|\overline{\mathbf{x}}\|_2^2 + \alpha x_1}{\|\overline{\mathbf{x}}\|_2^2 + 2\alpha x_1 + \alpha^2}\right)\overline{\mathbf{x}} - \alpha\frac{2\overline{\mathbf{v}}^T\overline{\mathbf{x}}}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\overline{\mathbf{e}}_1.
\end{aligned}$$

Innét egyszerre adódik, hogy $\overline{\mathbf{x}}$ szorzója nullává tehető az $\alpha = \pm\|\overline{\mathbf{x}}\|_2$ választással. Tehát adott $\overline{\mathbf{x}} \neq \mathbf{0}$ esetén $\overline{\mathbf{v}} = \overline{\mathbf{x}} \pm \|\overline{\mathbf{x}}\|_2\overline{\mathbf{e}}_1$ jó választás. Ekkor ugyanis

$$\mathbf{H}\overline{\mathbf{x}} = \mp\|\overline{\mathbf{x}}\|_2\frac{2(\overline{\mathbf{x}} \pm \|\overline{\mathbf{x}}\|_2\overline{\mathbf{e}}_1)^T\overline{\mathbf{x}}}{(\overline{\mathbf{x}} \pm \|\overline{\mathbf{x}}\|_2\overline{\mathbf{e}}_1)^T(\overline{\mathbf{x}} \pm \|\overline{\mathbf{x}}\|_2\overline{\mathbf{e}}_1)}\overline{\mathbf{e}}_1 = \mp\|\overline{\mathbf{x}}\|_2\frac{2\|\overline{\mathbf{x}}\|_2^2 \pm 2\|\overline{\mathbf{x}}\|_2 x_1}{2\|\overline{\mathbf{x}}\|_2^2 \pm 2\|\overline{\mathbf{x}}\|_2 x_1}\overline{\mathbf{e}}_1 = \mp\|\overline{\mathbf{x}}\|_2\overline{\mathbf{e}}_1.$$

3.8.2. megjegyzés. Ha $x_1 \neq 0$, akkor célszerű a $\overline{\mathbf{v}} = \overline{\mathbf{x}} + \mathrm{sgn}(x_1)\|\overline{\mathbf{x}}\|_2\overline{\mathbf{e}}_1$ választás a kiegyszerűsödés elkerülése érdekében.[9]

Mivel $\overline{\mathbf{v}}$-nek csak az iránya fontos, a nagysága nem, azért célszerű úgy normálni, hogy az első eleme 1 legyen. Ekkor a $\overline{\mathbf{v}}$ vektor tárolható az $\overline{\mathbf{x}}$ vektor lenullázott elemeinek helyén.

[9] Az sgn rövidítés a signum függvényt jelöli: $\mathrm{sgn}(x)$ pozitív x értékekre $+1$, negatívokra -1 és $x = 0$ esetén 0.

Legyen $\mathbf{C}$ egy tetszőleges mátrix. Ekkor a $\mathbf{HC}$ szorzatot a

$$\mathbf{HC} = \left(\mathbf{E} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\right)\mathbf{C} = \mathbf{C} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\mathbf{C} = \mathbf{C} + \overline{\mathbf{v}}\underbrace{\left(-\frac{2\overline{\mathbf{v}}^T\mathbf{C}}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\right)}_{=:\overline{\mathbf{w}}^T} = \mathbf{C} + \overline{\mathbf{v}}\overline{\mathbf{w}}^T$$

képlettel érdemes számítani a jóval több műveletet igénylő tényleges mátrixszorzás helyett. ◇

3.8.3. definíció.

Legyen $\overline{\mathbf{x}} \in \mathbb{R}^m$ tetszőleges vektor. Ekkor azt a $\mathbf{H}$ mátrixot, amely az $\overline{\mathbf{x}}$ vektort az $\overline{\mathbf{e}}_1$ egységvektorral párhuzamos vektorba viszi (az $\overline{\mathbf{x}}$ vektorhoz tartozó) *Householder*[10]*-tükrözésnek* nevezzük.

A korábbiak alapján adott $\overline{\mathbf{x}}$ vektorra úgy állítható elő a Householder-tükrözés, hogy az $\overline{\mathbf{x}}$ vektorból meghatározzuk a $\overline{\mathbf{v}} = \overline{\mathbf{x}} \pm \|\overline{\mathbf{x}}\|_2\overline{\mathbf{e}}_1$ képlettel a tükörsík normálvektorát, majd ezzel a $\overline{\mathbf{v}}$ vektorral képezzük a $\mathbf{H}$ tükrözési mátrixot. Látható, hogy egy adott $\overline{\mathbf{x}}$ vektorhoz több Householder-mátrix is létezik.

Householder-tükrözésekkel könnyen előállítható egy teljes oszloprangú mátrix QR-felbontása.

3.8.4. tétel.

Legyen $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(m \geq n)$ egy teljes oszloprangú mátrix. Ekkor léteznek olyan $\mathbf{Q} \in \mathbb{R}^{m\times m}$ ortogonális és $\mathbf{R} \in \mathbb{R}^{m\times n}$ felső háromszögmátrixok, melyekkel $\mathbf{A} = \mathbf{QR}$.

Bizonyítás. A tételt úgy bizonyítjuk, hogy Householder-tükrözésekkel előállítjuk a keresett felbontást. Legyen $\mathbf{H}_1$ az $\mathbf{A}(1:m,1)$ oszlophoz tartozó Householder-tükrözés. Ekkor $\mathbf{A}^{(2)} := \mathbf{H}_1\mathbf{A}$ első oszlopának $2:m$ elemei nullák. Legyen $\tilde{\mathbf{H}}_2$ az $\mathbf{A}^{(2)}(2:m,2)$ oszlophoz tartozó Householder-mátrix. Legyen továbbá $\mathbf{H}_2 = \operatorname{diag}(1, \tilde{\mathbf{H}}_2)$. Ekkor $\mathbf{A}^{(3)} := \mathbf{H}_2\mathbf{A}^{(2)}$ első oszlopának $2:m$ elemei ill. második oszlopának $3:m$ elemei nullák. A teljes rang miatt az eljárás tovább folytatható. Így a

$$\mathbf{H}_n \cdot \cdots \cdot \mathbf{H}_1 \cdot \mathbf{A} = \mathbf{R}$$

előállítást nyerjük, ahol $\mathbf{R}$ felső háromszögmátrix. A $\mathbf{Q}^T := \mathbf{H}_n \cdot \cdots \cdot \mathbf{H}_1$ mátrix ortogonális, így a fenti jelölésekkel $\mathbf{A} = \mathbf{QR}$. ■

3.8.2. QR-felbontás Givens-forgatásokkal

Ismert, hogy a θ szögű elforgatottja egy $\mathbb{R}^2$-beli $\overline{\mathbf{x}}$ vektornak az az $\overline{\mathbf{x}}'$ vektor, melyre

$$\overline{\mathbf{x}}' = \underbrace{\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}}_{=:\mathbf{G}_\theta}\overline{\mathbf{x}}.$$

Jelöljük a forgatási mátrixot $\mathbf{G}_\theta$-val. Könnyen ellenőrizhető, hogy ez a mátrix ortogonális.

Vizsgáljuk meg, hogy a $\mathbf{G}_\theta$ mátrix milyen θ szög esetén visz egy kételemű $\overline{\mathbf{x}}$ oszlopvektort olyan vektorba, melynek második eleme nulla, azaz milyen $\mathbf{G}_\theta$ mátrixszal lesz

$$\begin{bmatrix} * \\ 0 \end{bmatrix} = \mathbf{G}_\theta \begin{bmatrix} * \\ * \end{bmatrix}.$$

Vegyük észre, hogy a θ szög értékére nincs is szükségünk, elegendő az $s = \sin\theta$ és $c = \cos\theta$ értékek meghatározása. Könnyen látszik, hogy ha $x_2 = 0$, akkor $s = 0$, $c = 1$ jó választás. Ha $x_2 \neq 0$, akkor az $sx_1 + cx_2 = 0$, $s^2 + c^2 = 1$ egyenletrendszer megoldásából

$$s = \frac{\pm x_2}{\sqrt{x_1^2 + x_2^2}}, \quad c = \frac{\mp x_1}{\sqrt{x_1^2 + x_2^2}}$$

adódik. Összefoglalva tehát, bármely kételemű vektor megszorozható úgy egy ortogonális mátrixszal, hogy az eredményvektor második eleme nulla legyen.

Térjünk át most a többdimenziós esetre. Milyen ortogonális mátrixszal szorozzunk meg egy adott $\overline{\mathbf{x}} \in \mathbb{R}^m$ vektort ahhoz, hogy a j-edik eleme lenullázódjon a szorzás után ($j = 2, \ldots, m$), és az i-edik elemen kívül a többi elem változatlan maradjon ($i < j$)? A kétdimenziós esetet általánosítva könnyen adódik a válasz. A keresett mátrixnak az alábbi alakúnak kell lennie.

$$\mathbf{G}(i,j,\theta) = \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & c & & & -s & & \\ & & & 1 & & & & \\ & & & & \ddots & & & \\ & & & & & 1 & & \\ & & s & & & & c & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{bmatrix} \in \mathbb{R}^{m\times m}, \tag{3.8.1}$$

ahol az $s = \sin\theta$ és $c = \cos\theta$ értékek az i-edik ill. j-edik sorokban és oszlopokban helyezkednek el, és a kétdimenziós esetnél ismertetett módon az $\overline{\mathbf{x}}$ vektor x_i és x_j elemeiből számíthatók. Ez a mátrix tulajdonképpen egy, az $\{(0,\ldots,0,\underbrace{u}_{i.},0,\ldots,0,\underbrace{v}_{j.},0\ldots,0) \in \mathbb{R}^m \mid u,v \in \mathbb{R}\}$ hipersíkbeli θ szögű forgatást ír le.

3.8.5. definíció.

Azt a $\mathbf{G}(i,j,\theta)$ (3.8.1) alakú mátrixot, amely egy adott $\overline{\mathbf{x}} \in \mathbb{R}^m$ vektort úgy változtat meg a vele való szorzás során, hogy a j-edik ($j = 2,\ldots,m-1$) eleme nulla lesz, és az i-edik elem kivételével a többi elem nem változik, (az $\overline{\mathbf{x}}$ vektorhoz tartozó (i,j) indexű) Givens-forgatásnak hívjuk.

A Givens[11]-forgatás nem egyértelmű, azaz egy adott vektor adott két eleméhez többfajta, a feltételeknek megfelelő mátrix is megadható.

Egy teljes oszloprangú mátrix QR-felbontása előállítható Givens-forgatások segítségével is. Először megkeressük az első oszlophoz tartozó $(m-1,m)$ indexű Givens-forgatást, majd ezzel szorozzuk a mátrixot. Ekkor a mátrix m-edik sorának első eleme lenullázódik. Ezután szorzunk az első oszlop $(m-2,m-1)$ indexű Givens-forgatásával, stb. Ezt folytatjuk addig, míg az első oszlopban az első elem kivételével minden elem lenullázódott. Ezután folytatjuk az eljárást a második oszloppal, kinullázva az oszlopban a főátló alatti elemeket. Stb. Az egyes lépések nem rontják el a korábban lenullázott elemeket. Az eljárás végén egy felső háromszögmátrixot kapunk, ez lesz az $\mathbf{R}$ mátrix, a Givens-forgatások szorzatának transzponáltja pedig a $\mathbf{Q}$ mátrix.

[11] Wallace Givens, 1910-1993 (USA)

Az eljárás sematikusan a következőképpen néz ki:

$$\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ 0 & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ * & * & * \\ 0 & * & * \\ 0 & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \end{bmatrix}$$

$$\begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \end{bmatrix}$$

Ahogy láttuk, a QR-felbontást háromféleképpen is elő tudjuk állítani: Gram-Schmidt ortogonalizációval, Householder-tükrözésekkel és Givens-forgatásokkal. Vajon ezek ugyanazt a felbontást állítják elő? Ezt vizsgálja az alábbi tétel.

3.8.6. tétel.

Egy teljes oszloprangú mátrix esetén a redukált QR-felbontás egyértelmű, ha megköveteljük, hogy $\mathbf{R}$ főátlójában pozitív elemek legyenek.

Bizonyítás. Legyen $\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1$ a teljes oszloprangú $\mathbf{A} \in \mathbb{R}^{m\times n}$ mátrix redukált QR-felbontása. Ekkor $\mathbf{A}^T\mathbf{A}$ szimmetrikus és a teljes oszloprang miatt pozitív definit is, továbbá

$$\mathbf{A}^T\mathbf{A} = (\mathbf{Q}_1\mathbf{R}_1)^T(\mathbf{Q}_1\mathbf{R}_1) = \mathbf{R}_1^T\mathbf{R}_1.$$

Ha $\mathbf{R}_1$ főátlójában pozitív elemeknek kell állniuk, akkor a fenti előállítás éppen az $\mathbf{A}^T\mathbf{A}$ mátrix Cholesky-felbontását adja, ami egyértelmű. Mivel $\mathbf{Q}_1 = \mathbf{A}\mathbf{R}_1^{-1}$, azért $\mathbf{Q}_1$ is egyértelműen meghatározott. ■

Householder-tükrözések esetén a QR-felbontáshoz szükséges tükrözési mátrixok előállításának műveletigénye $2n^2(m - n/3)$ flop, míg a Givens-forgatások előállítása $3n^2(m - n/3)$ flop. Így azt mondhatjuk, hogy a Hoseholder tükrözések alkalmazása általában gyorsabb, mint a Givens-forgatásoké. A Givens-forgatások előnye viszont abban áll, hogy segítségükkel elemenként tudjuk lenullázni az eredeti mátrix elemeit. Pl. felső Hessenberg mátrixok QR-felbontásánál ez nagyon praktikus, hiszen csak a szubdiagonál elemeit kell lenulláznunk. Sematikusan ez a következő módon tehető meg, az i-edik lépésben az i-edik oszlophoz keressük meg az $(i, i + 1)$ indexű Givens-forgatást:

$$\begin{bmatrix} * & * & * \\ * & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \end{bmatrix}$$

A QR-felbontást a következő fejezetben túlhatározott lineáris egyenletrendszerek megoldására fogjuk használni. További alkalmazásával a sajátértékmeghatározási módszereknél fogunk majd találkozni.

3.9. Túlhatározott rendszerek megoldása

Ebben a fejezetben olyan lineáris egyenletrendszerek megoldását keressük meg, melyek legalább annyi egyenletet tartalmaznak ahány ismeretlenje van az egyenletrendszernek. Az ilyen egyenletrendszereket túlhatározott egyenletrendszereknek nevezzük. Az egyszerűség kedvéért ide soroljuk

a négyzetes mátrixú egyenletrendszereket is, azaz azt az esetet, amikor ugyanannyi egyenlet van mint ismeretlen. Az együtthatómátrixról feltesszük továbbá, hogy teljes oszloprangú. Összefoglalva tehát az

$$\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}, \quad \mathbf{A} \in \mathbb{R}^{m\times n},\ m \geq n,\ r(\mathbf{A}) = n$$

alakú egyenletrendszereket vizsgáljuk.

Ismert, hogy a fenti lineáris egyenletrendszernek vagy csak egy megoldása van (ha $\overline{\mathbf{b}}$ előáll $\mathbf{A}$ oszlopvektoraival) vagy nincs megoldása (ha nem áll elő). Ha van megoldás, akkor azt a Gauss-módszerrel határozhatjuk meg, hasonlóan ahhoz az esethez, amikor az együtthatómátrix invertálható négyzetes mátrix volt. Gauss-transzformációk és esetleges sorcserék segítségével az $\mathbf{A}$ mátrixot felső háromszög alakra hozzuk, melynek főátlójában nullától különböző elemek fognak állni. Ha most a $\overline{\mathbf{b}}(n+1:m)$ nem nullvektor, akkor az egyenletrendszernek nincs megoldása, ha pedig nullvektor, akkor az egyértelmű megoldás egyszerű visszahelyettesítéssel adódik.

A fentiek alapján általában azt mondhatjuk tehát, hogy egy túlhatározott lineáris egyenletrendszernek nincs megoldása. Hasznos azonban a megoldás fogalmát egy kicsit kiterjeszteni. Nevezzük egy túlhatározott lineáris egyenletrendszer esetén legkisebb négyzetek (angolul least square, rövidítve LS) értelemben legjobb megoldásnak, röviden LS-megoldásnak, azt az $\overline{\mathbf{x}} \in \mathbb{R}^n$ vektort, melyre $\|\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}\|_2^2$ a lehető legkisebb.

Legyen

$$\phi(\overline{\mathbf{x}}) = \|\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}\|_2^2,$$

továbbá legyen $\overline{\mathbf{z}}$ egy tetszőleges nemnulla vektor ($\mathbf{A}\overline{\mathbf{z}} \neq \mathbf{0}$), és α egy tetszőleges valós szám. Ekkor

$$\begin{aligned}\phi(\overline{\mathbf{x}} + \alpha\overline{\mathbf{z}}) &= \|\mathbf{A}(\overline{\mathbf{x}} + \alpha\overline{\mathbf{z}}) - \overline{\mathbf{b}}\|_2^2 \\ &= \|\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}\|_2^2 + \alpha^2\|\mathbf{A}\overline{\mathbf{z}}\|_2^2 + 2\alpha\overline{\mathbf{z}}^T\mathbf{A}^T(\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}) \\ &= \|\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}\|_2^2 + \alpha\|\mathbf{A}\overline{\mathbf{z}}\|_2^2\left(\alpha + \frac{2\overline{\mathbf{z}}^T\mathbf{A}^T(\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}})}{\|\mathbf{A}\overline{\mathbf{z}}\|_2^2}\right).\end{aligned}$$

Megmutatjuk, hogy az LS-megoldás nem lesz más, mint az

$$\mathbf{A}^T\mathbf{A}\overline{\mathbf{x}} = \mathbf{A}^T\overline{\mathbf{b}} \tag{3.9.1}$$

ún. normálegyenlet megoldása ($\mathbf{A}^T\mathbf{A}$ négyzetes mátrix), amely nyilvánvalóan létezik, hiszen $\mathbf{A}$ teljes oszloprangú. Jelölje ezt a megoldást $\overline{\mathbf{x}}_{LS}$. Ekkor

$$\phi(\overline{\mathbf{x}}_{LS} + \alpha\overline{\mathbf{z}}) = \|\mathbf{A}\overline{\mathbf{x}}_{LS} - \overline{\mathbf{b}}\|_2^2 + \alpha^2\|\mathbf{A}\overline{\mathbf{z}}\|_2^2 = \phi(\overline{\mathbf{x}}_{LS}) + \alpha^2\|\mathbf{A}\overline{\mathbf{z}}\|_2^2.$$

Azt kell megmutatnunk, hogy csak egy vektor minimalizálhatja ϕ-t. Ha $\overline{\mathbf{y}}$-nál is minimum lenne, akkor legyen $\alpha = 1$ és $\overline{\mathbf{z}} = \overline{\mathbf{y}} - \overline{\mathbf{x}}_{LS}$, azaz

$$\begin{aligned}&\phi(\overline{\mathbf{x}}_{LS} + (\overline{\mathbf{y}} - \overline{\mathbf{x}}_{LS})) \\ &= \phi(\overline{\mathbf{y}}) = \phi(\overline{\mathbf{x}}_{LS}) + \|\mathbf{A}(\overline{\mathbf{y}} - \overline{\mathbf{x}}_{LS})\|_2^2 \neq \phi(\overline{\mathbf{x}}_{LS}),\end{aligned}$$

ha $\overline{\mathbf{y}} \neq \overline{\mathbf{x}}_{LS}$.

Megoldás a normálegyenlet segítségével

Az LS-megoldás tehát a (3.9.1) egyenlet segítségével határozható meg $\overline{\mathbf{x}}_{LS} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\overline{\mathbf{b}}$ alakban. Természetesen a gyakorlatban nem így számoljuk. Mivel $\mathbf{A}^T\mathbf{A}$ szimmetrikus, pozitív definit mátrix, ezért előállítható a Cholesky-felbontása $\mathbf{L}\mathbf{L}^T$ alakban. Ezután megoldjuk az $\mathbf{L}\overline{\mathbf{y}} = \mathbf{A}^T\overline{\mathbf{b}}$ egyenletrendszert, majd $\overline{\mathbf{x}}_{LS}$ az $\mathbf{L}^T\overline{\mathbf{x}} = \overline{\mathbf{y}}$ egyenletrendszer megoldásaként adódik. Az eljárás műveletigény: $(m + n/3)n^2$ flop.

Megoldás a QR-felbontás segítségével

Legyen $\mathbf{A} = \mathbf{QR}$ az $\mathbf{A}$ mátrix QR-felbontása. Az

$$\begin{aligned}\|\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}\|_2^2 &= \|\mathbf{QR}\overline{\mathbf{x}} - \overline{\mathbf{b}}\|_2^2 = \|\mathbf{Q}^T(\mathbf{QR}\overline{\mathbf{x}} - \overline{\mathbf{b}})\|_2^2 \\ &= \|\mathbf{R}\overline{\mathbf{x}} - \mathbf{Q}^T\overline{\mathbf{b}}\|_2^2 = \|\mathbf{R}_1\overline{\mathbf{x}} - \overline{\mathbf{c}}\|_2^2 + \|\overline{\mathbf{d}}\|_2^2\end{aligned}$$

egyenlőség miatt, ahol $\mathbf{R}_1 = \mathbf{R}(1:n, 1:n)$, $\overline{\mathbf{c}} = (\mathbf{Q}^T\overline{\mathbf{b}})(1:n, :)$ és $\overline{\mathbf{d}} = (\mathbf{Q}^T\overline{\mathbf{b}})(n+1:m, :)$, az LS-megoldás az alábbi módon határozható meg.

- Képezzük az $\mathbf{A}$ mátrix QR-felbontását.
- Meghatározzuk az $\mathbf{R}_1 = \mathbf{R}(1:n, 1:n)$ mátrixot.
- Meghatározzuk a $\overline{\mathbf{c}} = (\mathbf{Q}^T\overline{\mathbf{b}})(1:n, :)$ vektort.
- Az $\overline{\mathbf{x}}_{LS}$ LS-megoldás az $\mathbf{R}_1\overline{\mathbf{x}} = \overline{\mathbf{c}}$ egyenletrendszer megoldásaként adódik.

Az eljárás műveletigénye: $2(m - n/3)n^2$ flop.

3.9.1. megjegyzés.

- $m >> n$ esetén a QR-felbontásos megoldás műveletszáma kb. kétszerese a normálegyenletesnek.
- Négyzetes, teljes rangú mátrixokra a műveletszám mindkét esetben $4n^3/3$, ami a Gauss-módszer kétszerese. Viszont ha a memóriakezelést is figyelembe vesszük, akkor a teljes futási idő összemérhető a Gauss-módszerrel, sőt stabil is, hiszen nincs növekedési faktor.
- Ranghiányos vagy közel ranghiányos esetben ezek a módszerek nem használhatók.
- A normálegyenletre alkalmazható a konjugált gradiens-módszer, de akkor az új rendszer kondíciószáma az eredeti négyzete lesz.
- Ha $\mathbf{A} \in \mathbb{R}^{n\times n}$, akkor $\overline{\mathbf{x}}_{LS} = \mathbf{A}^{-1}\overline{\mathbf{b}}$, ami az egyenletrendszer klasszikus értelemben vett megoldása.

◇

3.10. Lineáris egyenletrendszerek megoldása a MATLAB-ban

Lineáris egyenletrendszerek megoldása a MATLAB programban egyetlen paranccsal (\) történik függetlenül attól, hogy egyértelmű megoldás van vagy esetleg túlhatározott az egyenletrendszer. Első lépésben megpróbálja a MATLAB előállítani a Cholesky-felbontást (`chol`), ha ez nem sikerül (hamar kiderül, hogy nem szimmetrikus, pozitív definit a mátrix, így ez nem vesz el sok időt), akkor az LU-felbontással határozza meg a megoldást (Gauss-módszer). Ha az egyenlet túlhatározott, akkor először meghatározza a QR-felbontást (Householder-tükrözésekkel), majd ebből a legkisebb négyzetek értelemben legjobban közelítő megoldást. A `sparse` parancs segítségével tudathatjuk a MATLAB-bal, hogy egy mátrix ritka. Ebben az esetben a mátrix szerkezetének megfelelő megoldási módot fogja használni a MATLAB. A \ parancsról részletes angol nyelvű leírás található a http://www.weizmann.ac.il/matlab/techdoc/ref/arithmeticoperators.html#8559 oldalon.

Lássunk néhány szemléletes példát!

```
>> A=[2,-1;-1,2]; B=[1;1]; b=[1;1]; % A mátrixok megadása.
>> A\b, B\b % Az egyenletrendszer és a túlhatározott egyenletek megoldása.

ans =

   1.00000000000000
   1.00000000000000

ans =

   1.00000000000000

>> chol(A) % Cholesky-felbontás

ans =

   1.41421356237310  -0.70710678118655
                  0   1.22474487139159

>> [L,U,P]=lu(A) % Az általános LU-felbontás meghatározása.

L =

   1.00000000000000                   0
  -0.50000000000000   1.00000000000000


U =

   2.00000000000000  -1.00000000000000
                  0   1.50000000000000


P =

     1     0
     0     1

>> [Q,R]=qr(B) % A B mátrix QR-felbontása

Q =

  -0.70710678118655  -0.70710678118655
  -0.70710678118655   0.70710678118655


R =
```

```
  -1.41421356237310
                  0
```

Az alábbi program a Gauss-módszert valósítja meg egy olyan $n \times n$-es lineáris egyenletrendszerre, melyre végigfut a módszer. Figyeljük meg, hogy az együtthatómátrixot az eljárás elején kibővítjük a jobb oldali vektorral, így végig az együtthatómátrix elemeiként tudunk minden elemre hivatkozni. Az eljárás végére az együtthatómátrix tartalmazza az LU-felbontás **L** és **U** mátrixait, utolsó sorában pedig a megoldást.

```
function [U,L,x]=gaussmodszer(A,b) % A: együtthatómátrix, b: jobb oldal
n=max(size(A));
A=[A,b];
for k=1:n-1
   for i=k+1:n
      A(i,k)=A(i,k)/A(k,k);
      A(i,k+1:n+1)=A(i,k+1:n+1)-A(i,k)*A(k,k+1:n+1);
   end;
end;

for j = n:-1:2,
   A(j,n+1)=A(j,n+1)/A(j,j);
   A(1:j-1,n+1)=A(1:j-1,n+1)-A(j,n+1)*A(1:j-1,j);
end;
A(1,n+1) = A(1,n+1)/A(1,1);
U=triu(A(:,1:n));
L=A(:,1:n)-U+eye(n);
x=A(:,n+1);
```

A Cholesky-felbontást egy szimmetrikus, pozitív definit mátrixra az alábbi módon lehet előállítani. Ebben a módszerben az eljárás a mátrix alsó háromszög részét írja át a Cholesky-felbontás **G** mátrixává.

```
function [G]=choleskyfelbontas(A) % A: poz. def. szimm. matrix
n=max(size(A));
for k=1:n
    if k>1
    A(k:n,k)=A(k:n,k)-A(k:n,1:k-1)*(A(k,1:k-1))';
    end;
    A(k:n,k)=A(k:n,k)/sqrt(A(k,k));
end;
G=tril(A);
```

3.11. Feladatok

Kondíciószám

3.11.1. feladat. Igazoljuk, hogy tetszőleges reguláris $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix esetén

$$\frac{1}{n}\kappa_2(\mathbf{A}) \le \kappa_1(\mathbf{A}) \le n\kappa_2(\mathbf{A}), \quad \frac{1}{n}\kappa_\infty(\mathbf{A}) \le \kappa_2(\mathbf{A}) \le n\kappa_\infty(\mathbf{A}),$$

$$\frac{1}{n^2}\kappa_1(\mathbf{A}) \le \kappa_\infty(\mathbf{A}) \le n^2\kappa_1(\mathbf{A}).$$

3.11.2. feladat. Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy olyan négyzetes mátrix, melyben a főátló "felett" -1-esek, "alatta" nullák, és a főátlóban 1-esek állnak. Számítsuk ki a mátrix determinánsát és a kondíciószámát maximumnormában!

3.11.3. feladat. Igazoljuk, hogy reguláris $\mathbf{A}$ mátrixra $\kappa_2(\mathbf{A}^T\mathbf{A}) = \kappa_2^2(\mathbf{A}) \ge \kappa_2(\mathbf{A})$!

3.11.4. feladat. Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy nemszinguláris mátrix, és $\mathbf{B} \in \mathbb{R}^{n\times n}$ egy szinguláris mátrix. Igazoljuk, hogy tetszőleges indukált norma esetén $\|\mathbf{A}^{-1}\| \ge 1/\|\mathbf{A}-\mathbf{B}\|$. Ezen képlet segítségével adjunk alsó becslést az

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

mátrix maximumnormabeli kondíciószámára!

3.11.5. feladat. Igazoljuk, hogy ha $\mathbf{A}$ és $\mathbf{B}$ ortogonálisan hasonló reguláris mátrixok, akkor $\|\mathbf{A}\|_2 = \|\mathbf{B}\|_2$ és $\kappa_2(\mathbf{A}) = \kappa_2(\mathbf{B})$!

3.11.6. feladat. Ismert, hogy egy mátrix spektrálsugara becsülhető a mátrix tetszőleges indukált normájával. Igazoljuk ennek segítségével, hogy tetszőleges $\mathbf{A}$ mátrixra $\|\mathbf{A}\|_2^2 \le \|\mathbf{A}\|_1\|\mathbf{A}\|_\infty$, és hogy tetszőleges invertálható $\mathbf{A}$ mátrix esetén

$$\kappa_2(\mathbf{A}) \le \sqrt{\kappa_1(\mathbf{A})\kappa_\infty(\mathbf{A})}\ !$$

Gauss-módszer, LU-felbontás, Cholesky-felbontás

3.11.7. feladat. Vizsgáljuk meg, hogy mekkora a műveletszáma a Gauss–Jordan-eliminációnak egyenletrendszer megoldása és inverz számítása esetén!

3.11.8. feladat. Ha egy felső Hessenberg-mátrixra alkalmazzuk a Gauss-módszert, akkor figyelembe vehetjük, hogy a főátló "alatt" csak a közvetlenül a főátló alatti elemek különböznek nullától. Mekkora lesz az ilyen mátrixok LU-felbontásának műveletszáma? Mit mondhatunk az $\mathbf{L}$ és $\mathbf{U}$ mátrixok szerkezetéről? Ha már a mátrix LU-felbontása elkészült, akkor mennyi műveletbe kerül egy egyenletrendszer megoldása?

3.11.9. feladat. Oldjuk meg az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszert a Gauss-módszer segítségével a lenti adatokkal! Adjuk meg az együtthatómátrix determinánsát is!

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{bmatrix}, \quad \overline{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}, \quad \overline{\mathbf{b}} = \begin{bmatrix} 2 \\ 10 \\ 44 \\ 190 \end{bmatrix}.$$

Adjuk meg az $\mathbf{A}$ mátrix LU-felbontását! Az $\mathbf{A}$ mátrix inverze a MATLAB jelöléseit használva

inv(A)=[4,-13/3,3/2,-1/6;-3,19/4,-2,1/4;4/3,-7/3,7/6,-1/6 ;-1/4,11/24,-1/4,1/24].

Számítsuk ki az $\mathbf{A}$ mátrix maximumnormabeli kondíciószámát!

3.11.10. feladat. Kézi számolás segítségével határozzuk meg a 3×3-as Hilbert-mátrix LU-felbontását!

3.11.11. feladat. Tekintsük azt az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrixot, melyre $a_{ij} = 1$, ha $i = j$ vagy $j = n$, $a_{ij} = -1$, ha $i > j$, különben nulla. Mutassuk meg, hogy $\mathbf{A}$-nak van LU-felbontása, $|l_{ij}| \leq 1$, és $u_{nn} = 2^{n-1}$.

3.11.12. feladat. Tekintsük az alábbi mátrixot

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ -1 & 3 & 0 \\ 0 & -1 & 3 \end{bmatrix}.$$

Diagonalizálható-e ez a mátrix? Válaszunkat indokoljuk! Határozzuk meg a mátrix LU-felbontását!

3.11.13. feladat. Tekintsünk egy olyan lineáris egyenletrendszert, melynek mátrixában csak az első oszlopban, az első sorban ill. a főátlóban vannak nemnulla elemek. Mi történik a mátrixszal a Gauss-módszer alkalmazása során? Adjunk javaslatot a jelenség elkerülésére!

3.11.14. feladat. Legyen

$$\mathbf{A} = \begin{bmatrix} 2 & -3 & 100 \\ 1 & 10 & -0.001 \\ 3 & -100 & 0.01 \end{bmatrix}, \ \overline{\mathbf{b}} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

és tekintsük az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszert! Oldjuk meg az egyenletrendszert részleges főelemkiválasztással, négyjegyű mantisszával számolva!

3.11.15. feladat. Oldjuk meg az egyenletrendszert a Gauss-módszerrel teljes főelemkiválasztással és anélkül, négyjegyű mantisszát használva! Mekkora a két megoldás eltérése maximumnormában?

$$\begin{aligned} 0.003x_1 + 59.14x_2 &= 59.17 \\ 5.291x_1 - 6.13x_2 &= 46.78 \end{aligned}$$

3.11.16. feladat. Tekinsük az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ egyenletrendszert, ahol

$$\mathbf{A} = \begin{bmatrix} 34 & 55 \\ 55 & 89 \end{bmatrix}, \ \overline{\mathbf{b}} = \begin{bmatrix} 21 \\ 34 \end{bmatrix}.$$

Az $\overline{\mathbf{r}} = \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}$ maradékvektort az $\overline{\mathbf{x}} = [-0.11, 0.45]^T$ vektorral kiszámítva $\overline{\mathbf{r}} = [-0.01, 0]^T$, míg az $\overline{\mathbf{x}} = [-0.99, 1.01]^T$ vektorral $\overline{\mathbf{r}} = [-0.89, -1.44]^T$. A megoldás melyik $\overline{\mathbf{x}}$ közelítése pontosabb? Adjunk alsó és felső becslést egy $\overline{\mathbf{x}}$ közelítés megoldástól való eltérésére a maradékvektor segítségével! Ellenőrizzük a becslést az adott egyenletrendszeren!

3.11.17. feladat. Oldjuk meg a $0.00001x + y = 1, x + y = 2$ egyenletrendszert pontosan ill. úgy, hogy csak 4 számjegyű mantisszával dolgozhatunk! Mit tapasztalunk? Segít-e a részleges főelemkiválasztás? Szorozzuk be az első egyenletet 10^5-nel, majd így is oldjuk meg a feladatot! Fogalmazzuk meg tapasztalatainkat!

3.11.18. feladat. Határozzuk meg, hogy lebegőpontos számokat használva mekkora lesz az osztás művelet abszolút hibája ($|(x \boxed{/} y) - (x/y)| =?$)! Milyen tanulsága van az eredménynek?

3.11.19. feladat. Kézi számolással határozzuk meg a 3×3-as Hilbert mátrix Cholesky-felbontását ill. $\mathbf{LDL}^T$ felbontását!

3.11.20. feladat. Határozzuk meg az alábbi $\mathbf{B}$ mátrix LDL^T és Cholesky-felbontását!

$$\mathbf{B} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

3.11.21. feladat. Az alábbi mátrix egy $\mathbf{A} \in \mathbb{R}^{4\times 4}$ szimmetrikus mátrix LU-felbontását tartalmazza úgy, hogy a főátló "alatti" rész az $\mathbf{L}$ mátrix megfelelő főátló alatti részét tartalmazza, a többi elem pedig az $\mathbf{U}$ mátrix megfelelő eleme. Létezik-e az $\mathbf{A}$ mátrixnak Cholesky-felbontása? Ha igen, akkor adjuk meg a $\mathbf{G}$ mátrixot ($\mathbf{A} = \mathbf{G}\mathbf{G}^T$)! Adjuk meg azt az $\overline{\mathbf{x}} \in \mathbb{R}^4$ vektort, melyre $\mathbf{A}\overline{\mathbf{x}} = [1,0,0,0]^T$!

$$\begin{bmatrix} 2 & 3 & 2 & 4 \\ 3/2 & 3/2 & 2 & 3 \\ 1 & 4/3 & 7/3 & 3 \\ 2 & 2 & 9/7 & 1/7 \end{bmatrix}$$

Iterációs egyenletrendszer-megoldás, gradiens-módszerek

3.11.22. feladat. Mely iterációs módszerrel oldható meg a

$$\begin{aligned} 2x - y &= 1 \\ -x + 2y &= 3 \end{aligned}$$

lineáris egyenletrendszer? Mely ω paraméterekkel lesz konvergens a JOR iteráció? Adjuk meg ω értékét úgy, hogy a konvergencia a lehető leggyorsabb legyen! A SOR módszer esetén milyen értékeket vehet fel ω ahhoz, hogy konvergens módszert nyerjünk? Válasszuk az $\omega = 0.2 : 0.2 : 1.8$ értékeket, és vizsgáljuk meg számítógép segítségével a konvergencia sebességét!

3.11.23. feladat. Adjunk becslést arra, hogy az előző feladatban a Jacobi-módszert használva, az $\overline{\mathbf{x}}_0 = [1,1]^T$ kezdővektorról indulva hány iteráció szükséges ahhoz, hogy a megoldást 10^{-6}-nál pontosabban megkapjuk maximumnormában!

3.11.24. feladat. Oldjuk meg a

$$\begin{aligned} -x + 5y - 2z &= 3 \\ x + y - 4z &= -9 \\ 4x - y + 2z &= -8 \end{aligned}$$

egyenletrendszert valamilyen iterációs módszerrel!

3.11.25. feladat. Az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszer együtthatómátrixa szimmetrikus pozitív definit mátrix. Ezután definiáljuk az $\overline{\mathbf{x}}^{(k+1)} = (\mathbf{E} - h\mathbf{A})\overline{\mathbf{x}}^{(k)} + h\overline{\mathbf{b}}$ konzisztens iterációt, ahol h valamilyen valós paraméter. Hogyan válasszuk meg h értékét, hogy az iteráció minden kezdővektorra az egyenletrendszer megoldásához tartson?

3.11.26. feladat. A

$$\mathbf{B} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

mátrixszal szeretnénk megoldani a Jacobi-iterációt használva a $\mathbf{B}\overline{\mathbf{x}} = [1,1]^T$ lineáris egyenletrendszert. Végezzünk el két iterációs lépést a nullvektorról indulva, és becsüljük meg, hogy hány iterációs lépés lenne szükséges ahhoz, hogy a kapott közelítésnek a maximumnorma-beli eltérése a pontos megoldástól 10^{-6}-nál kisebb legyen!

3.11.27. feladat. Legyen $x_0 = 1$ és $x_{20} = 0$ és

$$x_k = \frac{3}{4}x_{k-1} + \frac{1}{4}x_{k+1},\ k = 1, \ldots, 19.$$

Igazoljuk, hogy az egyenletrendszer megoldása $x_k = 1-(3^k-1)/(3^{20}-1)$! Oldjuk meg az egyenletrendszert Gauss–Seidel-módszerrel! Mit tapasztalunk, javítja-e a konvergenciát az alulrelaxálás?

3.11.28. feladat. Az alábbi egyenletrendszert szeretnénk megoldani a Jacobi-módszer relaxálásával.

$$\begin{bmatrix} 3 & 1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Hogyan válasszuk meg ω értékét, hogy a leggyorsabban konvergáljon az eljárás? Számítsuk ki, hogy a nulvektorról indulva a leggyorsabb módszerrel mennyit kellene iterálni, hogy a megoldást 10^{-6}-nál jobban megközelítsük maximumnormában!

3.11.29. feladat. A Jacobi- vagy a Gauss–Seidel-iteráció konvergál gyorsabban az alábbi egyenletrendszerre?

$$\begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix} \overline{\mathbf{x}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Adjunk felső becslést arra, hogy hány iterációs lépést kellene elvégeznünk a gyorsabb módszerrel a $[0,0]^T$ kezdővektorral indulva, hogy a megoldást 10^{-6}-nál jobban megközelítse a sorozat 2-es normában!

3.11.30. feladat. A konjugált gradiens-módszert alkalmazzuk a $\operatorname{tridiag}(-1,2,-1)\overline{\mathbf{x}} = [1,0,1]^T$ egyenletrendszer megoldására. A nullvektort választva kezdővektornak számítsuk ki az $\overline{\mathbf{x}}_2$ vektort, majd számítsuk ki a hozzá tartozó maradékvektort! Mit tapasztalunk?

3.11.31. feladat. Alkalmazzunk két-két iterációs lépést a gradiens- és a konjugált gradiens-módszerrel a

$$4x + 2y = 7$$
$$2x + 3y = 10$$

lineáris egyenletrendszerre!

Householder- és Givens-transzformációk. QR-felbontás. Túlhatározott rendszerek.

3.11.32. feladat. Keressük meg azt a Householder-mátrixot, amellyel a $[2,6,-3]^T$ vektort beszorozva, annak utolsó két eleme nulla lesz! Mi lesz a Householder-mátrix a $[-3,1,-5,1]^T$ vektor esetén?

3.11.33. feladat. Megegyezhet-e egymással egy Householder-tükrözés és egy Givens-forgatás mátrixa?

3.11.34. feladat. Szorozzuk meg a $[2,6,-3]^T$ vektort egy ortogonális mátrixszal úgy, hogy az eredményvektor utolsó komponense 0 legyen!

3.11.35. feladat. Adjuk meg az

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 1 \\ 6 & 2 & 0 \\ -3 & -1 & -1 \end{bmatrix},\ \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 1 & 3 \\ 0 & 2 \end{bmatrix},\ \mathbf{C} = \begin{bmatrix} 3 & 3 & 0 \\ 3 & 5 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

mátrixok Householder QR-felbontásait!

3.11.36. feladat. Képezzük a fenti **A** mátrix QR-felbontását Givens-forgatások segítségével! Hány Givens-forgatásra van szükség?

3.11.37. feladat. Az alábbi táblázat nyolc síkbeli pont koordinátapárjait tartalmazza. Adjuk meg azt a legfeljebb harmadfokú $p(x)$ polinomot, melyre $\sum_{i=1}^{8}(p(x_i)-y_i)^2$ minimális lesz! Ábrázoljuk a pontokat és a polinomot is!

x_i	-4	-2	-1	0	1	3	4	6
y_i	-35.1	15.1	15.9	8.9	0.1	0.1	21.1	135

3.11.38. feladat. Két mennyiséget (x és y) mértünk, ill. ezek különbségét és összegét. Az eredmények: $x=a$, $y=b$, $x-y=c$ és $x+y=d$. Oldjuk meg ezt a túlhatározott egyenletrendszert!

3.11.39. feladat. Oldjuk meg az

$$\begin{bmatrix} 1 & 1 \\ 10^{-k} & 0 \\ 0 & 10^{-k} \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 10^{-k} \\ 1+10^{-k} \\ 1-10^{-k} \end{bmatrix}$$

túlhatározott egyenletrendszert $k=6,7,8$ esetén először papíron számolva, majd az $A\backslash b$ (QR-felbontást használja) és az $(A'*A)\backslash(A'*b)$ utasításokkal (Cholesky-felbontásos megoldás). Hasonlítsuk össze az eredményt!

Ellenőrző kérdések

1. Hogyan értelmezzük a mátrixok kondíciószámát?
2. Hogyan függ egy lineáris egyenletrenszer megoldásának hibája az együtthatómátrix és az egyenlet jobb oldalának hibájától?
3. Ismertessük a Gauss-módszert, és adjuk meg műveletszámát!
4. Milyen mátrixok esetén fut végig a Gauss-módszer elakadás nélkül?
5. Mi az az inga-módszer?
6. Hogyan állítjuk elő egy mátrix LU-felbontását, és mire lehet ezt használni?
7. Miért van szükség főelemkiválasztásra, és hogy hajtjuk ezt végre?
8. Mi az a Cholseky-felbontás?
9. Milyen mátrixok esetén érdemes egy lineáris egyenletrendszert iterációs módszerrel megoldani?
10. Hogy néz ki általánosan egy iterációs módszer?
11. Ismertessük a Jacobi- és Gauss–Seidel-iterációkat!
12. Mitől függ egy iterációs módszer konvergenciájának sebessége?
13. Mit jelent a relaxálás módszere?

14. A SOR módszer szimmetrikus, pozitív definit mátrixokra milyen relaxációs paraméter esetén lesz konvergens?
15. Hogyan alakítható át egy lineáris egyenletrendszer megoldása variációs feladattá?
16. Ismertessük a gradiens-módszert és tulajdonságait!
17. Ismertessük a konjugált gadiens módszert és tulajdonságait!
18. Mi az a Householder-tükrözés és Givens-forgatás?
19. Hogyan állítható elő egy mátrix QR-felbontása?
20. Mit jelent, hogy egy egyenletrendszer túlhatározott? Hogyan lehet egy túlhatározott egyenletrendszert megoldani?

4. Sajátérték-feladatok numerikus megoldása

Ebben a fejezetben megvizsgáljuk, hogyan lehet numerikusan meghatározni a mátrixok sajátértékeit és sajátvektorait. Kétféle módszertípus van. Az egyik típussal az egyszeresen domináns sajátértéket és a hozzá tartozó sajátvektort tudjuk meghatározni, a másik típussal pedig az összes sajátérték közelítését egyszerre határozhatjuk meg.

4.1. Sajátérték-feladatok kondicionáltsága

Sok gyakorlati probléma vezet sajátérték-feladatra, azaz olyan feladatra, amikor egy négyzetes mátrix sajátértékeit és sajátvektorait kell meghatároznunk. Sajátérték-feladat megoldását igényli pl. a szilárdtest fizika, a kvantuummechanika, a gráfelmélet, a differenciálegyenletek, a dinamikai rendszerek, a digitális jel- és képfeldolgozás tudományterületek számos feladata, de pl. a Google is ilyen feladatot old meg a honlapok rangsorolásához [6].

Az 1.2.1. fejezetben foglalkoztunk már a sajátérték-feladattal, de az ott említett karakterisztikus egyenlet segítségével történő sajátérték- és sajátvektor-meghatározási mód gyakorlati problémák esetén nem alkalmazható, hiszen négynél nagyobb fokszámú polinomok megoldására nincs megoldóképletünk. Ez azt jelenti, hogy a sajátértékeket általában direkt módszerrel (azaz olyannal, ami pontosan számolva véges sok lépésen belül pontos sajátértékeket adna) nem lehet meghatározni. Ezért a gyakorlatban mindig iterációs módszereket alkalmazunk. Elegendő a megfelelő sajátértéket vagy a sajátvektort közelíteni, mert az egyik közelítés ismeretében a másik közelítése mindig egyszerűen megadható.

Mielőtt ismertetnénk a sajátérték-feladatok numerikus megoldásának lehetőségeit, vizsgáljuk meg a sajátérték-feladat kondicionáltságát: vajon egy $\mathbf{A}$ négyzetes mátrix elemeit kicsit megváltoztatva, mekkorát változhatnak a sajátértékek? A következő tétel diagonalizálható mátrixokra vizsgálja meg ezt a kérdést.

4.1.1. tétel. (Bauer[1]-Fike, 1960)
Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy diagonalizálható mátrix, azaz $\mathbf{A}$ felírható $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$ alakban, ahol az egyszerűség kedvéért feltesszük, hogy $\mathbf{V}$ oszlopai normálva vannak a $\|.\|_p$ vektornormában, és $\mathbf{D}$ diagonális mátrix. Legyen $\delta\mathbf{A}$ egy tetszőleges mátrix, és legyen μ az $\mathbf{A}+\boldsymbol{\delta}\mathbf{A}$ mátrix egy sajátértéke. Ekkor

$$\min_{\lambda\ \mathbf{A}\ \text{sajátértéke}} |\lambda - \mu| \leq \kappa_p(\mathbf{V})\|\boldsymbol{\delta}\mathbf{A}\|_p.$$

Bizonyítás. Ha μ sajátértéke $\mathbf{A}$-nak is, akkor triviális az állítás. Tegyük fel tehát, hogy nem sajátértéke. Mivel $\mathbf{A}+\boldsymbol{\delta}\mathbf{A}-\mu\mathbf{E}$ szinguláris, ezért

$$\mathbf{V}^{-1}(\mathbf{A}+\boldsymbol{\delta}\mathbf{A}-\mu\mathbf{E})\mathbf{V} = \mathbf{D}+\mathbf{V}^{-1}\boldsymbol{\delta}\mathbf{A}\mathbf{V}-\mu\mathbf{E}$$

[1] Friedrich Ludwig Bauer (1923-), német matematikus.

is szinguláris, így van olyan $\overline{\mathbf{x}} \neq \mathbf{0}$ vektor, mellyel

$$(\mathbf{D} - \mu\mathbf{E} + \mathbf{V}^{-1}\boldsymbol{\delta}\mathbf{A}\mathbf{V})\overline{\mathbf{x}} = \mathbf{0}.$$

A $\mathbf{D} - \mu\mathbf{E}$ mátrix inverzével balról szorozva azt kapjuk, hogy

$$(\mathbf{E} + (\mathbf{D} - \mu\mathbf{E})^{-1}\mathbf{V}^{-1}\boldsymbol{\delta}\mathbf{A}\mathbf{V})\overline{\mathbf{x}} = \mathbf{0},$$

azaz

$$\overline{\mathbf{x}} = -(\mathbf{D} - \mu\mathbf{E})^{-1}\mathbf{V}^{-1}\boldsymbol{\delta}\mathbf{A}\mathbf{V}\overline{\mathbf{x}}.$$

Így

$$\|\overline{\mathbf{x}}\|_p \leq \|(\mathbf{D} - \mu\mathbf{E})^{-1}\|_p \|\mathbf{V}^{-1}\|_p \|\boldsymbol{\delta}\mathbf{A}\|_p \|\mathbf{V}\|_p \|\overline{\mathbf{x}}\|_p$$

és

$$\begin{aligned} 1 &\leq \|(\mathbf{D} - \mu\mathbf{E})^{-1}\|_p \kappa_p(\mathbf{V}) \|\boldsymbol{\delta}\mathbf{A}\|_p \\ &= \max_i \frac{1}{|\lambda_i - \mu|} \kappa_p(\mathbf{V}) \|\boldsymbol{\delta}\mathbf{A}\|_p = \frac{1}{\min_i |\lambda_i - \mu|} \kappa_p(\mathbf{V}) \|\boldsymbol{\delta}\mathbf{A}\|_p. \end{aligned}$$

Ebből már következik az állítás. ■

Érdemes megjegyezni, hogy a sajátérték-feladat kondicionáltságát nem az eredeti mátrix, hanem a diagonalizáló mátrix (a sajátvektorok mátrixa) kondíciószáma határozza meg. Ez speciálisan azt jelenti, hogy mivel a szimmetrikus mátrixok ortogonális mátrixokkal diagonalizálhatók, és az ortogonális mátrixok 2-es normája 1, szimmetrikus mátrixokra a

$$\min_{\lambda \text{ s.é.-e } \mathbf{A}\text{-nak}} |\lambda - \mu| \leq \|\boldsymbol{\delta}\mathbf{A}\|_2$$

becslés érvényes. Érdekes példaként említhetjük a $\mathbf{H}_n$ Hilbert-mátrixot, amely egyenletrendszer megoldásakor rosszul, viszont sajátérték-feladatok esetén jól kondicionált, hiszen a sajátértékek maximum akkorát változhatnak, mint a perturbáló mátrix 2-es normája.

4.2. A sajátértékeket egyenként közelítő eljárások

A numerikus matematikában a sajátérték-meghatározási módszereket két nagy csoportra szokás osztani. A sajátértékeket egyenként közelítő eljárásokra ill. a sajátértékeket egyszerre közelítőkre. Az első típusba tartozó módszerek mindig csak egy-egy megfelelő sajátértéket közelítenek, míg a második típusba tartozók olyan eljárást adnak, mellyel egyszerre az összes sajátértékre kapunk közelítést. Ebben a fejezetben a sajátértékeket egyenként közelítő módszereket tárgyaljuk.

A sajátértékeket egyenként közelítő eljárások egy olyan vektorsorozatot állítanak elő, amely egy meghatározott sajátvektorhoz tart. A sajátvektort ekkor ezen sorozat egy határértékhez elegendően közeli elemével közelíthetjük. Mielőtt rátérünk arra, hogy hogyan lehet egy adott sajátvektorhoz tartozó sorozatot előállítani, nézzük meg, hogy hogyan mondhatunk megfelelő becslést a sajátértékre, ha ismerjük a sajátvektor egy becslését? Az első gondolatunk az lehet, hogy ha van egy $\hat{\mathbf{v}}$ közelítésünk a sajátvektorra, akkor az $\mathbf{A}\hat{\mathbf{v}} \approx \lambda\hat{\mathbf{v}}$ becslés soraiból kaphatunk közelítést a sajátértékre. Vizsgáljuk meg ezt a módszert egy példa segítségével!

4.2.1. példa. Legyen $\mathbf{A}$ a 3×3-as Hilbert-mátrix, és tegyük fel, hogy $\hat{\mathbf{v}}^T = [0.82694986, 0.45995562, 0.32341112]$ egyik sajátvektorának közelítését adja. Adjunk becslést az adott sajátvektorhoz tartozó sajátértékre. Mivel

$$\begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix} \begin{bmatrix} 0.82694986 \\ 0.45995562 \\ 0.32341112 \end{bmatrix} = \begin{bmatrix} 1.16473138 \\ 0.64764625 \\ 0.45532108 \end{bmatrix} \approx \lambda \begin{bmatrix} 0.82694986 \\ 0.45995562 \\ 0.32341112 \end{bmatrix},$$

a fenti közelítés mindhárom sorából kaphatunk közelítést a sajátértékre. Ezek rendre $\lambda_1 = 1.40846675$, $\lambda_2 = 1.40806247$, $\lambda_3 = 1.40787084$, vagy vehetjük ezek átlagát, ami $\lambda_4 = 1.40813335$ értéket ad. Ez utóbbi érték $1.85576699.10^{-4}$-nel tér csak el a pontos sajátértéktől ($\lambda = 1.40831893$). $\diamond$

Adhatunk-e az előző példában nyert értéknél jobb közelítést a sajátértékre? Pl. az $\mathbf{A}\hat{\mathbf{v}} \approx \lambda\hat{\mathbf{v}}$ egyenlőséget balról $\hat{\mathbf{v}}^T$-tal szorozva, majd λ-t kifejezve kapjuk, hogy $\lambda \approx \hat{\mathbf{v}}^T\mathbf{A}\hat{\mathbf{v}}/(\hat{\mathbf{v}}^T\hat{\mathbf{v}})$. Vizsgáljuk meg ezt a közelítést az előző példán.

4.2.2. példa. A fenti példában adott sajátvektor-közelítéssel kiszámítva a $\lambda \approx \hat{\mathbf{v}}^T\mathbf{A}\hat{\mathbf{v}}/(\hat{\mathbf{v}}^T\hat{\mathbf{v}})$ hányadost $\lambda = 1.40831889$ adódik. Ennek a pontos sajátértéktől való eltérése csak $3.87698300.10^{-8}$, amely sokkal kisebb, mint az előző példában nyert közelítő érték. $\diamond$

Látható tehát, hogy ez a második közelítés sokkal pontosabb, mint az első. Ennek okát az alábbi tétel világítja meg.

4.2.3. tétel.
Legyen adva az $\mathbf{0} \neq \overline{\mathbf{x}} \in \mathbb{R}^n$ vektor és az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix. Ekkor

$$\min_{\alpha\in\mathbb{R}} \|\mathbf{A}\overline{\mathbf{x}} - \alpha\overline{\mathbf{x}}\|_2^2 = \|\mathbf{A}\overline{\mathbf{x}} - R(\overline{\mathbf{x}})\overline{\mathbf{x}}\|_2^2,$$

ahol

$$R(\overline{\mathbf{x}}) = \frac{\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}}}{\overline{\mathbf{x}}^T\overline{\mathbf{x}}}.$$

Bizonyítás. A bal oldalon egy α-tól függő egyváltozós függvény áll. Számítsuk ki ennek értékét!

$$\begin{aligned}\|\mathbf{A}\overline{\mathbf{x}} - \alpha\overline{\mathbf{x}}\|_2^2 &= (\overline{\mathbf{x}}^T\mathbf{A}^T - \alpha\overline{\mathbf{x}}^T)(\mathbf{A}\overline{\mathbf{x}} - \alpha\overline{\mathbf{x}}) = \overline{\mathbf{x}}^T\mathbf{A}^T\mathbf{A}\overline{\mathbf{x}} - 2\alpha\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}} + \alpha^2\overline{\mathbf{x}}^T\overline{\mathbf{x}} \\ &= \alpha^2\overline{\mathbf{x}}^T\overline{\mathbf{x}} - 2\alpha\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}} + \overline{\mathbf{x}}^T\mathbf{A}^T\mathbf{A}\overline{\mathbf{x}}.\end{aligned}$$

Mivel $\overline{\mathbf{x}}^T\overline{\mathbf{x}} > 0$, ha $\overline{\mathbf{x}} \neq \mathbf{0}$, ezért a függvény a minimumát az

$$\alpha_{\min} = \frac{\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}}}{\overline{\mathbf{x}}^T\overline{\mathbf{x}}} = R(\overline{\mathbf{x}})$$

esetben veszi fel. ■

4.2.4. definíció.
Legyen $\mathbf{0} \neq \overline{\mathbf{x}} \in \mathbb{R}^n$ és $\mathbf{A} \in \mathbb{R}^{n\times n}$. Ekkor az

$$R(\overline{\mathbf{x}}) = \frac{\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}}}{\overline{\mathbf{x}}^T\overline{\mathbf{x}}}$$

számot az $\overline{\mathbf{x}}$ vektorhoz tartozó *Rayleigh-hányadosnak* hívjuk.

A tétel szerint tehát 2-es normában egy vektor Rayleigh-hányados-szorosa lesz legközelebb $\overline{\mathbf{x}}$ számszorosai közül az $\mathbf{A}\overline{\mathbf{x}}$ vektorhoz.

4.2.5. megjegyzés. Szimmetrikus mátrixok esetén a Rayleigh-hányados mindig a legkisebb és a legnagyobb sajátérték között helyezkedik el, azaz

$$\lambda_{\min} \leq R(\overline{\mathbf{x}}) \leq \lambda_{\max},$$

továbbá

$$\lambda_{\max} = \max_{\mathbf{0}\neq\overline{\mathbf{x}}\in\mathbb{R}^n} R(\overline{\mathbf{x}}), \quad \lambda_{\min} = \min_{\mathbf{0}\neq\overline{\mathbf{x}}\in\mathbb{R}^n} R(\overline{\mathbf{x}}).$$

Ez utóbbi állítás az ún. Courant–Fischer-tétel. ◇

Az előzőek alapján tehát, ha van egy közelítésünk a sajátvektorra, akkor a Rayleigh-hányados segítségével kaphatunk jó becslést a sajátvektorhoz tartozó sajátértékre. Ha $\overline{\mathbf{x}}$ normája 1, akkor a Rayleigh-hányados az $R(\overline{\mathbf{x}}) = \overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}}$ alakra egyszerűsödik.

4.2.1. A hatványmódszer

A hatványmódszer alapötlete a következő. Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy adott mátrix, és tegyük fel, hogy a mátrixnak van egy egyszeresen domináns λ_1 sajátértéke, azaz egy olyan λ_1 sajátérték, mellyel a többi $(\lambda_2, \ldots, \lambda_n)$ sajátértékre igaz a

$$|\lambda_1| > |\lambda_2| \geq \ldots |\lambda_n|$$

egyenlőtlenség. Ebben az esetben nyilván $\lambda_1 \in \mathbb{R}$, továbbá a hozzá tartozó $\overline{\mathbf{v}}_1$ sajátvektor valósnak választható. Tegyük fel, hogy az $\mathbf{A}$ mátrix normális. Ekkor a mátrixnak van ortonormált sajátvektorrendszere. Legyen ez $\overline{\mathbf{v}}_1, \ldots, \overline{\mathbf{v}}_n$. Legyen $\overline{\mathbf{x}} \in \mathbb{R}^n$ olyan, hogy az $\overline{\mathbf{x}} = \alpha_1\overline{\mathbf{v}}_1 + \alpha_2\overline{\mathbf{v}}_2 + \cdots + \alpha_n\overline{\mathbf{v}}_n$ előállításban $\alpha_1 = \overline{\mathbf{v}}_1^T\overline{\mathbf{x}} \neq 0$ $(\alpha_1 \in \mathbb{R})$. Ekkor

$$\begin{aligned}\mathbf{A}^k\overline{\mathbf{x}} &= \alpha_1\lambda_1^k\overline{\mathbf{v}}_1 + \alpha_2\lambda_2^k\overline{\mathbf{v}}_2 + \cdots + \alpha_n\lambda_n^k\overline{\mathbf{v}}_n \\ &= \lambda_1^k\left(\alpha_1\overline{\mathbf{v}}_1 + \underbrace{\alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\overline{\mathbf{v}}_2}_{\to 0} + \cdots + \underbrace{\alpha_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\overline{\mathbf{v}}_n}_{\to 0}\right).\end{aligned}$$

A jelölt tagok k növelésével nullához tartanak, ami azt jelenti, hogy az $\mathbf{A}^k\overline{\mathbf{x}}$ vektor k növelésével egyre inkább a $\overline{\mathbf{v}}_1$ sajátvektor által meghatározott sajátirányba fog mutatni. Ha a domináns sajátérték nem ± 1, akkor az így nyert vektorok hossza vagy nullához, vagy végtelenhez fog tartani, így alul- vagy felülcsordulás léphet fel amikor számítógépen hajtjuk végre az iterációt. Emiatt az egyes lépések után a keletkező vektort célszerű normálni. Így jutunk el a hatványmódszerhez, melynek algoritmusa a következő.

Hatványmódszer, $\mathbf{A}$ normális, $\overline{\mathbf{y}}^{(0)}$ olyan kezdővektor, melyre $\overline{\mathbf{v}}_1^T\overline{\mathbf{y}}^{(0)} \neq 0$, $\|\overline{\mathbf{y}}^{(0)}\|_2 = 1$

for $k := 1 : k_{\max}$ **do**
  $\overline{\mathbf{x}}^{(k)} := \mathbf{A}\overline{\mathbf{y}}^{(k-1)}$
  $\overline{\mathbf{y}}^{(k)} := \overline{\mathbf{x}}^{(k)}/\|\overline{\mathbf{x}}^{(k)}\|_2$
  $\nu^{(k)} := (\overline{\mathbf{y}}^{(k)})^T\mathbf{A}\overline{\mathbf{y}}^{(k)}$
end for

A $\nu^{(k)}$ értékek a Rayleigh-hányadossal számított sajátérték közelítések. Ezen értékeket, azon túl, hogy közelítik a $\overline{\mathbf{v}}_1$ vektorhoz tartozó sajátértéket, felhasználhatjuk a leállási feltétel megadására is. Pl. ha értéke már egy adott toleranciaszintnél kevesebbet változik, akkor leállíthatjuk az

iterációt. A hatványmódszer (angolul power method) nyilvánvalóan onnét kapta a nevét, hogy az $\overline{\mathbf{y}}^{(0)}$ vektort az $\mathbf{A}$ mátrix egyre nagyobb hatványaival szorozzuk meg az eljárás során. A következő tétel azt mutatja, hogy a fenti algoritmus valóban a várt eredményt szolgáltatja.

4.2.6. tétel.
A fenti algoritmusban

$$\overline{\mathbf{y}}^{(k)} = \frac{\mathbf{A}^k \overline{\mathbf{y}}^{(0)}}{\|\mathbf{A}^k \overline{\mathbf{y}}^{(0)}\|_2},$$

$\nu^{(k)} \to \lambda_1$ $(k \to \infty)$, továbbá létezik olyan $\{\gamma_k\} \subset \mathbb{R}$ sorozat, hogy $|\gamma_k| = 1$ $(k = 1, \dots)$ és

$$\gamma_k \overline{\mathbf{y}}^{(k)} \to \overline{\mathbf{v}}_1.$$

Bizonyítás. A tétel első állítása teljes indukcióval egyszerűen igazolható.
A harmadik állításhoz a Parseval-egyenlőséget fogjuk használni, mely szerint ha

$$\overline{\mathbf{x}} = \alpha_1 \overline{\mathbf{v}}_1 + \ldots + \alpha_n \overline{\mathbf{v}}_n,$$

ahol $\overline{\mathbf{v}}_1, \dots, \overline{\mathbf{v}}_n$ 2-es normában ortonormált vektorrendszer, akkor $\|\overline{\mathbf{x}}\|_2 = \sqrt{\sum_{i=1}^n |\alpha_i|^2}$. Ez az egyenlőség az alábbi módon igazolható.

$$\overline{\mathbf{x}}^H \overline{\mathbf{x}} = \left(\sum_{i=1}^n \overline{\alpha}_i \overline{\mathbf{v}}_i^H\right)\left(\sum_{i=1}^n \alpha_i \overline{\mathbf{v}}_i\right) = \sum_{i=1}^n |\alpha_i|^2.$$

Legyen tehát $\overline{\mathbf{y}}^{(0)} = \alpha_1 \overline{\mathbf{v}}_1 + \alpha_2 \overline{\mathbf{v}}_2 + \cdots + \alpha_n \overline{\mathbf{v}}_n$, és tegyük fel a tétel feltételeinek megfelelően, hogy $\alpha_1 \neq 0$. Így

$$\begin{aligned}\overline{\mathbf{y}}^{(k)} &= \frac{\lambda_1^k \left(\alpha_1 \overline{\mathbf{v}}_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k \overline{\mathbf{v}}_2 + \cdots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^k \overline{\mathbf{v}}_n\right)}{\sqrt{\sum_{i=1}^n |\alpha_i|^2 |\lambda_i|^{2k}}} \\ &= \frac{\lambda_1^k \alpha_1 \left(\overline{\mathbf{v}}_1 + \frac{\alpha_2}{\alpha_1} \left(\frac{\lambda_2}{\lambda_1}\right)^k \overline{\mathbf{v}}_2 + \cdots + \frac{\alpha_n}{\alpha_1} \left(\frac{\lambda_n}{\lambda_1}\right)^k \overline{\mathbf{v}}_n\right)}{|\lambda_1|^k |\alpha_1| \sqrt{1 + \sum_{i=2}^n |\frac{\alpha_i}{\alpha_1}|^2 |\frac{\lambda_i}{\lambda_1}|^{2k}}}.\end{aligned}$$

Tehát

$$\overbrace{\frac{|\lambda_1|^k |\alpha_1|}{\lambda_1^k \alpha_1}}^{=:\gamma_k} \overline{\mathbf{y}}^{(k)} = \frac{\left(\overline{\mathbf{v}}_1 + \frac{\alpha_2}{\alpha_1} \left(\frac{\lambda_2}{\lambda_1}\right)^k \overline{\mathbf{v}}_2 + \cdots + \frac{\alpha_n}{\alpha_1} \left(\frac{\lambda_n}{\lambda_1}\right)^k \overline{\mathbf{v}}_n\right)}{\sqrt{1 + \sum_{i=2}^n |\frac{\alpha_i}{\alpha_1}|^2 |\frac{\lambda_i}{\lambda_1}|^{2k}}} \to \overline{\mathbf{v}}_1,$$

ahol $|\gamma_k| = 1$ $(k = 1, \dots)$.

A második állítás igazolásához pedig induljunk ki a

$$(\gamma_k \overline{\mathbf{y}}^{(k)})^T \mathbf{A} (\gamma_k \overline{\mathbf{y}}^{(k)}) - \overline{\mathbf{v}}_1^T \mathbf{A} \overline{\mathbf{v}}_1 \to 0$$

határértékből, mellyel $k \to \infty$ esetén

$$\nu^{(k)} - \lambda_1 = (\overline{\mathbf{y}}^{(k)})^T \mathbf{A} \overline{\mathbf{y}}^{(k)} - \lambda_1 = |\gamma_k|^2 (\overline{\mathbf{y}}^{(k)})^T \mathbf{A} \overline{\mathbf{y}}^{(k)} - \lambda_1 = (\gamma_k \overline{\mathbf{y}}^{(k)})^T \mathbf{A} (\gamma_k \overline{\mathbf{y}}^{(k)}) - \overline{\mathbf{v}}_1^T \mathbf{A} \overline{\mathbf{v}}_1 \to 0.$$

Ezzel a tétel állításait igazoltuk. ■

4.2.7. megjegyzés. A tétel bizonyításából az is látható, hogy a generált vektorsorozatból hogyan lehet előállítani a sajátvektorhoz tartó vektorsorozatot.

- Ha $\lambda_1, \alpha_1 > 0$, akkor $\overline{\mathbf{y}}^{(k)} \to \overline{\mathbf{v}}_1$.
- Ha $\lambda_1 > 0, \alpha_1 < 0$, akkor $-\overline{\mathbf{y}}^{(k)} \to \overline{\mathbf{v}}_1$.
- Ha $\lambda_1 < 0, \alpha_1 > 0$, akkor $(-1)^k \overline{\mathbf{y}}^{(k)} \to \overline{\mathbf{v}}_1$.
- Ha $\lambda_1 < 0, \alpha_1 < 0$, akkor $(-1)^{k+1} \overline{\mathbf{y}}^{(k)} \to \overline{\mathbf{v}}_1$.

$\diamond$

4.2.8. megjegyzés. Legyen $\overline{\mathbf{e}}^{(k)} = \overline{\mathbf{y}}^{(k)} - \overline{\mathbf{v}}_1$ a sajátvektor k-adik közelítésének hibája. Ekkor elegendően nagy k értékekre $\|\overline{\mathbf{e}}^{(k+1)}\|_2 \approx |\lambda_2/\lambda_1| \|\overline{\mathbf{e}}^{(k)}\|_2$, ami a lineáris konvergencián kívül azt is mutatja, hogy általában a $|\lambda_2/\lambda_1|$ hányados határozza meg a konvergencia sebességét. $\diamond$

A hatványmódszer konvergenciáját az egyszerűség kedvéért csak normális mátrixok esetén igazoltuk, de az eljárás alkalmazható pl. diagonalizálható mátrixok esetén is. Természetesen az a feltétel, hogy λ_1 egyszeresen domináns sajátérték legyen, nehezen ellenőrizhető előre, hiszen nem ismerjük a mátrix sajátértékeit (éppen ezek meghatározása a feladat). Mindenesetre az $\mathbf{A}$ mátrixnak lehetnek képzetes résszel rendelkező sajátértékei is, melyek konjugáltja is sajátérték lesz, így nem sok esély van arra, hogy egyszeresen domináns sajátértéke legyen a mátrixnak. Ha viszont a mátrix szimmetrikus, akkor annak esélye, hogy a domináns sajátérték többszörös lesz, kicsi. Így a továbbiakban mindig feltesszük, hogy az a mátrix, melynek a sajátértékeit és sajátvektorait keressük, szimmetrikus.

4.2.9. példa. Határozzuk meg a $\mathrm{tridiag}(-1, 2, -1) \in \mathbb{R}^{20\times 20}$ mátrix domináns sajátértékét és a hozzá tartozó sajátvektort a hatványmódszer segítségével! Mivel a mátrix szimmetrikus, így hacsak nem lesz két egyforma legnagyobb abszolút értékű sajátérték, akkor a hatványmódszer valóban megtalálja a domináns sajátértéket és a hozzá tartozó sajátvektort. Ha a `powmeth.m` programot alkalmazzuk a megoldáshoz az $\overline{\mathbf{y}}^{(0)} = [1/\sqrt{20}, \ldots, 1/\sqrt{20}]^T$ kezdővektorral és 10^{-6} toleranciaszinttel, akkor a 83. lépés után áll le az iteráció. Ekkor a sajátértékbecslés 3.97765118. $\diamond$

4.2.10. megjegyzés. A hatványmódszer alkalmazásához eddig feltettük, hogy az $\overline{\mathbf{y}}^{(0)}$ kezdővektor olyan, hogy $\overline{\mathbf{v}}_1^T \overline{\mathbf{y}}^{(0)} \neq 0$, azaz hogy a kezdővektornak van az első sajátvektor irányába eső komponense. Hogyan biztosítható ez, ha nem ismerjük a sajátvektorokat? Szerencsére a gyakorlatban a $\overline{\mathbf{v}}_1^T \overline{\mathbf{y}}^{(0)} \neq 0$ feltétel nem játszik komoly szerepet, hiszen ha az első lépésben nem is teljesül a feltétel, a kerekítési hibák miatt az iteráció során lesz az iterációs vektornak $\overline{\mathbf{v}}_1$ irányú komponense, és azzal már elindul az iteráció. $\diamond$

4.2.2. Inverz iteráció

Az előző fejezetben láttuk, hogy hogyan határozható meg a hatványmódszerrel az egyszeresen domináns sajátérték és a hozzá tartozó sajátvektor. Ebben a fejezetben azt vizsgáljuk meg, hogy más sajátértékek hogyan határozhatók meg a hatványmódszer kis átalakításával. Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy nemszinguláris szimmetrikus mátrix λ_i sajátértékekkel és $\overline{\mathbf{v}}_i$ ortonormált sajátvektorokkal $(i = 1, \ldots, n)$. Ha $\mu \neq \lambda_i$ $(i = 1, \ldots, n)$, akkor az $\mathbf{A} - \mu\mathbf{E}$ mátrix invertálható, és $(\mathbf{A} - \mu\mathbf{E})^{-1}$ sajátvektorai megegyeznek $\mathbf{A}$ sajátvektoraival, sajátértékei pedig $(\lambda_i - \mu)^{-1}$. Ha μ elegendően közel van valamelyik λ_j sajátértékhez, akkor a domináns sajátérték $(\lambda_j - \mu)^{-1}$ lesz, és az $(\mathbf{A} - \mu\mathbf{E})^{-1}$

mátrixszal végrehajtva a hatványmódszert, λ_j és $\overline{\mathbf{v}}_j$ meghatározható. Az algoritmus tehát a következő.

Inverz iteráció, $\mathbf{A}$ szimmetrikus, $\|\overline{\mathbf{y}}^{(0)}\|_2^2 = 1$
for $k := 1 : k_{\max}$ **do**
  $(\mathbf{A} - \mu\mathbf{E})\overline{\mathbf{x}}^{(k)} = \overline{\mathbf{y}}^{(k-1)}$ megoldása $\overline{\mathbf{x}}^{(k)}$-ra
  $\overline{\mathbf{y}}^{(k)} := \overline{\mathbf{x}}^{(k)} / \|\overline{\mathbf{x}}^{(k)}\|_2$
  $\nu^{(k)} := (\overline{\mathbf{y}}^{(k)})^T \mathbf{A} \overline{\mathbf{y}}^{(k)}$
end for

Az eljárást nyilvánvalóan azért hívjuk inverz iterációnak, mert az $\mathbf{A}$ mátrix helyett az $\mathbf{A} - \mu\mathbf{E}$ mátrix inverzével hajtjuk végre a hatványmódszert. Ha $\mu = 0$, akkor a nullához legközelebbi, azaz a legkisebb abszolút értékű sajátértéket találja meg a módszer.

Természetesen az inverz iteráció sokkal nagyobb műveletigényű, mint a hatványmódszer, hiszen az előbbinél minden iterációs lépésben meg kell oldanunk egy-egy lineáris egyenletrendszert. A műveletszám csökkentésére jól alkalmazható az LU-felbontás. Mivel minden iterációs lépés egyenletrendszerében az $\mathbf{A} - \mu\mathbf{E}$ mátrix az együtthatómátrix, így ennek LU-felbontását az első lépésben $2n^3/3 + \mathcal{O}(n^2)$ flop művelettel kiszámítva a többi lépésben már csak $2n^2$ flop műveletre van szükség.

Fontos észrevétel, hogy míg a Rayleigh-hányados segítségével sajátvektor közelítéséből sajátértéket tudtunk közelíteni, addig, ha adott egy becslés egy sajátértékre, akkor a hozzá tartozó sajátvektort az inverz iteráció segítségével tudjuk meghatározni (természetesen ekkor a sajátértékbecslés is pontosodik).

4.2.3. Rayleigh-hányados iteráció

Az inverz iteráció eljárását módosíthatjuk az alábbi módon. Ha végrehajtjuk az inverz iteráció egy lépését egy sajátérték-közelítésből kiindulva, akkor kapunk egy becslést a sajátvektorra, melyből Rayleigh-hányados segítségével mondhatunk egy újabb sajátértékbecslést. Az inverz iteráció következő lépésében már az új sajátértékbecslést használhatjuk. Ezt az eljárást Rayleigh-hányados iterációnak nevezzük. Az algoritmusa a következő.

Rayleigh-hányados iteráció, $\mathbf{A}$ szimm., $\|\overline{\mathbf{y}}^{(0)}\|_2^2 = 1$
for $k := 1 : k_{\max}$ **do**
  $R(\overline{\mathbf{y}}^{(k-1)})$ kiszámítása
  $(\mathbf{A} - R(\overline{\mathbf{y}}^{(k-1)})\mathbf{E})\overline{\mathbf{x}}^{(k)} = \overline{\mathbf{y}}^{(k-1)}$ megoldása $\overline{\mathbf{x}}^{(k)}$-ra
  $\overline{\mathbf{y}}^{(k)} := \overline{\mathbf{x}}^{(k)} / \|\overline{\mathbf{x}}^{(k)}\|_2$
end for

Az algoritmus során minden lépésben egy új lineáris egyenletrendszert kell megoldanunk. Cserébe gyorsabb konvergenciát nyerünk, nevezetesen a konvergencia harmadrendű lesz. Érdemes megjegyezni, hogy az iterációban attól függ, hogy melyik sajátértéket és sajátvektort találja meg a módszer, hogy milyen kezdővektorról indítjuk azt. A módszer általában a kezdővektor által meghatározott Rayleigh-hányadoshoz legközelebbi sajátértékhez és a hozzá tartozó sajátvektorhoz konvergál. Ha egy adott μ értékhez legközelebbi sajátértékre van szükségünk, akkor érdemes először az inverz iterációt alkalmazni, majd amikor már elég közel kerültünk a keresett sajátértékhez, akkor áttérhetünk a Rayleigh-hányados iterációra.

4.2.11. megjegyzés. A Rayleigh-hányados iterációban minél közelebb vagyunk a keresett sajátértékhez, annál közelebb van az $\mathbf{A} - R(\overline{\mathbf{y}}^{(k-1)})\mathbf{E}$ mátrix egy szinguláris mátrixhoz. Ha tehát az iterációs mátrixunk szingulárissá válik a számítógépes eljárás során, az azt jelzi, hogy nagyon közel vagyunk a keresett sajátértéket. Ez használható tehát leállási feltételként. ⋄

4.2.4. Deflációs eljárások

Tegyük fel, hogy egy $\mathbf{A} \in \mathbb{R}^{n\times n}$ valós szimmetrikus mátrixnak minden sajátértéke abszolút értékben különböző: $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$. Ha már meghatároztuk a mátrix szigorúan domináns λ_1 sajátértékét és a hozzá tartozó $\overline{\mathbf{v}}_1$ sajátvektort (pl. a hatványmódszer segítségével), akkor néhány egyszerű eljárással előállíthatunk egy olyan mátrixot, melynek λ_2 lesz a domináns sajátértéke, így azt a hatványmódszerrel meg tudjuk határozni most már az új mátrixra alkalmazva azt. Ez az eljárás addig folytatható, amíg a kellő számú sajátértéket meg nem határoztuk. Ezt az eljárást *deflációnak*[2] nevezzük. Háromfajta deflációs eljárást mutatunk most be.

Householder-defláció

Tegyük fel, hogy már meghatároztuk az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix szigorúan domináns λ_1 sajátértékét és a hozzá tartozó euklideszi normában normált $\overline{\mathbf{v}}_1$ sajátvektort. Határozzuk meg ezután azt a $\mathbf{H}$ Householder-tükrözési mátrixot, mellyel $\mathbf{H}\overline{\mathbf{v}}_1 = \overline{\mathbf{e}}_1$. Ekkor

$$\mathbf{HAH}\overline{\mathbf{e}}_1 = \mathbf{HAHH}\overline{\mathbf{v}}_1 = \mathbf{HA}\overline{\mathbf{v}}_1 = \mathbf{H}\lambda_1\overline{\mathbf{v}}_1 = \lambda_1\mathbf{H}\overline{\mathbf{v}}_1 = \lambda_1\overline{\mathbf{e}}_1.$$

A $\mathbf{HAH}$ mátrix tehát a

$$\mathbf{HAH} = \begin{bmatrix} \lambda_1 & \overline{\mathbf{b}}^T \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}$$

alakot ölti. Mivel hasonlósági transzformációt hajtottunk végre, így $\mathbf{A}_2$ sajátértékei λ_1 kivételével megegyeznek az $\mathbf{A}$ mátrix sajátértékeivel. Így tehát ha $|\lambda_2| > |\lambda_3|$, akkor az $\mathbf{A}_2$ mátrixszal végrehajtva a hatványmódszert, megkaphatjuk λ_2 közelítő értékét. Legyen ez $\tilde{\lambda}_2$.

A sajátvektor meghatározását az $\mathbf{A} - \tilde{\lambda}_2\mathbf{E}$ mátrixra vonatkozó inverz iterációval végezhetjük el. Ekkor közelítőleg megkapjuk a $\overline{\mathbf{v}}_2$ sajátvektort, és a λ_2-re adott becslés is pontosodik.

Ha most meg tudnánk mondani az $\mathbf{A}_2$ mátrix λ_2-höz tartozó sajátvektorát is, akkor hasonlóan folytathatnánk a többi sajátpár meghatározását, mint ahogy λ_1 és $\overline{\mathbf{v}}_1$ segítségével meghatároztuk a $\lambda_2, \overline{\mathbf{v}}_2$ sajátpárt. Az $\mathbf{A}_2$ mátrix λ_2-höz tartozó sajátvektora könnyen meghatározható. Mivel

$$\mathbf{HAH}(\mathbf{H}\overline{\mathbf{v}}_2) = \mathbf{HA}\overline{\mathbf{v}}_2 = \lambda_2(\mathbf{H}\overline{\mathbf{v}}_2),$$

így $\mathbf{H}\overline{\mathbf{v}}_2$ sajátvektora a $\mathbf{HAH}$ mátrixnak λ_2 sajátértékkel. Azaz $(\mathbf{H}\overline{\mathbf{v}}_2)(2:n)$ sajátvektora $\mathbf{A}_2$-nek, szintén λ_2 (szigorúan domináns) sajátértékkel.

Rangdefláció

Tegyük fel ismét, hogy meghatároztuk az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix szigorúan domináns λ_1 sajátértékét és a hozzá tartozó normált (euklideszi normában) $\overline{\mathbf{v}}_1$ sajátvektort. Tekintsük az $\mathbf{A} - \lambda_1\overline{\mathbf{v}}_1\overline{\mathbf{v}}_1^T$ mátrixot. Ennek sajátértékei megegyeznek $\mathbf{A}$ sajátértékeivel azzal a különbséggel, hogy λ_1 helyett nulla szerepel. A sajátvektorok ugyanazok. Így ha λ_2 szigorúan domináns a maradék sajátértékek között, akkor az $\mathbf{A} - \lambda_1\overline{\mathbf{v}}_1\overline{\mathbf{v}}_1^T$ mátrixszal végrehajtva a hatványmódszert, megkaphatjuk a λ_2 sajátértéket és a $\overline{\mathbf{v}}_2$ sajátvektort.

[2] A defláció szó jelentése leapasztás vagy leeresztés.

Blokk háromszögmátrix defláció

Egy $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrixot blokk felső háromszögmátrixnak nevezünk, ha vannak olyan $\mathbf{A}_{kl}$ $(k,l = 1,\ldots,m)$ mátrixok ($\mathbf{A}_{11},\ldots,\mathbf{A}_{mm}$ kvadratikus mátrixok), melyekkel

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} & \ldots & \mathbf{A}_{1m} \\ \mathbf{0} & \mathbf{A}_{22} & \mathbf{A}_{23} & \ldots & \vdots \\ \mathbf{0} & \mathbf{0} & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & 0 & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{mm} \end{bmatrix}.$$

A blokk alsó háromszögmátrix hasonlóan definiálható. Ha egy mátrix blokk háromszögmátrix alakú, akkor a sajátértékei a diagonálisban szereplő $\mathbf{A}_{11},\ldots,\mathbf{A}_{mm}$ mátrixok sajátértékeinek uniójaként adódnak. (Ez könnyen igazolható a karakterisztikus polinom vizsgálatával.) Így a sajátérték-meghatározást visszavezethetjük kisebb méretű mátrixok sajátértékeinek meghatározására. Különösen gyakran alkalmazzuk az eljárást akkor, ha $m = 2$, és $\mathbf{A}_{11}$ $(n-1)\times(n-1)$-es, $\mathbf{A}_{22}$ pedig egy 1×1-es mátrix. Gyakori az az eset is, hogy a mátrix egyszerű permutációkkal végzett hasonlósági transzformációkkal blokk háromszögmátrix alakra hozható (lásd pl. 4.5.2. feladat).

4.3. A sajátértékeket egyszerre közelítő eljárások

Most rátérünk azokra a módszerekre, amelyek a mátrixok minden sajátértékére egyszerre mondanak közelítést. Ezen módszerek közös alapgondolata az, hogy ha egy mátrixszal hasonlósági transzformációt végzünk, akkor annak sajátértékei nem változnak meg. Tehát ha pl. elő tudnánk állítani egy mátrix Schur-felbontását, azaz ha fel tudnánk írni az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrixot $\mathbf{A} = \mathbf{STS}^H$ alakban, ahol $\mathbf{T}$ felső háromszögmátrix és $\mathbf{S}$ unitér mátrix, akkor az $\mathbf{S}^H\mathbf{AS} = \mathbf{T}$ egyenlőség miatt $\mathbf{A}$ sajátértékei megegyeznének $\mathbf{T}$ diagonális elemeivel. A fő problémát az $\mathbf{S}$ unitér mátrix megkeresése jelenti. Direkt módszerrel $\mathbf{S}$ általában nem határozható meg. Így iterációs módszereket kell találnunk, melyek az $\mathbf{S}$ mátrixot egy mátrixsorozat határértékeként állítják elő. Két módszert ismertetünk most ennek megvalósítására. A továbbiakban ismét valós szimmetrikus mátrixokkal foglalkozunk csak, így a Schur-felbontás tulajdonképpen az $\mathbf{A} = \mathbf{S\Lambda S}^T$ felbontást jelenti, ahol $\mathbf{S}$ az ortonormált sajátvektorok mátrixa, $\mathbf{\Lambda}$ pedig a sajátértékek diagonálmátrixa.

4.3.1. A Jacobi-módszer

A most ismertetendő módszert először Jacobi írta le 1845-ben.[3] Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy valós szimmetrikus mátrix. Keresünk egy olyan $\mathbf{S}$ ortogonális mátrixot és $\mathbf{\Lambda}$ diagonális mátrixot, melyekkel $\mathbf{S}^T\mathbf{AS} = \mathbf{\Lambda}$.

Az $\mathbf{S}$ mátrixot 2×2-es szimmetrikus mátrixokra könnyen meghatározhatjuk. Legyen pl.

$$\mathbf{A} = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

$(b \neq 0)$ és keressük az $\mathbf{S}$ mátrixot

$$\mathbf{S} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$$

[3] Elméleti eredményként jó darabig feledésbe merült a módszer. A számítógépek elterjedésével az 1950-es évektől kezdték nagyméretű mátrixok sajárértékeinek meghatározására használni. Manapság párhuzamosíthatósága miatt népszerű.

alakban, ahol $s = \sin\theta$ és $c = \cos\theta$, valamilyen θ paraméterrel. (Az $\mathbf{S}$ mátrix elemeinek ezen megválasztása mellett $\mathbf{S}$ valóban ortogonális mátrix lesz.) A c és s értékeknek olyanoknak kell lenniük, hogy a

$$\begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} a & b \\ b & d \end{bmatrix} \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$$
$$= \begin{bmatrix} * & sca + c^2b - s^2b - scd \\ sca - s^2b + c^2b - scd & * \end{bmatrix}$$

mátrix diagonális legyen, azaz igaz legyen a

$$0 = sca - s^2b + c^2b - scd = sc(a-d) + b(c^2 - s^2) = (a-d)\sin(2\theta)/2 + b\cos(2\theta)$$

egyenlőség. Továbbá nyilvánvalóan $s^2 + c^2 = 1$.

Ha $a = d$, akkor a $\cos(2\theta) = 0$ egyenlőségnek kell teljesülnie. Ezt és az $s^2 + c^2 = 1$ összefüggést felhasználva s és c meghatározható.

Ha $a \neq d$, akkor $\mathrm{tg}(2\theta) = -2b/(a-d)$. Ebből és az $s^2 + c^2 = 1$ összefüggésből s és c ismét meghatározható (lásd 4.5.8. feladat).

Ezzel a 2×2-es szimmetrikus mátrixokhoz megadtunk olyan ortogonális mátrixot, amely a hasonlósági transzformáció során azt diagonális mátrixba viszi. Nagyobb méretű mátrixoknál azonban ez nehéz feladat. A Jacobi-módszer alapötlete az, hogy nagyobb méretű mátrixok esetén is a 2×2-es esetre levezetett diagonalizáló mátrixot használjuk, csak sokszor egymás után. Az algoritmus a következő.

- Válasszuk ki a mátrix főátlója felett a legnagyobb abszolút értékű elemet. Legyen ez az a_{ij} elem ($i < j$).
- Határozzuk meg a diagonalizáló $\mathbf{S}$ mátrixot az

  $$\begin{bmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{bmatrix}$$

  2×2-es szimmetrikus mátrixra, azaz adjuk meg a benne szereplő s és c paramétereket.
- Definiáljuk az alábbi $n \times n$-es ortogonális mátrixot:

  $$\mathbf{S}_{ij} = [\overline{\mathbf{e}}_1, \ldots, \overline{\mathbf{e}}_{i-1}, c\overline{\mathbf{e}}_i - s\overline{\mathbf{e}}_j, \overline{\mathbf{e}}_{i+1}, \ldots, \overline{\mathbf{e}}_{j-1}, s\overline{\mathbf{e}}_i + c\overline{\mathbf{e}}_j, \overline{\mathbf{e}}_{j+1}, \ldots, \overline{\mathbf{e}}_n],$$

  azaz

  $$\mathbf{S}_{ij} = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & c & & & & s & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & -s & & & & c & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix}.$$

- Az $\mathbf{A}^{(1)} = \mathbf{S}_{ij}^T \mathbf{A} \mathbf{S}_{ij}$ mátrixban az $a_{ij}^{(1)}$ és $a_{ji}^{(1)}$ elemek nullák lesznek (így választottuk $\mathbf{S}$-t). Azt mondjuk, hogy a fenti eljárással végrehajtottunk egy (i, j) elemhez tartozó Jacobi-transzformációt.

- Ezután visszalépünk az eljárás elejére újra kiválasztva a főátló felett a legnagyobb abszolút értékű elemet.

Könnyű látni, hogy minden lépésben az (i,j) és (j,i) helyeken álló főátlón kívüli elemek lenullázódnak, de egy következő lépés újra nullától különbözővé teheti őket. Mutassuk meg, hogy az algoritmus valóban a várt eredményt adja! Ehhez igazolnunk kell, hogy az algoritmussal nyert $\mathbf{S}_{ij}$ mátrixok szorzata határértékben éppen az $\mathbf{S}$ diagonalizáló mátrixhoz fog tartani.

4.3.1. tétel.
Tegyük fel, hogy a fent ismertetett módon végrehajtottunk egy (i,j) elemhez tartozó Jacobi-transzformációt. Ekkor az $\mathbf{A}^{(1)}$ mátrix főátlón kívüli elemeinek négyzetösszege $2a_{ij}^2$-tel kevesebb lesz, mint az $\mathbf{A}$ mátrix esetén.

Bizonyítás: A Jacobi-transzformáció csak az i-edik és a j-edik sort és oszlopot változtathatja meg. Nyilvánvalóan $l \neq i,j$ esetén ezen oszlopok és sorok elemei az új mátrixban

$$a_{il}^{(1)} = ca_{il} - sa_{jl},\ a_{jl}^{(1)} = ca_{jl} + sa_{il},$$
$$a_{li}^{(1)} = ca_{li} - sa_{lj},\ a_{lj}^{(1)} = ca_{lj} + sa_{li}$$

alakúak. Igaz továbbá, hogy

$$(a_{il}^{(1)})^2 + (a_{jl}^{(1)})^2 = (a_{il})^2 + (a_{jl})^2,$$
$$(a_{li}^{(1)})^2 + (a_{lj}^{(1)})^2 = (a_{li})^2 + (a_{lj})^2.$$

Az első egyenlőség az

$$(a_{il}^{(1)})^2 + (a_{jl}^{(1)})^2 = (ca_{il} - sa_{jl})^2 + (ca_{jl} + sa_{il})^2$$
$$= (c^2+s^2)(a_{il})^2 + (s^2+c^2)(a_{jl})^2 = (a_{il})^2 + (a_{jl})^2$$

módon látható, a másik pedig hasonlóan igazolható. Számítsuk ki most a négyzetösszeg-változást a mátrix főátló feletti részén:

$$-a_{ij}^2 + \sum_{l=i+1,\neq j}^{n} [(\overbrace{ca_{il} - sa_{jl}}^{a_{il}^{(1)}})^2 - (a_{il})^2] + \sum_{l=j+1}^{n} [(\overbrace{ca_{jl} + sa_{il}}^{a_{jl}^{(1)}})^2 - (a_{jl})^2]$$
$$+ \sum_{l=1}^{i-1} [(\overbrace{ca_{li} - sa_{lj}}^{a_{li}^{(1)}})^2 - (a_{li})^2] + \sum_{l=1,\neq i}^{j-1} [(\overbrace{ca_{lj} + sa_{li}}^{a_{lj}^{(1)}})^2 - (a_{lj})^2]$$
$$= -a_{ij}^2 + \sum_{l=i+1}^{j-1} ((a_{il}^{(1)})^2 - (a_{il})^2) + \sum_{l=i+1}^{j-1} \underbrace{((a_{lj}^{(1)})^2 - (a_{lj})^2)}_{(a_{jl}^{(1)})^2 - (a_{jl})^2} = -a_{ij}^2.$$

Így a teljes főátlón kívüli rész négyzetösszeg-változása valóban $-2a_{ij}^2$. ■

Legyen a főátlón kívüli elemeket tartalmazó mátrix a k-adik lépésben $\mathbf{B}_k = \mathbf{A}^{(k)} - \mathrm{diag}(\mathrm{diag}(\mathbf{A}^{(k)}))$. Ekkor

$$\|\mathbf{B}_k\|_F^2 = \|\mathbf{B}_{k-1}\|_F^2 - 2(a_{ij}^{(k-1)})^2 \leq \|\mathbf{B}_{k-1}\|_F^2 - 2\frac{\|\mathbf{B}_{k-1}\|_F^2}{n(n-1)}$$

$$= \|\mathbf{B}_{k-1}\|_F^2 \left(1 - \frac{2}{n(n-1)}\right),$$

így

$$\|\mathbf{B}_k\|_F^2 \leq \|\mathbf{B}_0\|_F^2 \left(1 - \frac{2}{n(n-1)}\right)^k.$$

Ez a becslés mutatja, hogy a módszer konvergenciarendje legalább egy (valójában másodrendű).

A Jacobi-transzformációkat végrehajtva tehát a mátrix főátlón kívüli elemei nullához tartanak. Így az $\mathbf{S}_{ij}$ mátrixok szorzata valóban az $\mathbf{S}$ diagonalizáló mátrixhoz fog tartani. Elegendően sok lépést végrehajtva a főátlón kívüli elemek annyira lecsökennek, hogy a főátlóban megkapjuk a sajátértékek közelítéseit a Gersgorin-tételt alkalmazva.

4.3.2. megjegyzés. A legnagyobb elem kiválasztása $n(n-1)/2$ összehasonlítást jelent, ami sokáig tart. Ezért inkább az a szokás, hogy a főátló feletti részen soronként haladunk végig a Jacobi-transzformációkkal (azaz ilyenkor nem a legnagyobb abszolút értékű elemet választjuk ki – ez a módosított Jacobi-módszer). A tapasztalat szerint 5-szöri végighaladás elég szokott lenni egy megfelelő sajátérték-közelítéshez. ⋄

4.3.3. megjegyzés. A módosított Jacobi-módszer esetén egy Jacobi-transzformáció $4n$ szorzást és $2n$ összeadást igényel, és $n(n-1)/2$ elemen 5-ször kell végighaladni. Ez kb. $15n^3$ flop műveletet igényel. ⋄

4.3.4. megjegyzés. Ha egymás után az $\mathbf{S}_{i_1 j_1}, \ldots, \mathbf{S}_{i_k j_k}$ ortogonális mátrixokat használtuk a Jacobi-transzformációhoz, akkor a λ_j sajátértékhez tartozó sajátvektor az $\mathbf{S}_{i_1 j_1} \cdot \ldots \cdot \mathbf{S}_{i_k j_k} \overline{\mathbf{e}}_j$ módon közelíthető. ⋄

4.3.5. megjegyzés. A Jacobi-módszer olyan mátrixokra is konvergál, melyek nem feltétlenül szimmetrikusak, de minden sajátértékük valós. ⋄

4.3.2. QR-iteráció

A módszer alapötlete az, hogy az ortogonális diagonalizáló mátrixot az $\mathbf{A}$ mátrix QR-felbontásának $\mathbf{Q}$ mátrixával közelítjük. Az algoritmus a következő:

QR-iteráció, $\mathbf{A}$ adott mátrix, $\mathbf{A}^{(0)} := \mathbf{A}$
for $k := 1 : k_{\max}$ **do**
  Készítsük el $\mathbf{A}^{(k-1)}$ QR-felbontását: $\mathbf{A}^{(k-1)} = \mathbf{Q}^{(k-1)}\mathbf{R}^{(k-1)}$
  $\mathbf{A}^{(k)} := \mathbf{R}^{(k-1)}\mathbf{Q}^{(k-1)}$
end for

Először is látnunk kell, hogy az iteráció $\mathbf{A}$-hoz hasonló mátrixokat készít, hiszen

$$\mathbf{A}^{(k)} = \mathbf{R}^{(k-1)}\mathbf{Q}^{(k-1)} = (\mathbf{Q}^{(k-1)})^T\mathbf{Q}^{(k-1)}\mathbf{R}^{(k-1)}\mathbf{Q}^{(k-1)} = (\mathbf{Q}^{(k-1)})^T\mathbf{A}^{(k-1)}\mathbf{Q}^{(k-1)}.$$

Tehát

$$\mathbf{A}^{(k)} = (\mathbf{Q}^{(k-1)})^T \ldots (\mathbf{Q}^{(0)})^T \mathbf{A} \underbrace{\mathbf{Q}^{(0)} \ldots \mathbf{Q}^{(k-1)}}_{=:\mathbf{Q}_k} = \mathbf{Q}_k^T \mathbf{A} \mathbf{Q}_k,$$

és így a hasonlósági transzformációk miatt $\mathbf{A}^{(k)}$ sajátértékei megegyeznek $\mathbf{A}$ sajátértékeivel.

Minden lépésben szükség van egy QR-felbontásra. Ez elég költséges eljárás, hiszen Householder-tükrözésekkel $4n^3/3$ flop a műveletigénye. A gyakorlatban általában úgy érdemes eljárni, hogy először a mátrixot Hessenberg-alakra transzformáljuk hasonlósági transzformáció segítségével (hogy a sajátértékek ne változzanak), majd pedig ezzel az új mátrixszal hajtjuk végre a QR-iterációt. Az iteráció során a Hessenberg-struktúra megmarad. A QR-felbontást így minden lépésben Givens-forgatásokkal hajtjuk végre ($3n^2$ flop iterációs lépésenként).

A Hessenberg alakra való transzformációt Householder-tükrözésekkel végezzük el úgy, hogy először $\tilde{\mathbf{H}}_1$ legyen az $\mathbf{A}(2:n,1)$ oszlophoz tartozó Householder-tükrözési mátrix, majd $\mathbf{H}_1 = \operatorname{diag}(1, \tilde{\mathbf{H}}_1)$, $\tilde{\mathbf{H}}_2$. Legyen $\mathbf{A}^{(2)} = \mathbf{H}_1\mathbf{A}\mathbf{H}_1$. Ezután legyen $\tilde{\mathbf{H}}_2$ az $\mathbf{A}^{(2)}(3:n,2)$ oszlophoz tartozó Householder-tükrözési mátrix, majd $\mathbf{H}_2 = \operatorname{diag}(1,1,\tilde{\mathbf{H}}_2)$, és definiáljuk az $\mathbf{A}^{(3)} = \mathbf{H}_2\mathbf{A}^{(1)}\mathbf{H}_2$ mátrixot. Hasonlóan eljárva kapjuk $n-1$ lépés után az $\mathbf{A}^{(n)} = \mathbf{H}_{n-1}\dots\mathbf{H}_2\mathbf{H}_1\mathbf{A}\mathbf{H}_1\mathbf{H}_2\dots\mathbf{H}_{n-1}$ Hessenberg alakú, $\mathbf{A}$-val hasonló mátrixot. Az eljárást, melynek műveletigénye $4n^3/3$ flop, sematikusan az alábbi módon foglalhatjuk össze:

$$\begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \end{bmatrix}.$$

Igazoljuk, hogy ha $\mathbf{A}$ egy felső Hessenberg-mátrix, akkor a QR-iteráció megtartja ezt a struktúrát. Legyen $\mathbf{A} = \mathbf{QR}$ felső Hessenberg-mátrix. Ekkor az

$$\mathbf{A}^{(1)} = \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{Q}^T\mathbf{Q}\mathbf{R}\mathbf{Q} = \mathbf{R}\mathbf{Q} = \mathbf{R}\mathbf{Q}\mathbf{R}\mathbf{R}^{-1} = \mathbf{R}\mathbf{A}\mathbf{R}^{-1}$$

mátrix is felső Hessenberg-mátrix, hiszen felső Hessenberg-mátrixok szorzata is felső Hessenberg-mátrix lesz. Itt feltettük, hogy $\mathbf{R}$, azaz $\mathbf{A}$ is invertálható mátrix. Különben a nulla sajátértéke lenne és deflációs eljárással kisebb méretű mátrix vizsgálatára térhetnénk át. A továbbiakban tegyük fel, hogy az eredeti $\mathbf{A}$ mátrix már Hessenberg alakra van transzformálva.

Az általunk vizsgált szimmetrikus mátrixokra a Hessenberg alakra transzformált mátrix egy szimmetrikus, tridiagonális mátrix lesz. Egy ilyen mátrixszal kell ezután végrehajtani a QR-iterációt. A továbbiakban feltesszük, hogy a QR-iterációt már mindig egy Hessenberg alakra hozott mátrixszal indítjuk.

Az, hogy a $\mathbf{Q}_k$-val jelölt $\mathbf{Q}^{(0)}\dots\mathbf{Q}^{(k-1)}$ mátrix valóban az $\mathbf{A}$-t diagonalizáló $\mathbf{S}$ mátrixhoz tart, egyáltalán nem nyilvánvaló az algoritmusból. Az alábbi konvergenciatételt bizonyítás nélkül közöljük.

4.3.6. tétel. (pl. Quarteroni–Sacco–Saleri [29], 202. oldal)
Ha az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrixnak minden sajátértéke valós és abszolút értékben különböző ($|\lambda_1| > |\lambda_2| > |\lambda_3| > \dots > |\lambda_n|$), akkor

$$\lim_{k\to\infty} \mathbf{A}^{(k)} = \begin{bmatrix} \lambda_1 & \tilde{a}_{12} & \dots & \tilde{a}_{1n} \\ 0 & \lambda_2 & \tilde{a}_{23} & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_n \end{bmatrix}$$

valamilyen megfelelő $\tilde{a}_{ij}$ konstansokkal, azaz a határértékmátrix egy felső háromszögmátrix. Ha $\mathbf{A}$ szimmetrikus, akkor $\{\mathbf{A}^{(k)}\}$ diagonális mátrixhoz tart.

4.3.7. megjegyzés. Bár mi csak valós, szimmetrikus mátrixok sajátértékeivel foglalkozunk, a fenti tétel olyan mátrixokra is vonatkozik, melyeknek minden sajátértéke abszolút értékben különböző. ◇

4.3.8. megjegyzés. Igazolható, hogy ha k elég nagy, akkor az $\mathbf{A}^{(k)}$ mátrixok szubdiagonálbeli elemeire (a Hessenberg alak miatt csak ezek különbözhetnek nullától a főátló alatt)

$$|a_{i,i-1}^{(k)}| \approx \left|\frac{\lambda_i}{\lambda_{i-1}}\right|^k, \ i = 2, \ldots, n.$$

Ez mutatja, hogy a QR-iteráció konvergenciarendje lineáris, és hogy a konvergenciasebességet az abszolút értékben egymást követő sajátértékek hányadosa határozza meg. Ha vannak abszolút értékben egymáshoz közeli sajátértékek, akkor lassan konvergál csak a módszer. ◇

4.3.9. példa. Határozzuk meg QR-iterációval a $\mathbf{H}_6$ Hilbert mátrix sajátértékeit! Állítsuk le akkor a QR-iterációt, ha a főátlón kívüli elemeket tartalmazó mátrix maximumnormája már csak legfeljebb 10^{-6}-szorosa a főátló legkisebb elem abszolút értékének! A 16. iteráció után az alábbi mátrixot nyerjük

$$\begin{bmatrix} 1.6189 & 7.1945 \times 10^{-14} & -2.0429 \times 10^{-16} & -4.8395 \times 10^{-17} & -8.0718 \times 10^{-17} & -1.1143 \times 10^{-17} \\ 7.1945 \times 10^{-14} & 0.2424 & 3.2897 \times 10^{-17} & 4.3090 \times 10^{-17} & 4.3450 \times 10^{-17} & -4.8171 \times 10^{-17} \\ 0 & 0 & 1.6322 \times 10^{-2} & -1.4419 \times 10^{-17} & -2.0930 \times 10^{-17} & 1.2691 \times 10^{-17} \\ 0 & 0 & -4.9 \times 10^{-25} & 6.1575 \times 10^{-4} & 1.6031 \times 10^{-17} & -5.3934 \times 10^{-18} \\ 0 & 0 & 0 & 0 & 1.2571 \times 10^{-5} & -9.3562 \times 10^{-18} \\ 0 & 0 & 0 & 0 & 0 & 1.0828 \times 10^{-7} \end{bmatrix}$$

Figyeljük meg, hogy az iteráció során abszolút értékben csökkenő módon kerülnek a sajátérték-közelítések a főátlóba. A legnagyobb abszolút értékű sajátértékről pl. a Gersgorin-tételt alkalmazva azt mondhatjuk, hogy értéke az $1.61889985892434 \pm 7.2289 \times 10^{-14}$ intervallumba esik (itt több tizedesjegyre pontosan írtuk ki a számokat, mint a mátrixban). ◇

Láttuk, hogy a QR-iteráció elsőrendű módszer sajátérték meghatározására. A módszer felgyorsítható az ún. eltolt QR-iteráció alkalmazásával. Az algoritmusban szereplő μ_k értékek megválasztásának módjáról később lesz szó.

Eltolt QR-iteráció, $\mathbf{A}$ Hessenberg alakú mátrix, $\mathbf{A}^{(0)} := \mathbf{A}$, μ_k adott paraméterek

for $k := 1 : k_{\max}$ **do**
  Készítsük el $\mathbf{A}^{(k-1)} - \mu_{k-1}\mathbf{E}$ QR-felbontását: $\mathbf{A}^{(k-1)} - \mu_{k-1}\mathbf{E} = \mathbf{Q}^{(k-1)}\mathbf{R}^{(k-1)}$
  $\mathbf{A}^{(k)} := \mathbf{R}^{(k-1)}\mathbf{Q}^{(k-1)} + \mu_{k-1}\mathbf{E}$
end for

Az eljárás valóban mindig $\mathbf{A}$-hoz hasonló mátrixot ad, ugyanis

$$\begin{aligned} \mathbf{A}^{(k)} &= \mathbf{R}^{(k-1)}\mathbf{Q}^{(k-1)} + \mu_{k-1}\mathbf{E} = (\mathbf{Q}^{(k-1)})^T\mathbf{Q}^{(k-1)}\mathbf{R}^{(k-1)}\mathbf{Q}^{(k-1)} + \mu_{k-1}\mathbf{E} \\ &= (\mathbf{Q}^{(k-1)})^T(\mathbf{A}^{(k-1)} - \mu_{k-1}\mathbf{E})\mathbf{Q}^{(k-1)} + \mu_{k-1}\mathbf{E} = (\mathbf{Q}^{(k-1)})^T\mathbf{A}^{(k-1)}\mathbf{Q}^{(k-1)} \end{aligned}$$

minden $k = 1, 2, \ldots$ esetén.

Az iteráció során ha az utolsó sor utolsó előtti eleme kicsivé válna, akkor a főátló utolsó eleme nagy pontossággal a mátrix sajátértékét adná. Ezek után már elegendő lenne csak a mátrix $(1 : n-1, 1 : n-1)$ blokkjának megkeresni a sajátértékeit (defláció).

Az utolsó sor utolsó előtti eleme a μ_k értékek ügyes megválasztásával tehető gyorsan kicsivé (gyakorlatilag nullává). Igazolható, hogy ez az elem elegendően nagy k értékekre

$$|(\lambda_n - \mu_k)/(\lambda_{n-1} - \mu_k)|^k$$

nagyságrendű. Ezért egy gyakran alkalmazott választás $\mu_k = a_{nn}^{(k)}$, azaz μ_k-nak a főátló utolsó elemét szokás választani. Ekkor ugyanis, ha a főátló utolsó eleme közelít egy λ_n sajátértékhez, akkor a $|(\lambda_n - \mu_k)/(\lambda_{n-1} - \mu_k)|$ hányados nagyon pici lesz, és ennek hatványai gyorsan nullához tartanak.

4.4. Sajátértékszámítás a MATLAB-ban

A MATLAB a sajátértékeket és a sajátvektorokat az `eig` parancs segítségével számolja. Ritka mátrixok esetén az `eigs` parancsot használjuk. Hasznos lehet még a `hess` parancs is, amely egy mátrixot transzformál felső Hessenberg alakra hasonlósági transzformációval.

```
>> A=[2,-1,0;-1,2,-1;0,-1,2] % Mátrix definiálás

A =

     2    -1     0
    -1     2    -1
     0    -1     2

>> [V,L]=eig(A) % V: sajátvektorok mátrixa, L: sajátértékek mátrixa

V =

    0.5000   -0.7071   -0.5000
    0.7071    0.0000    0.7071
    0.5000    0.7071   -0.5000


L =

    0.5858         0         0
         0    2.0000         0
         0         0    3.4142

>> A=rand(3) % Egy véletlen 3x3-as mátrix

A =

    0.9501    0.4860    0.4565
```

```
    0.2311    0.8913    0.0185
    0.6068    0.7621    0.8214

>> [S,H]=hess(A) % Hessenberg alakra hozás parancsa (H=SAS^T)

S =              % ortogonális mátrix

    1.0000         0         0
         0   -0.3559   -0.9345
         0   -0.9345    0.3559


H =              % Hessenberg alakú mátrix

    0.9501   -0.5996   -0.2917
   -0.6494    1.0899    0.6864
         0   -0.0571    0.6228
```

Végül közlünk egy egyszerű programot a hatványmódszer megvalósítására. A program egyszerű módosítással átírható úgy, hogy az inverz iterációt vagy a Rayleigh-hányados iterációt hajtsa végre. A programban `A` az adott mátrix, `kmax` a maximális iterációszám és `toll` a toleranciaszint. Ez utóbbi azt jelenti, hogy ha két egymás utáni sajátérték-közelítés közelebb van egymáshoz, mint ez az érték, akkor leállítjuk az iterációt. Az `iter` érték a végrehajtott iterációs lépések száma. Ha `iter`=`kmax`, akkor amiatt állt le az eljárás, mert elértük a maximális iterációszámot, és nem amiatt, mert kicsi a hiba. Az `y` kimeneti érték a sajátvektor-közelítés, `nu` a sajátérték-közelítés.

```
function [y,nu,iter]=hatvanymodszer(A,kmax,toll);
   n=max(size(A));
   y=(ones(n,1));
   y=y/norm(y);
   nu=y'*A*y;
   err=1;
   iter=0;
while err>toll & iter<kmax
   iter=iter+1;
   y=A*y;
   y=y/norm(y);
   nuold=nu;
   nu=y'*A*y;
   err=abs(nu-nuold);
end;
```

4.5. Feladatok

4.5.1. feladat. Igazoljuk, hogy az $\mathbf{A}$ mátrix sajátértékei mind valósak (a sajátértékek kiszámítása nélkül), ill. adjunk meg egy lehetőleg rövid intervallumot, amiből a $\mathbf{B}$ mátrix sajátértékei kikerülhetnek!

$$\mathbf{A} = \begin{bmatrix} 3 & 5 & 2 \\ 0 & 4 & 1 \\ 0 & 1 & 5 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 7 & 0 \\ -1 & 0 & 5 \end{bmatrix}$$

4.5.2. feladat. A sajátértékek kiszámítása nélkül mondjuk meg, hogy hány nem valós sajátértéke van az alábbi mátrixnak? (Alkalmazzunk hasonlósági transzformációkat permutációs mátrixokkal!)

$$\begin{bmatrix} -4 & 0 & 0 & 0.5 & 0 \\ 2 & 2 & 4 & -3 & 1 \\ 0.5 & 0 & -1 & 0 & 0 \\ 0.5 & 0 & 0.2 & 3 & 0 \\ 2 & 0.5 & -1 & 3 & 4 \end{bmatrix}$$

4.5.3. feladat. Tekintsük az

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 2 & 9 \\ 1 & 2 \end{bmatrix}$$

mátrixokat. Becsüljük meg, hogy ha ezen mátrixok második sorának első elemeihez 0.1-et adunk, akkor legfeljebb mennyivel térnek el az új mátrixok sajátértékei az eredetiektől! Ellenőrizzük az eredményt a MATLAB segítségével! (Használjuk a Bauer–Fike-tételt!)

4.5.4. feladat. Alkalmazzuk a hatványmódszert az $\mathbf{A} = \text{tridiag}(-1, 2, -1)$ mátrixra (sajátértékek: $2, 2+\sqrt{2}, 2-\sqrt{2}$, sajátvektorok: $[-1, 0, 1]^T$, $[1, -\sqrt{2}, 1]^T$, $[1, \sqrt{2}, 1]^T$). Állítsuk le a számítógépen végrehajtott iterációt a 20. lépés után. Becsüljük meg λ_1 értékét az $\overline{\mathbf{y}}^{(20)}$ vektorral úgy, hogy az elemeivel osztjuk az $\mathbf{A}\overline{\mathbf{y}}^{(20)}$ vektor elemeit! Hasonlítsuk össze ezeket az értékeket a Rayleigh-hányados által adott értékkel!

4.5.5. feladat. Határozzuk meg az előző feladat mátrixának legkisebb abszolút értékű sajátértékét a korábbi program módosításával, ha tudjuk, hogy értéke kb. 0.5! Használjuk az inverz iterációt!

4.5.6. feladat. Határozzuk meg az $\mathbf{A} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{50\times 50}$ mátrix legnagyobb és legkisebb abszolút értékű sajátértékét és a hozzájuk tartozó sajátvektorokat. A legkisebb abszolút értékű sajátértéket kétféleképpen is határozzuk meg! Először a mátrix inverzének és a hatványmódszernek a használatával, és másodszor azt az információt használva, hogy értéke kb. 0.003! Határozzuk meg a mátrix 1.1-hez legközelebbi sajátértékét!

4.5.7. feladat. Egy $\mathbf{A}$ 4×4-es mátrixról tudjuk, hogy sajátértékei a 20, 10, 5 és 1 számok közelében vannak. Milyen α számmal alkalmazzuk a hatványmódszert az $\mathbf{A} - \alpha\mathbf{E}$ mátrixra, hogy az az 1 közeli sajátértéket és a hozzá tartozó sajátvektort adja meg?

4.5.8. feladat. A Jacobi-iterációs módszernél keresnünk kell egy olyan θ szöget, melyre teljesül a $\text{ctg}(2\theta) = (d-a)/(2b)$ egyenlőség, majd ezután meg kell határoznunk a $\sin\theta$ és $\cos\theta$ értékeket. Ezek az értékek megkaphatók a θ szög kiszámolása nélkül is. Igazoljuk először, hogy olyan szögekre, melyekre a szereplő függvények értelmezve vannak, igaz a

$$\text{t}g^2\theta + 2\text{ctg}(2\theta) \cdot \text{t}g\theta - 1 = 0$$

egyenlőség. Hogyan lehet ebből $\sin\theta$ és $\cos\theta$ értékét kiszámítani?

4.5.9. feladat. Alkalmazzuk a Jacobi-iterációt az 5×5-ös Hilbert-mátrixra ill. az

$$\mathbf{A} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{50\times 50}$$

mátrixra számítógép segítségével!

4.5.10. feladat. Írjunk MATLAB programot a QR-iterációs és eltolt QR-iterációs sajátérték-meghatározásra, és határozzuk meg újra az 5×5-ös Hilbert-mátrix és az $\mathbf{A} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{50\times 50}$ mátrix sajátértékeit!

Ellenőrző kérdések

1. Milyen két nagy csoportja van a sajátérték-meghatározási módszereknek?
2. Ismertessük a hatványmódszer működését!
3. Hogyan módosítható a hatványmódszer úgy, hogy a domináns sajátértékeken kívül más sajátértékeket is megtaláljon?
4. Ha adott egy mátrix egyik sajátértékére egy közelítés, akkor hogyan határozhatjuk meg a hozzá tartozó sajátvektort?
5. Ha adott egy mátrix egyik sajátvektorára egy közelítés, akkor hogyan határozhatjuk meg a hozzá tartozó sajátértéket?
6. Mi az a Rayleigh-hányados, és milyen tulajdonságai vannak?
7. Ismertessük a Jacobi-módszert!
8. Ismertessük a QR-iteráció lényegét! Milyen módszerrel lehet a konvergenciáját felgyorsítani?

5. Nemlineáris egyenletek és egyenletrendszerek megoldása

Ebben a fejezetben azt vizsgáljuk meg, hogy hogyan lehet egy nemlineáris egyenlet megoldásait numerikusan közelíteni. Megismerjük az intervallumfelezési-, a húr-, a szelő- és a Newton-módszereket, valamint a fixpont-iterációs módszereket, és megvizsgáljuk ezen módszerek tulajdonságait.

5.1. Nemlineáris egyenletek

Gyakorlati problémák megoldása során gyakran találkozunk azzal a feladattal, hogy egyenletet vagy egyenletrendszert kell megoldanunk. Sok esetben a megoldás explicit módon előállítható, azaz véges sok alapművelet segítségével megadható. A legegyszerűbb ilyen eset az $ax = b$ lineáris egyenlet vagy az $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ lineáris egyenletrendszer. A nemlineáris egyenletek megoldása komplikáltabb. Egy bizonyos részük explicit módon könnyen megoldható (ilyen polinom-, trigonometrikus-, exponenciális- vagy logaritmikus egyenletekkel találkoztunk középiskolai tanulmányaink során), de sok olyan egyenlet is van, amely nem ilyen. Például az $x^2 = 4\sin x$ vagy az $x = \cos x$ egyenleteket nem tudjuk megoldani. Mivel a legalább ötödfokú polinomok zérushelyének meghatározására nincs megoldóképlet (olyan képlet, ami az együtthatókból megmondaná a zérushelyeket), ezért ezek megoldása sem lehetséges általában explicit módon. Ilyen polinom például az $x^5 - 4x^4 + x^3 - x^2 + 4x - 4 = 0$ polinom, amely racionális zérushelykereséssel, vagy valamilyen más egyszerű módszerrel nem alakítható szorzattá, így nem lehet a zérushely keresését visszavezetni alacsonyabbfokú polinomokéira. Explicit megoldás hiányában a nemlineáris egyenletek megoldására általában iterációs módszereket használunk, melyek lényege, hogy olyan sorozatot állítunk elő, amely valamelyik megoldáshoz konvergál.

Azok a módszerek, amelyeket be fogunk mutatni ebben a fejezetben, több száz évvel ezelőtt keletkeztek. Ennek ellenére mégis fontos őket megismernünk, mert még ma is ezeket a módszereket használjuk nemlineáris egyenletek megoldására. Természetesen ma már a fáradtságos számolásokat elvégzik helyettünk a számítógépek, de pl. a nemlineáris egyenletrendszerek megoldása (főleg a legalább három ismeretlent tartalmazóké) még mindig nehéz feladat.

5.1.1. A gyökök elkülönítése

A nemlineáris egyenletek megoldása során az első lépés általában az, hogy megadunk olyan intervallumokat, melyekben biztosan van megoldása az egyenletnek. Ezt a lépést röviden úgy hívjuk, hogy elkülönítjük a gyököket. Számítógéppel a gyökök elkülönítése gyorsan megoldható, hiszen a grafikonból könnyen látható, hogy hol vannak az egyenlet gyökei. Amikor nem áll rendelkezésünkre számítógép, akkor az alábbi tétel segítségével végezhető el az elkülönítés.

5.1.1. tétel.
Ha egy folytonos függvény esetén $f(a) \cdot f(b) < 0$ $(a < b)$, akkor van olyan $c \in (a, b)$, melyre $f(c) = 0$.

Ez a tétel a Bolzano-tétel közvetlen következménye. Ha tehát bizonyos pontokban kiszámítjuk a függvényértékeket, és két egymást követő pontban ellentétes a függvényértékek előjele, akkor a pontok között van zérushely. Ha azt is tudjuk még, hogy a függvény szigorúan monoton az adott intervallumon (pl. deriválható, és a deriváltja nem vált előjelet), akkor azt is tudjuk már, hogy az adott intervallumban pontosan egy zérushely lesz.

A gyökök elkülönítésében segíthet az is, ha az egyenlet két oldalán szereplő függvényeket tudjuk külön-külön ábrázolni, mert akkor a két grafikon metszéspontjainak abszcisszái adják a megoldásokat. Ez a módszer alkalmazható pl. az $x^2 = 4\sin x$ vagy az $x = \cos x$ egyenletek esetén.

Ha polinomok zérushelyeit keressük, akkor hasznos tudni, hogy a valós zérushelyek melyik intervallumból kerülhetnek ki egyáltalán. Ezt a kérdést vizsgálja meg az alábbi tétel.

5.1.2. tétel.

A $p(x) = a_n x^n + \ldots + a_1 x + a_0$ $(a_n, a_0 \neq 0)$ polinom zérushelyei az origó közepű $R = 1 + A/|a_n|$ és $r = 1/(1 + B/|a_0|)$ sugarak által meghatározott körgyűrűben vannak, ahol

$$A = \max\{|a_{n-1}|, \ldots, |a_0|\},\ B = \max\{|a_n|, \ldots, |a_1|\}.$$

Bizonyítás. Írjuk át a polinomot a

$$p(x) - a_{n-1}x^{n-1} - \ldots - a_1 x - a_0 = a_n x^n$$

alakba, és alkalmazzuk a háromszög-egyenlőtlenséget

$$|p(x)| + |a_{n-1}||x|^{n-1} + \ldots + |a_1||x| + |a_0| \geq |a_n||x|^n.$$

Ezt átrendezve kapjuk, hogy

$$\begin{aligned} |p(x)| &\geq |a_n||x|^n - |a_{n-1}||x|^{n-1} - \ldots - |a_1||x| - |a_0| \\ &\geq |a_n||x|^n - A(|x|^{n-1} + \ldots + |x| + 1) \\ &= |a_n||x|^n - A\frac{1-|x|^n}{1-|x|} = |x|^n \left(|a_n| - A\frac{1-|x|^n}{(1-|x|)|x|^n} \right). \end{aligned}$$

Ha most x_k olyan zérushely, melyre $|x_k| > 1$, akkor

$$0 = |p(x_k)| \geq |x_k|^n \left(|a_n| - A\frac{1-|x_k|^n}{(1-|x_k|)|x_k|^n} \right) \geq |x_k|^n \left(|a_n| + \frac{A}{1-|x_k|} \right).$$

Itt a második tényező nem lehet pozitív, és ebből az

$$|x_k| \leq 1 + \frac{A}{|a_n|} = R$$

egyenlőtlenséget kapjuk. Tehát a zérushelyek a komplex számsíkon az origó középpontó R sugarú körön kívül nem helyezkedhetnek el.

Az alsó becsléshez legyen $x = 1/y$. Ekkor a polinom

$$\begin{aligned} p(x) = p\left(\frac{1}{y}\right) &= a_n \left(\frac{1}{y}\right)^n + \ldots + a_1 \left(\frac{1}{y}\right) + a_0 \\ &= \frac{1}{y^n}(a_n + \ldots + a_1 y^{n-1} + a_0 y^n) \end{aligned}$$

alakban írható, ahol a zárójelben lévő polinomra alkalmazhatjuk a felső korlátra előbb bizonyított állítást. Mivel $a_0 \neq 0$, így a $p(x)$ polinomnak a nulla nem zérushelye. Ezért, ha x_k zérushelye $p(x)$-nek, akkor $y_k = 1/x_k$ zérushelye a zárójelben szereplő polinomnak, azaz

$$\frac{1}{|x_k|} = |y_k| \leq 1 + B/|a_0|.$$

Így

$$|x_k| \geq \frac{1}{1 + B/|a_0|} = r.$$

Ezt akartuk megmutatni. ■

5.1.3. példa. A $p(x) = x^5 - 4x^4 + x^3 - x^2 + 4x - 4$ polinom esetén $A = 4$ és $B = 4$, és így a polinom zérushelye az $1/2 \leq |x| \leq 5$ intervallumból kerül ki. ◇

A polinomok helyettesítési értékét a Horner[1]-módszerrel lehet gyorsan számolni. Ennek alapötlete a polinom

$$a_n x^n + \ldots + a_1 x + a_0 = (\ldots((a_n x + a_{n-1})x + a_{n-2})\ldots)x + a_0$$

alakú átírása. Ezzel az eredetileg szükséges n összeadás és $n(n+1)/2$ szorzás helyett csak n összeadás és n szorzás szükséges a helyettesítési érték meghatározásához. Ahogy azt az alábbi tételek mutatják, ezek a szorzási és összeadási számok optimálisak is.

5.1.4. tétel. (Ostrowski, 1954)
Egy n-edfokú polinom helyettesítési értékének kiszámításához legalább n összeadás kell.

5.1.5. tétel. (Victor Pan, 1966)
Egy n-edfokú polinom helyettesítési értékének kiszámításához legalább n szorzás kell.

5.1.2. Nemlineáris egyenletek megoldásának kondicionáltsága

Tekintsük az $f(x) = d$ egyenletet, melyről tegyük fel, hogy a feladat korrekt kitűzésű. Jelölje a megoldófüggvényt (f inverzét) G. Ekkor tehát $x = G(d)$. Ha a $G(d)$ megoldófüggvény differenciálható, akkor a feladat relatív és abszolút kondíciószáma

$$\kappa(d) = \left|\frac{G'(d)d}{G(d)}\right| = \left|\frac{d}{f'(G(d))G(d)}\right| = \left|\frac{d}{f'(x)x}\right|, \quad \kappa_{abs}(d) = \frac{1}{|f'(G(d))|} = \frac{1}{|f'(x)|}.$$

Ez mutatja, hogy ha $|f'(x)|$ kicsi, akkor rosszul kondicionált, ha nagy, akkor jól kondicionált a feladat. A rossz (jó) kondicionáltság azt jelenti, hogy d értékét kicsit megváltoztatva relatívan sokat (csak keveset) változhat a megoldás értéke.

[1] William George Horner (1786-1837, angol)

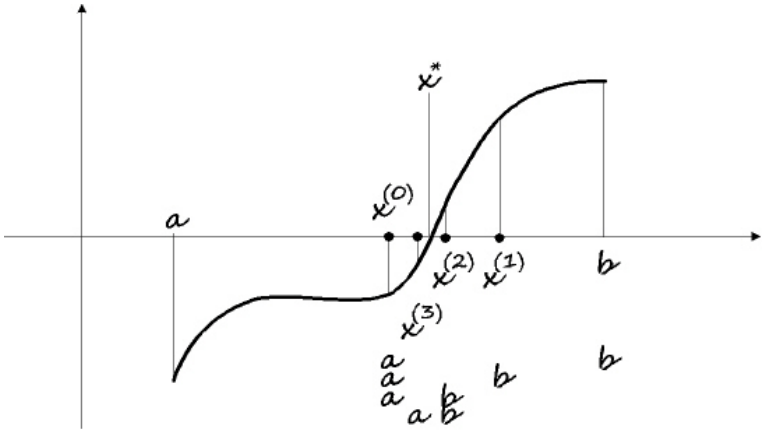
5.1.3. Geometriai módszerek

Tekintsük az $f(x) = 0$ egyenletet. Ennek egy megoldását jelölje $x^\star$, azaz $f(x^\star) = 0$. A nemlineáris egyenletek megoldási módszereinek nagy része azon alapul, hogy a megoldáshoz szemléletesen az f függvény grafikonjának és az x-tengelynek a metszéspontját kell megmondanunk. Így tehát egy olyan sorozatot kell előállítanunk, amely a metszéspont abszcisszájához tart. Ezeket a módszereket geometriai módszereknek nevezzük.

Intervallumfelezési módszer

Tegyük fel, hogy egy $f : [a, b] \to \mathbb{R}$ folytonos függvényre teljesül, hogy $f(a)f(b) < 0$. Ekkor nyilvánvalóan az f függvénynek legalább egy zérushelye van az $[a, b]$ intervallum belsejében. Itt feltesszük, hogy $f(a)$ negatív és $f(b)$ pozitív. Ha nem így lenne, akkor vizsgáljuk a $-f(x) = 0$ egyenletet, melynek nyilván ugyanazok a megoldásai, mint az eredeti egyenletnek.

Az ún. intervallumfelezési eljárás úgy keresi meg a grafikon és az x-tengely metszéspontját, hogy a felezőpont segítségével két részintervallumra bontja az eredeti $[a, b]$ intervallumot, majd ezen intervallumok végpontjaiban megvizsgálva a függvény előjeleit megállapítja, hogy melyik részintervallumba esik a metszéspont. Ezután ezzel az intervallummal végrehajtjuk ugyanezt az eljárást, stb. A módszert az 5.1.1. ábrán szemléltettük. Itt az $x^{(0)}$ pontban a függvényérték negatív, így az $x^{(0)}$-tól jobbra elhelyezkedő intervallum veszi át az eredeti intervallum szerepét a következő lépésben. Ennek $x^{(1)}$ felezőpontjában a függvényérték pozitív, így a tőle balra elhelyezkedő (az eredeti intervallum harmadik negyedelő intervalluma) intervallummal folytatjuk tovább az iterációt, stb. Az ábrán bejelöltük, hogy az egyes lépésekben mely pontok játsszák az a és b végpontok szerepét.



5.1.1. ábra: Az intervallumfelezési módszer szemléltetése.

Az alábbi algoritmus az intervallumfelezési eljárást hajtja végre egy olyan $f : [a, b] \to \mathbb{R}$ folytonos függvény esetén, melyre $f(a) < 0$ és $f(b) > 0$. Akkor áll le az algoritmus, ha az iterációszám eléri a $k_{\max}$ értéket, vagy az intervallumhossz *toll* alá csökken.

Intervallumfelezési módszer, a, b $(a < b)$, *toll*, $k = 0$, $k_{\max}$ ill. f adott, $f(a) < 0 < f(b)$.
while $k < k_{\max}$ **and** $b - a > toll$ **do**
  $x := a + (b - a)/2$

```
    f := f(x)
    if f = 0 then
      end
    else
      if f > 0 then
        b = x
      else
        a = x
      end if
    end if
  end while
```

A konstrukcióból nyilvánvaló, hogy az eljárás által generált sorozat tartani fog az $[a, b]$ intervallum valamelyik zérushelyéhez. Ha több zérushely van, akkor valamelyiket biztosan megtalálja az eljárás.

Azt is fontos észrevenni, hogy a generált sorozat nem monoton módon tart általában a zérushelyhez. Az 5.1.1. ábrán például az $x^{(0)}$ közelítés közelebb van a zérushelyhez, mint $x^{(1)}$. Emiatt az 1.3.1. definíció értelmében nem definiálható a módszer konvergenciarendje. Jelölje $\tilde{e}_k$ a k. közelítés hibájának egy abszolút értékben vett felső becslését. A $\tilde{e}^{(0)} = (b-a)/2$ nyilvánvalóan jó választás, és így $\tilde{e}^{(k)} = (b-a)/2^{k+1}$. Bár a sorozat konvergenciarendjét nem tudtuk definiálni, de az látszik, hogy a hiba felső becsléseinek sorozata elsőrendben konvergens. Mivel $\tilde{e}^{(k+1)} = \tilde{e}^{(k)}/2$, így a hibabecslésünk kb. három iterációs lépésenként javul egy nagyságrendet.

Az

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

becslést ($e^{(k)} = x^{(k)} - x^\star$) használhatjuk a leállási feltételhez is. A jobb oldali kifejezésből meg tudjuk mondani, hogy ha az iterációszám már nagyobb, mint $k_{\max} = \frac{\ln((b-a)/\varepsilon)}{\ln 2} - 1$, akkor a sorozatelem ε-nál jobban megközelítette a zérushelyet.

5.1.6. megjegyzés. Más leállási feltétel lehet még a módszerhez az, hogy akkor állítjuk le az iterációt, ha

$$\frac{|x^{(k)} - x^{(k-1)}|}{|x^{(k-1)}|} \leq \text{toleranciaszint}$$

vagy

$$|f(x^{(k)})| \leq \text{toleranciaszint}.$$

◇

5.1.7. megjegyzés. Az algoritmusban a felezőpontot a $x := a+(b-a)/2$, és nem az $x := (a+b)/2$ módon számoltuk. Ennek oka az, hogy az első módszer lebegőpontos számokkal számolva mindig a és b közé eső értéket ad, míg a másik nem. Kétjegyű mantisszával dolgozva pl. $(0.67+0.69)/2 = 1.36/2 \approx 1.4/2 = 0.7$, ami nem esik a két szám közé. Viszont $0.67 + (0.69 - 0.67)/2 = 0.67 + 0.02/2 = 0.67 + 0.01 = 0.68$. ◇

Húrmódszer

A most tárgyalandó húrmódszer és a későbbi szelő- ill. Newton-módszerek az alábbi általános megfontoláson alapulnak. Tegyük fel, hogy az f differenciálható függvény $x^\star$ zérushelyéhez tartó

sorozatot szeretnénk előállítani. Tegyük fel, hogy a sorozat k-adik eleme $x^{(k)}$, és hogy $f'(x)$ egyik pontban sem nulla az $x^{(k)}$ és $x^\star$ pontok között. Fejtsük sorba a függvényt az $x^\star$ megoldás körül!

$$f(x^{(k)}) = \overbrace{f(x^\star)}^{=0} + f'(\xi_k)(x^{(k)} - x^\star),$$

ahol ξ_k az $x^{(k)}$ és $x^\star$ pontok közé esik. Innét

$$x^\star = x^{(k)} - \frac{f(x^{(k)})}{f'(\xi_k)}.$$

Azaz ha tudnánk ξ_k értékét, akkor a következő iterációs lépésben egyszerre a zérushelyre tudna lépni a sorozat. Mivel ezt általában nem tudjuk, ezért $f'(\xi_k)$ értékét minden lépésben közelítjük. Legyen q_k a k-adik lépésben az $f'(\xi_k)$ derivált közelítése. Ekkor az

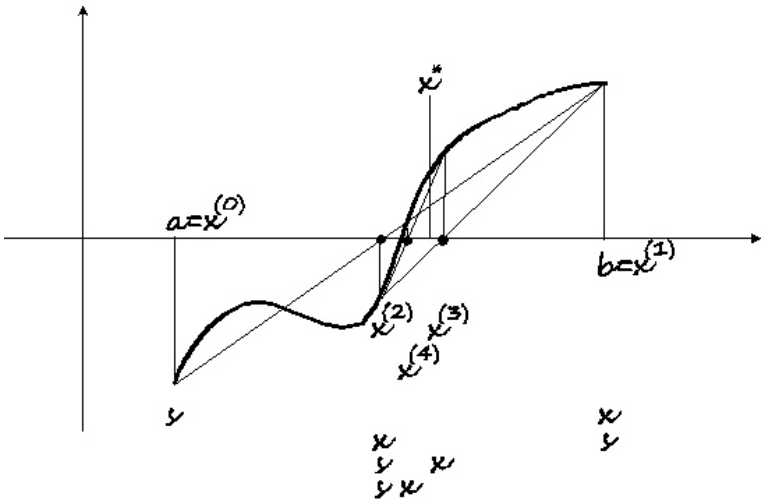
$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{q_k} = x^{(k)} - \frac{1}{q_k} f(x^{(k)})$$

iterációkat lehet használni. A különböző tárgyalandó módszerek abban különböznek, hogy hogyan határozzuk meg a q_k közelítéseket. Vegyük még észre, hogy geometriailag $x^{(k+1)}$ az $(x^{(k)}, f(x^{(k)}))$ ponton átmenő q_k meredekségű egyenes x-tengellyel alkotott metszéspontja lesz.

A *húrmódszer* esetén a q_k értéket az adott k-adik sorozatelemhez tartozó grafikonpontot és a hozzá legközelebbi korábbi ellenkező előjelű függvényértéket adó sorozatelemhez tartozó grafikonpontot összekötő húr meredekségének választjuk. A módszert az 5.1.2. ábrán szemléltettük. Az ábrán az $x^{(0)}$ és $x^{(1)}$ értékek a zérushelye(ke)t tartalmazó intervallum két végpontja. Az $f(x^{(0)})f(x^{(1)}) < 0$ feltétel biztosítja, hogy valóban van zérushely az adott intervallumon. Az $x^{(2)}$ sorozatelemet az $x^{(0)}$ és $x^{(1)}$ pontokhoz tartozó grafikonpontokat összekötő húr x-tengellyel alkotott metszéspontja adja. Ez nyilván $x^{(0)}$ és $x^{(1)}$ közé esik. Mivel $f(x^{(2)}) < 0$ és $f(x^{(1)}) > 0$ ezért a következő közelítést az ezen két ponthoz tartozó grafikonpontokat összekötő húr x-tengellyel alkotott metszéspontja adja. Így kapjuk $x^{(3)}$-at, ami $x^{(2)}$ és $x^{(1)}$ közé esik. Mivel $f(x^{(3)}) > 0$ és $f(x^{(2)}) < 0$, azért a következő közelítést az ezen két ponthoz tartozó grafikonpontokat összekötő húr x-tengellyel alkotott metszéspontja adja. Így kapjuk $x^{(4)}$-et, ami $x^{(2)}$ és $x^{(3)}$ közé esik. Mivel $f(x^{(4)})$ is pozitív, így a következő lépésben az $x^{(4)}$ és újra az $x^{(2)}$ ponthoz tartozó grafikonpontokat összekötő húr x-tengellyel alkotott metszéspontja adja $x^{(5)}$-öt. Stb. Az ábrán x-szel jelöltük az aktuális iterációs lépést, és s-sel azt a legközelebbi korábbi iterációs lépést, mellyel $f(s)f(x) < 0$.

Az algoritmus az alábbi módon foglalható össze.

Húrmódszer, a, b ill. f adott, $f(a)f(b) < 0$.
$x := b$, $s := a$
for $k := 1 : k_{\max}$ **do**
  $xnew := x - \frac{x-s}{f(x)-f(s)} f(x)$ $(q_k = \frac{f(x)-f(s)}{x-s})$
  if $f(xnew) = 0$ **then**
    end
  else
    if $f(xnew)f(x) < 0$ **then**
      $s = x$
    end if
  end if
  $x := xnew$
end for



5.1.2. ábra: A húrmódszer iterációjának szemléltetése.

A módszer konstrukciójából következik, hogy a kiindulási intervallumban elhelyezkedő zérushelyek egyikét biztosan meg fogja találni (csakúgy, mint az intervallumfelezési módszer).

Több tételben hasonló feltevéseket kell majd tennünk, ezért a tételek megfogalmazásának megkönnyítésére bevezetjük az alábbi definíciót.

5.1.8. definíció.

Azt mondjuk, hogy az f függvény *kielégíti az alapfeltevéseket* az $[a,b]$ intervallumon, ha van $[a,b]$ belsejében zérushelye, legalább kétszer folytonosan deriválható, és sem az első, sem a második deriváltja nem vesz fel nulla értéket $[a,b]$-n.

A fenti feltételekből következik, hogy f első és második deriváltjának előjele jól meghatározott az intervallumon, és hogy vannak olyan m_1, m_2, M_1 és M_2 pozitív konstansok, melyekkel $0 < m_1 \le |f'(x)| \le M_1 < \infty$ és $0 < m_2 \le |f''(x)| \le M_2 < \infty$ minden adott intervallumbeli x értékre. Figyeljük meg, hogy m az alsó becsléseket, M a felső becsléseket jelöli, az index pedig azt adja meg, hogy a becslés hányadik deriváltra vonatkozik. A szigorú monotonitásból következik, hogy csak egy zérushely lehet az $[a,b]$ intervallum belsejében. A zérushely jelölésére, ahogy korábban, $x^\star$-t fogunk használni.

Ezután megfogalmazzuk a húmódszerre vonatkozó konvergenciatételt.

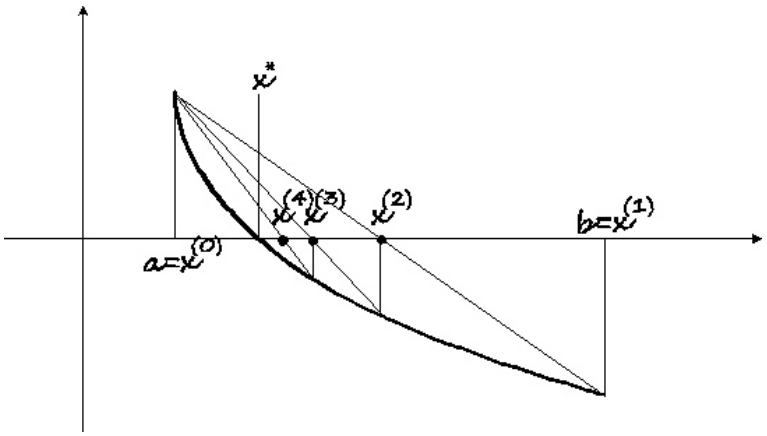
5.1.9. tétel.

Elégítse ki f az 5.1.8. definícióbeli alapfeltevéseket az $[a,b]$ intervallumon. Ekkor a húrmódszerrel előállított sorozat az $f(x) = 0$ egyenlet $x^\star$ megoldásához konvergál, a konvergencia elsőrendű, és

$$|e^{(k+1)}| \le C|e^{(k)}|, \tag{5.1.1}$$

ahol $C = |e^{(0)}|M_2/(2m_1)$.

Bizonyítás. A deriváltak előjele szempontjából négy különböző eset lehetséges csak. Vizsgáljuk csak azt az esetet, amikor $f' < 0$ és $f'' > 0$ (5.1.3. ábra). A másik három eset hasonlóan igazolható.



5.1.3. ábra: A húrmódszer iterációja, ha $f' < 0$ és $f'' > 0$.

A szigorú konvexitás ($f'' > 0$) miatt $\{x^{(k)}\}$ szigorúan monoton csökkenő ($s = x^{(0)}$ $(k \geq 1)$), továbbá minden sorozatelemre $x^{(k)} \geq x^\star$ $(k \geq 1)$. Így a sorozat konvergens. Legyen a határértéke y. Az

$$\underbrace{x^{(k+1)}}_{\to y} = \underbrace{x^{(k)}}_{\to y} - \frac{\overbrace{x^{(k)}}^{\to y} - x^{(0)}}{\underbrace{f(x^{(k)})}_{\to f(y)} - f(x^{(0)})} \underbrace{f(x^{(k)})}_{\to f(y)}$$

iteráció vizsgálatával látható, hogy a határértéknek az

$$y = y - \frac{y - x^{(0)}}{f(y) - f(x^{(0)})} f(y) = y - \frac{1}{f'(\eta)} f(y)$$

egyenletet kell kielégítenie, ahol η $x^{(0)}$ és y közé esik (Lagrange-középértéktétel). Mivel $f'(\eta)$ negatív, ezért az egyenlőség csak úgy teljesülhet, ha $f(y) = 0$, azaz $y = x^\star$.

A konvergenciarend megállapításához ki kell fejeznünk a $(k+1)$-edik közelítés hibáját a k-adik közelítés hibájával. A $(k+1)$-edik közelítés hibája

$$x^{(k+1)} - x^\star = x^{(k)} \frac{f(x^{(k)}) - f(x^{(0)})}{f(x^{(k)}) - f(x^{(0)})}$$

$$- \frac{x^{(k)} - x^{(0)}}{f(x^{(k)}) - f(x^{(0)})} f(x^{(k)}) - x^\star \frac{f(x^{(k)}) - f(x^{(0)})}{f(x^{(k)}) - f(x^{(0)})}$$

$$= \frac{f(x^{(0)})(x^\star - x^{(k)}) - f(x^{(k)})(x^\star - x^{(0)})}{f(x^{(k)}) - f(x^{(0)})}$$

alakban írható. Ezután a nulladik és k-adik közelítés hibájával osztva kapjuk, hogy

$$\frac{e^{(k+1)}}{e^{(k)} e^{(0)}} = \frac{-f(x^{(0)})/e^{(0)} + f(x^{(k)})/e^{(k)}}{f(x^{(k)}) - f(x^{(0)})}$$

$$= \frac{f(x^{(k)})/e^{(k)} - f(x^{(0)})/e^{(0)}}{f(x^{(k)}) - f(x^{(0)})} \underbrace{=}_{\text{Cauchy-k.é.t.}} \frac{f'(\xi_k)(\xi_k - x^\star) - f(\xi_k)}{(\xi_k - x^\star)^2 f'(\xi_k)},$$

ahol ξ_k $x^{(0)}$ és $x^{(k)}$ közé eső megfelelő érték a Cauchy-középértéktétel szerint. A számlálóban lévő kifejezést az f függvény ξ_k pont körül felírt és $x^\star$ pontban vett

$$0 = f(x^\star) = f(\xi_k) + f'(\xi_k)(x^\star - \xi_k) + \frac{f''(\eta_k)}{2}(x^\star - \xi_k)^2$$

Taylor-polinomjának értékéből fejezhetjük ki, ahol η_k ξ_k és $x^\star$ közé esik. Így

$$\frac{e^{(k+1)}}{e^{(k)}e^{(0)}} = \frac{f'(\xi_k)(\xi_k - x^\star) - f(\xi_k)}{(\xi_k - x^\star)^2 f'(\xi_k)}$$

$$= \frac{f'(\xi_k)(\xi_k - x^\star) - (-f'(\xi_k)(x^\star - \xi_k) - f''(\eta_k)(x^\star - \xi_k)^2/2)}{(\xi_k - x^\star)^2 f'(\xi_k)} = \frac{f''(\eta_k)}{2f'(\xi_k)}.$$

Azaz

$$|e^{(k+1)}| = \frac{|f''(\eta_k)|}{2|f'(\xi_k)|}|e^{(k)}||e^{(0)}|.$$

A sorozat szigorúan monoton csökkenése és a tétel deriváltakra vonatkozó feltételei miatt

$$0 < \frac{m_2}{2M_1}|e^{(0)}| \le \frac{|f''(\eta_k)|}{2|f'(\xi_k)|}|e^{(0)}| < 1 \ (k \ge 1),$$

amiből az 1.3.3. tételre tekintettel következik az elsőrendű konvergencia. Továbbá

$$|e^{(k+1)}| = \frac{|f''(\eta_k)|}{2|f'(\xi_k)|}|e^{(0)}||e^{(k)}| \le \underbrace{|e^{(0)}|\frac{M_2}{2m_1}}_{=:C}|e^{(k)}| = C|e^{(k)}|$$

a $C = |e^{(0)}|M_2/(2m_1)$ jelölés bevezetésével. Ezzel a tételt igazoltuk. ■

5.1.10. megjegyzés. Az (5.1.1) becslés nyilvánvalóan csak akkor használható leállási feltétel megadására, ha $C < 1$. Ez elérhető úgy, ha $x^{(0)}$-t elegendően közel választjuk a zérushelyhez. ◇

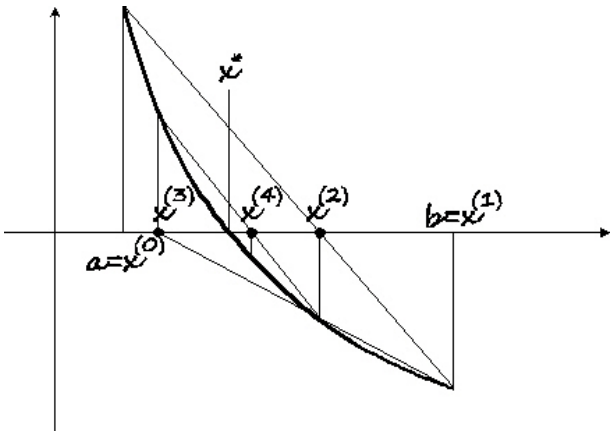
Szelőmódszer

A szelőmódszer annyiban különbözik a húrmódszertől, hogy itt nem figyelünk arra, hogy mindig két ellenkező előjelű függvényértéket adó grafikonpont legyen összekötve. Egyszerűen két egymás utáni sorozatelemhez tartozó grafikonponton át húzunk szelőt, és ahol az metszi (ha van metszéspont) az x-tengelyt, az lesz a következő sorozatelem. Ezen módszernek veszélye, hogy nem mindig találja meg a kívánt zérushelyet, cserébe viszont magasabbrendű konvergenciát kapunk. A szelőmódszer esetén a ξ_k pontbeli első deriváltra tehát a

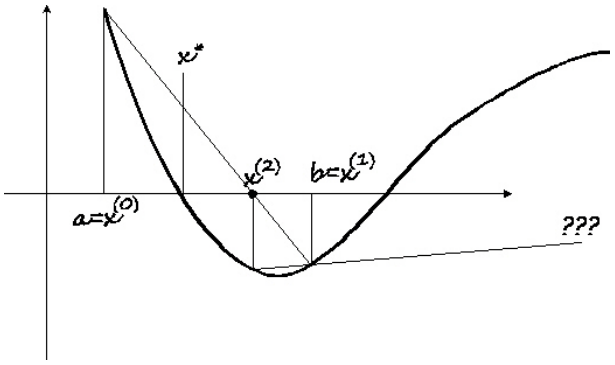
$$q_k = \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}$$

közelítést használjuk.

Az 5.1.4. ábrán szemléltettük a szelőmódszer iterációját. Ha az $[a, b]$ intervallumban van zérushely, akkor választhatjuk a-t $x^{(0)}$-nak és b-t $x^{(1)}$-nek. A két értékhez tartozó grafikonponton át berajzolt szelő x-tengellyel alkotott metszéspontja lesz $x^{(2)}$. Ezután az $x^{(1)}$ és $x^{(2)}$ pontokhoz tartozó grafikonpontokat kötjük össze. Az összekötő szelő x-tengellyel alkotott metszéspontja lesz $x^{(3)}$. (Megjegyezzük, hogy a húrmódszer esetén ebben a lépésben az $x^{(0)}$ és $x^{(2)}$ pontokhoz tartozó grafikonpontokat kötnénk össze.) Ezután $x^{(2)}$-vel és $x^{(3)}$-mal tesszük ugyanezt. Stb.



5.1.4. ábra: A szelőmódszer iterációjának szemléltetése.



5.1.5. ábra: A szelőmódszer nem mindig találja meg az intervallumban lévő zérushelyet.

Arra, hogy a szelőmódszer nem mindig találja meg az intervallumban lévő zérushelyet, az 5.1.5. ábrán adtunk példát. Az $x^{(2)}$ és $x^{(1)}$ pontokhoz tartozó szelő már nem az $[a, b]$ intervallumban metszi az x-tengelyt, hanem tőle távol.

A szelőmódszer iterációs lépéseit az alábbi algoritmus tartalmazza.

Szelőmódszer, a, b ill. f adott, $f(a)f(b) < 0$.
$x := b$, $s := a$
for $k := 1 : k_{\max}$ **do**
  $xnew := x - \frac{x-s}{f(x)-f(s)} f(x)$ $(q_k = \frac{f(x)-f(s)}{x-s})$
  if $f(xnew) = 0$ **then**
    end
  else
    $s := x$, $x := xnew$
  end if
end for

5.1.11. tétel.
Teljesítse az f függvény az 5.1.8. definícióbeli alapfeltevéseket az $[a,b]$ intervallumon. Ekkor, ha $\max\{|a-x^\star|, |b-x^\star|\} < 2m_1/M_2$, akkor a szelőmódszerrel előállított sorozat monoton módon $x^\star$-hoz tart, és a konvergencia rendje $(1+\sqrt{5})/2 \approx 1.618$. Továbbá érvényes az

$$|e^{(k+1)}| \leq C|e^{(k)}||e^{(k-1)}|$$

becslés a $C = M_2/(2m_1)$ választással.

Bizonyítás. Igazoljuk először, hogy ha az $x^{(k-1)}$ és $x^{(k)}$ sorozatelemek közelebb vannak $x^\star$-hoz, mint $2m_1/M_2$, akkor az $x^{(k+1)}$ elem is közelebb lesz. Így mivel $x^{(0)}$-ra és $x^{(1)}$-re is teljesül a feltétel, minden sorozatelem közelebb lesz, mint $2m_1/M_2$ (indukció).

Számítsuk ki a $(k+1)$-edik iterációs lépés hibáját! (Ez ugyanaz a számolás lesz, mint a húrmódszer esetén, csak $x^{(0)}$ helyett $x^{(k-1)}$-et kell írni.)

$$x^{(k+1)} - x^\star = x^{(k)} \frac{f(x^{(k)}) - f(x^{(k-1)})}{f(x^{(k)}) - f(x^{(k-1)})}$$

$$- \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k)}) - x^\star \frac{f(x^{(k)}) - f(x^{(k-1)})}{f(x^{(k)}) - f(x^{(k-1)})}$$

$$= \frac{f(x^{(k-1)})(x^\star - x^{(k)}) - f(x^{(k)})(x^\star - x^{(k-1)})}{f(x^{(k)}) - f(x^{(k-1)})}.$$

Ezután a $(k-1)$-edik és k-adik közelítés hibájával osztva kapjuk, hogy

$$\frac{e^{(k+1)}}{e^{(k)}e^{(k-1)}} = \frac{-f(x^{(k-1)})/e^{(k-1)} + f(x^{(k)})/e^{(k)}}{f(x^{(k)}) - f(x^{(k-1)})}$$

$$= \frac{f(x^{(k)})/e^{(k)} - f(x^{(k-1)})/e^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} \underbrace{=}_{\text{Cauchy-k.é.t.}} \frac{f'(\xi_k)(\xi_k - x^\star) - f(\xi_k)}{(\xi_k - x^\star)^2 f'(\xi_k)},$$

ahol ξ_k $x^{(k-1)}$ és $x^{(k)}$ közé eső megfelelő érték a Cauchy-középértéktétel szerint. A számlálóban lévő kifejezést az f függvény ξ_k pont körüli, $x^\star$ pontban vett

$$0 = f(x^\star) = f(\xi_k) + f'(\xi_k)(x^\star - \xi_k) + \frac{f''(\eta)}{2}(x^\star - \xi_k)^2$$

Taylor-polinomjának értékéből fejezhetjük ki, ahol η_k ξ_k és $x^\star$ közé esik. Így

$$\frac{e^{(k+1)}}{e^{(k)}e^{(k-1)}} = \frac{f'(\xi_k)(\xi_k - x^\star) - f(\xi_k)}{(\xi_k - x^\star)^2 f'(\xi_k)}$$

$$= \frac{f'(\xi_k)(\xi_k - x^\star) - (-f'(\xi_k)(x^\star - \xi_k) - f''(\eta_k)(x^\star - \xi_k)^2/2)}{(\xi_k - x^\star)^2 f'(\xi_k)} = \frac{f''(\eta_k)}{2f'(\xi_k)}.$$

Ebből adódik tehát, hogy

$$|e^{(k+1)}| \leq C|e^{(k)}||e^{(k-1)}|, \tag{5.1.2}$$

ahol $C = M_2/(2m_1)$.

Mivel a feltétel miatt $|e^{(k-1)}| < 2m_1/M_2$, ezért kapjuk, hogy

$$|e^{(k+1)}| < |e^{(k)}|,$$

azaz a hibasorozat monoton csökkenő. Így az $|e^{(k)}| < 2m_1/M_2$ feltétel miatt $|e^{(k+1)}| < 2m_1/M_2$ is teljesül.

Most megmutatjuk, hogy a hibasorozat nullához tart. Vezessük be a $d := C\max\{|a-x^\star|, |b-x^\star|\}$ jelölést. Nyilván $0 < d < 1$. Az (5.1.2) becslés miatt

$$\begin{aligned}
|e^{(0)}| &\le d/C,\\
|e^{(1)}| &\le d/C,\\
|e^{(2)}| &\le C|e^{(1)}||e^{(0)}| < d^2/C,\\
|e^{(3)}| &\le C|e^{(2)}||e^{(1)}| < d^3/C,\\
|e^{(4)}| &\le C|e^{(3)}||e^{(2)}| < d^5/C,\\
&\vdots\\
|e^{(k)}| &\le C|e^{(k-1)}||e^{(k-2)}| < d^{u_{k+1}}/C,
\end{aligned}$$

ahol u_{k+1} a Fibonacci-sorozat $k+1$-edik eleme. Mivel $0 < d < 1$, és a Fibonacci-sorozat végtelenhez tart, a fenti becslések mutatják a sorozat konvergenciáját.

Most már csak a konvergenciarend megállapítása van hátra. Jelöljük C_k-val az $f''(\eta_k)/(2f'(\xi_k))$ hányadost. Ezzel a jelöléssel a hibaegyenlet az

$$e^{(k+1)} = C_k e^{(k)} e^{(k-1)}$$

alakban írható. Vegyük mindkét oldal abszolútértékét, majd logaritmusát (tegyük fel, hogy k olyan nagy, hogy a hiba már egynél kisebb lesz)

$$\ln|e^{(k+1)}| = \ln|C_k| + \ln|e^{(k)}| + \ln|e^{(k-1)}|.$$

Osszuk el mindkét oldalt $\ln|e^{(k)}|$-val!

$$\frac{\ln|e^{(k+1)}|}{\ln|e^{(k)}|} = \frac{\ln|C_k|}{\ln|e^{(k)}|} + 1 + \frac{\ln|e^{(k-1)}|}{\ln|e^{(k)}|}.$$

A konvergenciarend a bal oldalon álló logaritmikus relatív csökkenés határértéke. Jelölje ezt r. A jobb oldali első tag határértéke 0, hiszen mivel $x^{(k)} \to x^\star$, ha $k \to \infty$, és a deriváltak folytonosak, ezért C_k-nak is van nullától különböző határértéke, így korlátos, a nevező pedig $-\infty$-hez tart. A második tag egy, a harmadik pedig nyilvánvalóan $1/r$-hez tart. Azaz a határérték egyértelműsége miatt r-nek ki kell elégítenie az $r = 1 + 1/r$ egyenletet, melynek egynél nagyobb megoldása $r = (1+\sqrt{5})/2$. Tehát a konvergenciarend valóban $(1+\sqrt{5})/2$. ■

Newton-módszer

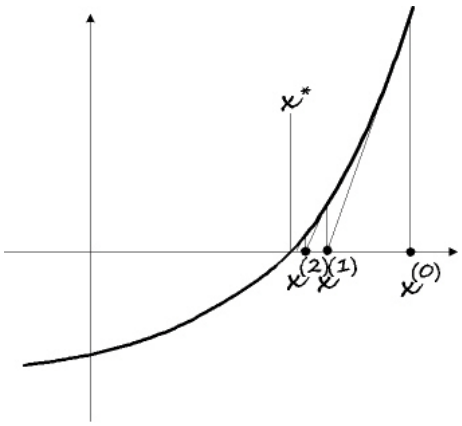
A most ismertetendő módszer első változatát Newton írta le 1671-ben Method of Fluxions című művében, de csak 1736-ban publikálta. Egy letisztultabb változata 1690-ben jelent meg Joseph Raphson[2] Analysis Aequationum című cikkében. A módszer mai alakja Thomas Simpson-tól származik. Róla a Numerikus integrálás fejezetben lesz majd még szó. A módszert gyakran hívják Newton–Raphson-módszernek is, de mi a továbbiakban *Newton-módszer*nek fogjuk nevezni.

A módszer alapötlete az, hogy az első derivált ξ_k pontbeli értékét az $x^{(k)}$ pontbeli érintő meredekségével közelítjük, azaz $q_k = f'(x^{(k)})$. Az iteráció az alábbi algoritmussal írható le.

[2] Joseph Raphson, 1648–1715. Bővebb életrajz található a http://numericalmethods.eng.usf.edu/anecdotes/raphson.html oldalon.

Newton-módszer, $x^{(0)}$ ill. f adott.
$x := x^{(0)}$
for $k := 1 : k_{\max}$ **do**
  $x := x - \frac{1}{f'(x)} f(x)$ $(q_k = f'(x))$
  if $f(x) = 0$ **then**
    end
  end if
end for

A Newton-módszer iterációs lépéseit az 5.1.6. ábrán szemléltettük. Az $x^{(0)}$ pontbeli grafikonpontban behúzzuk a grafikon érintőjét. Ahol ez metszi az x-tengelyt, az lesz az $x^{(1)}$ közelítés. Ezután az ehhez a ponthoz tartozó grafikonpontban húzunk érintőt. Ahol ez metszi az x-tengelyt, az lesz az $x^{(2)}$ közelítés, stb.



5.1.6. ábra: A Newton-módszer iterációjának szemléltetése.

Fontos különbség a korábbi módszerekhez képest, hogy míg korábban minden lépésben csak egyszer kellett kiszámolni az f függvény függvényértékét, addig a Newton-módszernél két függvényértéket kell számolni minden lépésben: az f és f' függvényekét. További különbség még, hogy a Newton-módszer végrehajtásához az f függvény deriváltjára is szükség van. Természetesen a Newton-módszer fenti "nehézségei" csak addig lassították az iterációs eljárás végrehajtását, amíg a számítógépek meg nem jelentek. Ezekért a nehézségekért a módszer másodrendű konvergenciát ad cserébe.

5.1.12. tétel.

Teljesítse az f függvény az 5.1.8. definícióbeli alapfeltevéseket az $[a, b]$ intervallumon. Az f függvény zérushelyének meghatározására alkalmazzuk a Newton-módszert olyan $x^{(0)}$ pontból indítva, melyre $e^{(0)} < \min\{|a - x^\star|, |b - x^\star|, 2m_1/M_2\} =: h_0$. Ekkor a módszer által előállított $\{x^k\}$ sorozat másodrendben és monoton módon konvergál a határértékhez, továbbá érvényes az

$$|e^{(k+1)}| \leq C|e^{(k)}|^2 \tag{5.1.3}$$

becslés a $C = M_2/(2m_1)$ választással.

Bizonyítás. Tegyük fel, hogy a sorozat $x^{(k)}$ eleme az $x^\star$ zérushely h_0 sugarú környezetébe esik. Ekkor f $x^{(k)}$-ban sorbafejthető. Számoljuk ki a közelítő polinom értékét $x^\star$-ban!

$$0 = f(x^\star) = f(x^{(k)}) + f'(x^{(k)})(x^\star - x^{(k)}) + \frac{f''(\xi_k)}{2}(x^\star - x^{(k)})^2,$$

ahol ξ_k $x^\star$ és $x^{(k)}$ közé esik. Innét

$$-\frac{f(x^{(k)})}{f'(x^{(k)})} = (x^\star - x^{(k)}) + \frac{f''(\xi_k)}{2f'(x^{(k)})}(x^\star - x^{(k)})^2,$$

amit a Newton-módszer $(k+1)$-edik lépésének

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}$$

képletébe helyettesítünk.

$$\begin{aligned} x^{(k+1)} - x^\star &= x^{(k)} - x^\star - \frac{f(x^{(k)})}{f'(x^{(k)})} \\ &= x^{(k)} - x^\star + (x^\star - x^{(k)}) + \frac{f''(\xi_k)}{2f'(x^{(k)})}(x^\star - x^{(k)})^2 \\ &= \frac{f''(\xi_k)}{2f'(x^{(k)})}(x^\star - x^{(k)})^2. \end{aligned} \tag{5.1.4}$$

A

$$0 < \frac{m_2}{2M_1} \le \frac{f''(\xi_k)}{2f'(x^{(k)})} \le \frac{M_2}{2m_1}$$

egyenlőtlenségből és a

$$\frac{M_2}{2m_1}|x^{(0)} - x^\star| < \frac{M_2}{2m_1}h_0 \le 1$$

becslésből következik a sorozat másodrendű konvergenciája $x^\star$-hoz (1.3.5. tétel).

Az (5.1.4) egyenlőség abszolútértékét véve és felső becslést végezve

$$|x^{(k+1)} - x^\star| = \left|\frac{f''(\xi_k)}{2f'(x^{(k)})}\right| |x^\star - x^{(k)}|^2 \le \frac{M_2}{2m_1}h_0|x^\star - x^{(k)}| < |x^\star - x^{(k)}|,$$

amiből következik a tételbeli (5.1.3) becslés és a monoton konvergencia. Mivel $x^{(0)}$ beleesett $x^\star$ h_0 sugarú környezetébe, így a többi sorozatelem is ebbe a környezetbe fog esni (teljes indukció). Ez mutatja, hogy jogosan tettük fel a tétel elején, hogy $x^{(k)}$-ban sorbafejthető f. ■

Az előző tétel feltételeit általában nehéz ellenőrizni. Most megadunk egy könnyebben ellenőrizhető feltételt, amely ráadásul szigorúan monoton sorozat határértékeként állítja elő az egyenlet megoldását.

5.1.13. tétel.

Tegyük fel, hogy az $f \in C^2[a,b]$ függvény első és második deriváltja sem vesz fel nulla értéket az $x^\star$ zérushely és az $x^{(0)}$ kezdőpontok által meghatározott intervallumon, és $f(x^{(0)})f''(x^{(0)}) > 0$. Ekkor a Newton-módszer által generált $\{x^{(k)}\}$ sorozat szigorúan monoton sorozat lesz, és tart $x^\star$-hoz.

Bizonyítás. Legyen például $x^{(0)} > x^\star$ és $f(x^{(0)}) > 0, f''(x^{(0)}) > 0$ (ekkor $f'(x) > 0$ $x^{(0)}$ és a $x^\star$ zérushely között). Az

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}$$

iterációs képletből látható, hogy $x^{(k+1)} < x^{(k)}$, azaz szigorúan monoton csökkenő lesz a sorozat (amíg a zérushelyet meg nem találja). Másrészt a szigorú konvexitásból következik, hogy $x^{(k)} \geq x^\star$. Így a sorozat konvergens lesz. Legyen a határértéke $\bar{x}^\star$. Ekkor az iterációs képlet mindkét oldalán álló sorozatnak ugyanoda kell tartania

$$\underbrace{x^{(k+1)}}_{\to \bar{x}^\star} = \underbrace{x^{(k)}}_{\to \bar{x}^\star} - \frac{\overbrace{f(x^{(k)})}^{\to f(\bar{x}^\star)}}{\underbrace{f'(x^{(k)})}_{\to f'(\bar{x}^\star)}},$$

amiből következik, hogy $\bar{x}^\star = x^\star$. ■

5.2. Fixpont-iterációk

A fixpont-iterációs egyenletmegoldás a Banach-féle fixponttételt (1.1.18. tétel) alkalmazva határozza meg egy nemlineáris egyenlet megoldását. Valós függvényekre a tétel azt mondja ki, hogy ha egy $F : [a, b] \to [a, b]$ függvény kontrakció a $0 \leq q < 1$ kontrakciós tényezővel, akkor bármilyen $x^{(0)} \in [a, b]$ kezdőelemre az

$$x^{(k)} = F(x^{(k-1)})$$

iterációval előállított sorozat tart az F függvény egyetlen $[a, b]$-beli $x^\star$ fixpontjához, továbbá

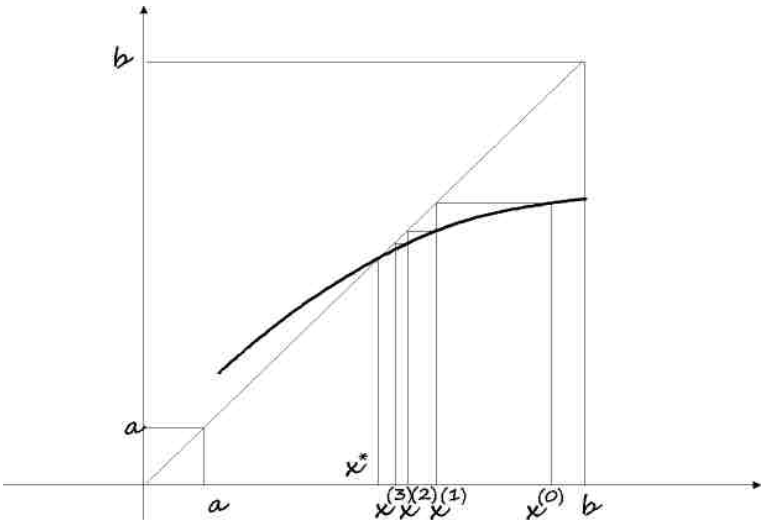
$$|x^{(k)} - x^\star| \leq \frac{q^k}{1-q}|x^{(1)} - x^{(0)}|. \tag{5.2.1}$$

A fixpont-iteráció lépései szemléltethetők az $y = F(x)$ és az $y = x$ függvények grafikonjának segítségével. Az $x^{(0)}$ értékkel kiszámítjuk az $F(x^{(0)})$ értéket, majd ezt az $y = x$ függvény segítségével az x-tengelyre vetítjük. Ezután $x^{(1)}$-gyel járunk el hasonlóan, stb (5.2.1. ábra).

Az (5.2.1) becslés használható leállási feltételként, hiszen meg tudjuk becsülni az első két elem segítségével, hogy hányadik sorozatelemtől leszünk közelebb a határértékhez, mint egy adott hibaszint.

Ha egy $f(x) = 0$ egyenlet megoldását szeretnénk megkeresni (f folytonos függvény), akkor először a feladatot átírjuk egy ekvivalens $F(x) = x$ fixpontproblémává, ahol most már az $F(x)$ függvény fixpontjának megkeresése a feladat. Az $F(x)$ függvényt többféleképpen előállíthatjuk. A legegyszerűbb az, ha ki tudjuk fejezni $f(x)$-ből az x változót. Példa lehet erre mondjuk az $x^2 - 2 = 0$ egyenlet megoldása, ahol $f(x) = x^2 - 2$, és x-et kifejezve kapjuk, hogy $x = 2/x$, vagyis az $F(x) = 2/x$ választás megfelelő. Egy általánosabb módszer fixpontprobléma előállítására az, ha egy tetszőleges, sehol sem nulla folytonos g függvénnyel képezzük az $F(x) = x - g(x) \cdot f(x)$ függvényt. Mivel g nem vehet fel nulla értéket, $x = F(x)$ pontosan olyan x-re teljesülhet csak, amire $f(x) = 0$.

Természetesen nem elegendő csak a megoldandó egyenlet fixpontproblémává transzformálása a megoldáshoz, hiszen az $F(x)$ függvénynek ki kell elégítenie az Banach-féle fixponttétel feltételeit. A gyakorlatban a kontrakciós tulajdonságot garantálja, ha F folytonos $[a, b]$-n és differenciálható (a, b)-ben, továbbá van olyan $0 \leq q < 1$ szám, mellyel $|F'(x)| \leq q$, $\forall x \in (a, b)$ (Lagrange-középértéktétel).



5.2.1. ábra: A fixpont-iteráció szemléltetése.

5.2.1. megjegyzés. A kontrakciós tulajdonságra nincs szükségünk a konvergencia garantálásához, ha az F $[a,b]$-ből $[a,b]$-be képező folytonos függvény monoton növő, és a-ból vagy b-ből indítjuk az iterációt. Ha pl. $a = x^{(0)}$-ból indítjuk, akkor $x^{(1)} = F(x^{(0)}) \geq x^{(0)}$, és a monotonitás miatt $x^{(2)} = F(x^{(1)}) \geq F(x^{(0)}) = x^{(1)}$, azaz a generált sorozat monoton növő lesz. Mivel minden elem felső korlátja b, így a sorozat konvergens lesz. A határérték F folytonossága miatt csak F egy fixpontja lehet (az a-hoz legközelebbi). $\diamond$

A fixpont-iteráció konvergenciájáról szól az alábbi tétel.

5.2.2. tétel.

Legyen $F : [a,b] \to [a,b]$ kontrakció, továbbá legyen F legalább r-szer folytonosan differenciálható úgy, hogy

$$F'(x^\star) = \ldots = F^{(r-1)}(x^\star) = 0,$$

és $F^{(r)}(x^\star) \neq 0$. Ekkor az F által meghatározott fixpontiteráció $[a,b]$ bármelyik pontjából indítva r-ed rendben tart az F függvény egyetlen $[a,b]$-beli fixpontjához.

Bizonyítás. Csak azt kell igazolnunk, hogy a konvergencia rendje r. A tétel többi állítása a Banach-féle fixponttétel következménye.

Az $x^\star$ körüli sorfejtésből (feltesszük, hogy $x^\star \neq a, b$, különben megtaláltuk a fixpontot)

$$F(x^{(k)}) = F(x^\star) + \frac{F^{(r)}(\xi)}{r!}(x^{(k)} - x^\star)^r,$$

azaz

$$F(x^{(k)}) - x^\star = \frac{F^{(r)}(\xi)}{r!}(x^{(k)} - x^\star)^r.$$

Mivel az r-edik derivált folytonos és $x^\star$-ban nem nulla, ezért $x^\star$ egy környezetében teljesülnek az 1.3.5. tétel feltételei. Így ezen tétel miatt a konvergenciarend valóban r. ■

Most megadunk egy olyan hibabecslő formulát, amely két egymást követő sorozatelem segítségével mond becslést a fixpponttól való távolságra.

5.2.3. tétel.

Az $f(x) = 0$ egyenlet egy $[a, b]$ intervallumbeli $x^\star$ megoldásának megkeresésére alkalmazzuk az $x^{(k+1)} = F(x^{(k)})$ fixpont-iterációt, ahol az $F(x)$ függvényt az $F(x) = x - g(x)f(x)$ módon választjuk. Tegyük fel, hogy F az $[a, b]$ intervallumot önmagára képezi, hogy f és g az $[a, b]$ intervallumon folytonos, (a, b)-ben deriválható függvény, hogy g-nek nincs zérushelye $[a, b]$-ben, valamint hogy teljesül az $m_1 := \min_{x\in(a,b)}\{|(gf)'(x)|\} > 0$ feltétel. Ekkor ha valamilyen $\varepsilon > 0$ számra és k indexre

$$\frac{|x^{(k+1)} - x^{(k)}|}{|x^{(k)}|} \leq \frac{\varepsilon}{1+\varepsilon} m_1, \tag{5.2.2}$$

akkor az

$$|x^{(k)} - x^\star| \leq \varepsilon |x^\star|$$

becslés is igaz.

Bizonyítás. Rendezzük át a k-adik iterációs lépés $x^{(k+1)} = x^{(k)} - g(x^{(k)})f(x^{(k)})$ képletét, majd alkalmazzuk a Lagrange-középértéktételt.

$$|x^{(k+1)} - x^{(k)}| = |g(x^{(k)})f(x^{(k)})| = |g(x^{(k)})f(x^{(k)}) - \underbrace{g(x^\star)f(x^\star)}_{=0}|$$
$$= |(gf)'(\xi_k)| \cdot |x^{(k)} - x^\star| \geq m_1 |x^{(k)} - x^\star|,$$

ahol ξ_k megfelelő konstans az $x^\star$ és $x^{(k)}$ pontok között (Lagrange-középértéktétel).

Becsüljük az (5.2.2) képletből m_1 értékét, majd helyettesítsük be azt a fenti képletbe

$$|x^{(k+1)} - x^{(k)}| \geq \frac{1+\varepsilon}{\varepsilon} \frac{|x^{(k+1)} - x^{(k)}|}{|x^{(k)}|} |x^{(k)} - x^\star|.$$

Az $|x^{(k+1)} - x^{(k)}|$ tényezővel egyszerűsítve az $|x^{(k)} - x^\star|$ hibára az alábbi becslés nyerhető

$$|x^{(k)} - x^\star| \leq \varepsilon |x^{(k)}| - \varepsilon |x^{(k)} - x^\star| = \varepsilon(||x^{(k)}| - |x^{(k)} - x^\star||) \leq \varepsilon |x^{(k)} - (x^{(k)} - x^\star)| = \varepsilon |x^\star|.$$

Az utolsó előtti lépésben felhasználtuk az 1.1.10. tételt. ■

A tételben bizonyított állítás felhasználható leállási feltételként az iteráció végrehajtása során. Mivel $|x^\star| \leq \max\{|a|, |b|\}$, ezért ha ε-t úgy választjuk meg, hogy az $\varepsilon \max\{|a|, |b|\}$ érték kisebb legyen, mint egy elérni kívánt hibaérték, akkor az (5.2.2) feltétel teljesüléséből következik, hogy $x^{(k)}$ az adott hibértéknél közelebb lesz a fixponthoz.

5.2.1. Aitken-gyorsítás

Az alábbi eljárás segítségével a fixponthoz tartó sorozat konvergenciáját gyorsíthatjuk fel. Az eljárást Aitken[3]-gyorsításnak nevezzük.

Tegyük fel, hogy van egy iterációval előállított elsőrendű konvergens sorozatunk. Ekkor az előző fejezet szerint

$$e^{(k+1)} = F'(\xi_k) e^{(k)},$$

ahol ξ_k $x^{(k)}$ és $x^\star$ közé esik. Hasonlóan

$$e^{(k)} = F'(\xi_{k-1}) e^{(k-1)}.$$

[3] Alexander Aitken, 1895-1967, Új-Zéland

Az F' függvény folytonossága miatt nyilvánvalóan $F'(\xi_{k-1}) \to F'(x^\star)$, azaz ha k értéke elég nagy, akkor a fenti két képletben szereplő konstansok kb. megegyeznek. Így

$$(e^{(k)})^2 = e^{(k-1)}e^{(k+1)},$$

azaz

$$(x^{(k)} - x^\star)^2 = (x^{(k-1)} - x^\star)(x^{(k+1)} - x^\star).$$

Fejezzük ki innét az $x^\star$ fixpontot.

$$(x^{(k)})^2 + (x^\star)^2 - 2x^{(k)}x^\star$$
$$= x^{(k-1)}x^{(k+1)} - x^\star(x^{(k-1)} + x^{(k+1)}) + (x^\star)^2$$

Ebből a zérushely

$$x^\star = \frac{x^{(k-1)}x^{(k+1)} - (x^{(k)})^2}{x^{(k-1)} + x^{(k+1)} - 2x^{(k)}} = \frac{x^{(k-1)}x^{(k+1)} - (x^{(k)})^2}{\underbrace{x^{(k+1)} - x^{(k)}}_{=:\Delta x_{k+1}} - \underbrace{(x^{(k)} - x^{(k-1)})}_{\Delta x_k}}$$

$$= \frac{x^{(k+1)}(x^{(k-1)} + x^{(k+1)} - 2x^{(k)}) - ((x^{(k+1)})^2 - 2x^{(k+1)}x^{(k)} + (x^{(k)})^2)}{\Delta x_{k+1} - \Delta x_k}$$

$$= x^{(k+1)} - \frac{(\Delta x_{k+1})^2}{\underbrace{\Delta x_{k+1} - \Delta x_k}_{=:\Delta^2 x_k}}.$$

A bevezetett jelölések segítségével tehát

$$x^\star = x^{(k+1)} - \frac{(\Delta x_{k+1})^2}{\Delta^2 x_k} =: \hat{x}^{(k-1)}$$

Természetesen az alkalmazott közelítés miatt ez nem adja meg pontosan $x^\star$ értékét, de az így kiszámolt $\{\hat{x}^{(k-1)}\}$ sorozat gyorsabban tart $x^\star$-hoz, mint az eredeti sorozat, nevezetesen igaz, hogy

$$\lim_{k\to\infty} \frac{\hat{x}^{(k)} - x^\star}{x^{(k)} - x^\star} = 0.$$

5.3. Mintafeladat

Ebben a fejezetben az $x^2 - 2 = 0$ egyenleten mutatjuk meg az egyes ismertetett módszerek alkalmazását. Határozzuk meg az egyenlet $x^\star = \sqrt{2}$ megoldását. Kiindulási intervallumnak mindig válasszuk az $[1, 2]$ intervallumot. Ebben előre tudhatjuk, hogy van megoldás, hiszen $f(x) = x^2 - 2$ folytonos, $f(1) = -1$ és $f(2) = 2$. Az iterációkat addig csináljuk, míg a megoldást 10^{-6}-nál kisebb hibával kapjuk már meg.

Intervallumfelezési módszer. Előre kiszámolható, hogy hány lépést kell tennünk az adott pontosság eléréséhez. $k_{\max} = (\ln(b-a)/10^{-6})/\ln 2 - 1 = (\ln(10^6))/\ln 2 - 1 \approx 18.9316$, azaz 19 iterációs lépés elegendő. Az iterációs lépéseket az 5.3.1. táblázatban foglaltuk össze. Érdemes észrevenni, hogy a konvergencia nem monoton, vagyis az utolsó oszlopbeli értékek nem monoton csökkenőek.

Húrmódszer. Mivel $f'(x) = 2x$ és $f''(x) = 2$, emiatt az első és második derivált nem vált előjelet az $[1, 2]$ intervallumban, így az 5.1.9. tétel feltételei teljesülnek. Az $m_1 = 2$ és $M_2 = 2$ választás megfelelő, kiindulásként megadható pl. az $|e^{(0)}| = |a - x^\star| = |1 - x^\star| \le 1$ felső becslés.

k	a	b	$x^{(k)}$	$\lvert x^{(k)} - x^\star \rvert$
0	1.00000000000000	2.00000000000000	1.50000000000000	0.08578643762690
1	1.00000000000000	1.50000000000000	1.25000000000000	0.16421356237310
2	1.25000000000000	1.50000000000000	1.37500000000000	0.03921356237310
3	1.37500000000000	1.50000000000000	1.43750000000000	0.02328643762690
4	1.37500000000000	1.43750000000000	1.40625000000000	0.00796356237310
5	1.40625000000000	1.43750000000000	1.42187500000000	0.00766143762690
6	1.40625000000000	1.42187500000000	1.41406250000000	0.00015106237310
7	1.41406250000000	1.42187500000000	1.41796875000000	0.00375518762690
8	1.41406250000000	1.41796875000000	1.41601562500000	0.00180206262690
9	1.41406250000000	1.41601562500000	1.41503906250000	0.00082550012690
10	1.41406250000000	1.41503906250000	1.41455078125000	0.00033721887690
11	1.41406250000000	1.41455078125000	1.41430664062500	0.00009307825190
12	1.41406250000000	1.41430664062500	1.41418457031250	0.00002899206060
13	1.41418457031250	1.41430664062500	1.41424560546875	0.00003204309565
14	1.41418457031250	1.41424560546875	1.41421508789063	0.00000152551753
15	1.41418457031250	1.41421508789063	1.41419982910156	0.00001373327153
16	1.41419982910156	1.41421508789063	1.41420745849609	0.00000610387700
17	1.41420745849609	1.41421508789063	1.41421127319336	0.00000228917974
18	1.41421127319336	1.41421508789063	1.41421318054199	0.00000038183110
19	1.41421318054199	1.41421508789063	1.41421413421631	0.00000057184321

5.3.1. táblázat: Az intervallumfelezési eljárás adatai a mintafeladatra.

Így az (5.1.1) becslés miatt $|e^{(k)}| \le (1/2)^k$, és a kívánt pontosság eléréséhez elegendő $k_{\max} = 20$ iterációs lépést végezni (a valóságban kevesebb is elég, de a fenti becslésünk ezt adja). Az iterációs lépéseket az 5.3.2. táblázatban foglaltuk össze. Az utolsó oszlop a logaritmikus relatív csökkenést adja meg. Látható, hogy ennek értéke 1-hez tart, hiszen elsőrendű a módszer. A sorozat hibája (második oszlop) $k \ge 2$ esetén monoton csökkenő.

Szelőmódszer. Az 5.1.11. tételnek megfelelően legyen $C = M_2/(2m_1) = 1/2$. Ekkor nyilván teljesül, hogy

$$\max\{|a - x^\star|, |b - x^\star|\} = \max\{|1 - x^\star|, |2 - x^\star|\} \le 1 < 1/C.$$

Így a szelőmódszer monoton módon konvergálni fog. Mivel $d = 1/2$, emiatt alkalmazhatjuk az

$$|e^{(k)}| \le \frac{d^k}{C} = 2d^k \le 10^{-6}$$

becslést. Ebből kapjuk, hogy $k_{\max} = 21$ iterációs lépés után 10^{-6}-nál pontosabb közelítést kapunk. Az 5.3.3. táblázatból látszik, hogy sokkal hamarabb (6. lépésben) is elérjük a kívánt pontosságot, sőt tovább már nem is tudunk menni, mert a MATLAB nullával való osztás miatt hibaüzenetet ad. A $k \ge 2$ értékekre jól látszik a monoton konvergencia (3. oszlop) és a konvergenciarend (4. oszlop).

Newton-módszer. Mivel $f'' = 2 > 0$, és $f(x) > 0$, ha $x > x^\star$, ezért a Newton-módszer bármilyen $x^{(0)} > x^\star$ pontból indítható (5.1.13. tétel). Indítsuk el az $x^{(0)} = 1.9$ pontból. Ekkor nyilván teljesülnek az 5.1.12. tétel feltételei. Így az (5.1.3) becslés miatt, ha az $|e^{(0)}| \le 1$ durva becslést használjuk, akkor $|e^{(1)}| \le C|e^{(0)}|^2 \le 1/2$, $|e^{(2)}| \le C|e^{(1)}|^2 \le 1/8$, $|e^{(3)}| \le C|e^{(2)}|^2 \le 1/128$, $|e^{(4)}| \le C|e^{(3)}|^2 \le 1/32768$ és $|e^{(5)}| \le C|e^{(4)}|^2 \le 1/(2(32768)^2) \approx 4.6566 \times 10^{-10} < 10^{-6}$. Ez mutatja, hogy négy lépés biztosan elég lesz az adott pontosság eléréséhez, ahogy ezt az 5.3.4. táblázatban szereplő adatok is mutatják. Figyeljük meg, hogy a logaritmikus relatív csökkenés valóban 2 közeli.

k	$x^{(k)}$	$\lvert x^{(k)} - \sqrt{2}\rvert$	log. rel. csökk.
0	1.00000000000000	0.41421356237310	-
1	2.00000000000000	0.58578643762690	-
2	1.33333333333333	0.08088022903976	4.70229223589887
3	1.40000000000000	0.01421356237310	1.69141982231166
4	1.41176470588235	0.00244885649074	1.41343626909880
5	1.41379310344828	0.00042045892482	1.29307890232953
6	1.41414141414141	0.00007214823168	1.22672843924426
7	1.41420118343195	0.00001237894114	1.18483435747867
8	1.41421143847487	0.00000212389823	1.15600171615800
9	1.41421319796954	0.00000036440355	1.13494961104049
10	1.41421349985132	0.00000006252177	1.11890364913929
11	1.41421355164605	0.00000001072704	1.10626800210629
12	1.41421356053263	0.00000000184047	1.09605990342007
13	1.41421356205732	0.00000000031577	1.08764109222480
14	1.41421356231892	0.00000000005418	1.08057902359452
15	1.41421356236380	0.00000000000930	1.07456970644472
16	1.41421356237150	0.00000000000159	1.06939397673514
17	1.41421356237282	0.00000000000027	1.06487465606826
18	1.41421356237305	0.00000000000005	1.06086537951130
19	1.41421356237309	0.00000000000001	1.05688612585613
20	1.41421356237309	0.00000000000000	1.05133725968926

5.3.2. táblázat: A húrmódszer lépései a tesztfeladatra.

k	$x^{(k)}$	$\lvert x^{(k)} - x^\star\rvert$	log. rel. csökk.
0	1.00000000000000	0.41421356237310	-
1	2.00000000000000	0.58578643762690	-
2	1.33333333333333	0.08088022903976	4.70229223589887
3	1.40000000000000	0.01421356237310	1.69141982231166
4	1.41463414634146	0.00042058396837	1.82761471235075
5	1.41421143847487	0.00000212389823	1.68027808569044
6	1.41421356205732	0.00000000031577	1.67474819872268

5.3.3. táblázat: A szelőmódszer iterációs lépései.

Fixpont iteráció. Az $x^2 - 2 = 0$ egyenletet átírhatjuk pl. az $x = 2/x =: F(x)$ alakra, mellyel definiálhatjuk az $x^{(k+1)} = F(x^{(k)})$ iterációt. Sajnos ez az iteráció nem fogja megtalálni a fixpontot, hiszen $x^{(0)}$-ról indítva $x^{(1)} = 2/x^{(0)}$ és $x^{(2)} = 2/(2/x^{(0)}) = x^{(0)}$, azaz az iteráció felváltva lépeget $x^{(1)}$-re és $x^{(0)}$-ra. Így más iterációs eljárást kell keresnünk.

Írjuk át az egyenletet az ekvivalens $x = x - g(x^2 - 2) =: F(x)$ alakba, ahol g meghatározandó pozitív konstans. Most $F'(x) = 1 - 2gx$. Az $[1, 2]$ intervallumon ennek akkor lesz a legkisebb az abszolútérték-maximuma, ha $g = 2/5$, és ekkor a maximum $3/5 = q < 1$. Azaz F valóban kontrakció lesz az $[1, 2]$ intervallumon. Továbbá, mivel F-nek stacionárius pontja van $x = 5/4$-nél és $F(1) = 1.4$, $F(2) = 1.2$ és $F(5/4) = 1.425$, ezért F az $[1, 2]$ intervallumot valóban az $[1, 2]$ intervallum belsejébe képezi. Így a Banach-féle fixponttétel miatt bármilyen $x^{(0)} \in [1, 2]$ pontból indítva az

$$x^{(k+1)} = x^{(k)} - \frac{2}{5}((x^{(k)})^2 - 2)$$

k	$x^{(k)}$	$\lvert x^{(k)} - x^\star \rvert$	log. rel. csökk.
0	1.90000000000000	0.48578643762690	-
1	1.47631578947368	0.06210222710059	3.84906734083784
2	1.41551974856928	0.00130618619619	2.38960316159481
3	1.41421416502183	0.00000060264874	2.15670829615046
4	1.41421356237322	0.00000000000013	2.07263104582362
5	1.41421356237310	0	-

5.3.4. táblázat: A Newton-módszer iterációs lépései.

iteráció a fixponthoz fog tartani.

Indítsuk el az iterációt példként az $x^{(0)} = 2$ pontból. Ekkor $x^{(1)} = 1.2$, $|x^{(1)} - x^{(0)}| = 0.8$, és a hibabecslő formulából

$$|x^{(k)} - x^\star| \leq \frac{q^k}{1-q}|x^{(1)} - x^{(0)}| = \frac{0.6^k}{0.4}0.8 \leq 10^{-6},$$

így $k_{\max} = 29$ iteráció elég az előre adott pontosság eléréséhez. Az 5.3.5. táblázat azt mutatja, hogy 8 iterációs lépés is elegendő. A táblázat utolsó előtti oszlopában kiszámoltuk az Aitken-gyorsítással kapott új sorozatot is és annak logaritmikus relatív csökkenését. Látható, hogy az Aitken-gyorsítással előállított elemek hibája kisebb, mint az eredeti sorozat megfelelő hibája. Az új sorozat is elsőrendben konvergál.

k	$x^{(k)}$	$\lvert x^{(k)} - x^\star \rvert$	log. rel. csökk.	Aitken gyorsítás	log. rel. csökk.
0	2.00000000000000	0.58578643762690	-	-	0
1	1.20000000000000	0.21421356237310	2.88104303658437	1.37500000000000	0
2	1.42400000000000	0.00978643762690	3.00286370704064	1.41341463414634	2.20216989083631
3	1.41288960000000	0.00132396237310	1.43234778523415	1.41420899509093	1.72408572874810
4	1.41438679128474	0.00017322891164	1.30688572006857	1.41421348125311	1.32779360045390
5	1.41419079314044	0.00002276923265	1.23429487859350	1.41421356097773	1.24883257677142
6	1.41421655337916	0.00000299100607	1.18987664275526	1.41421356234091	1.18486140714539
7	1.41421316943851	0.00000039293459	1.15957057942449	1.41421356233394	0.99188777543203
8	1.41421361399318	0.00000005162009	1.13761248562765	-	-

5.3.5. táblázat: A fixpont-iteráció iterációs lépései.

5.4. Nemlineáris egyenletrendszerek megoldása

Tegyük fel, hogy $\bar{\mathbf{f}} : H \to H$ egy adott kellően sokszor folytonosan differenciálható függvény, ahol $H \subset \mathbb{R}^n$ egy zárt halmaz. Keressük azon $\overline{\mathbf{x}} \in H$ vektorokat, melyekre $\bar{\mathbf{f}}(\overline{\mathbf{x}}) = \mathbf{0}$. Ha $\bar{\mathbf{f}}$ lineáris, akkor egy lineáris egyenletrendszert kell megoldanunk, különben nemlineáris egyenletrendszert kapunk. A nemlineáris egyenletrendszerek megoldása általában nehéz feladat. Már annak eldöntése sem egyszerű, hogy van-e, és ha igen, hány megoldása van az egyenletrendszernek. Magasabb dimenzióban az a lehetőségünk is megszűnik, hogy számítógéppel ábrázolva a függvényeket meg tudjuk sejteni a megoldás helyét. Kétváltozós, kétismeretlenes egyenletrendszerek esetén segíthet, ha számítógépes program segítségével megrajzoljuk $\bar{\mathbf{f}}$ koordinátafüggvényeinek a nulla értékhez tartozó szintvonalait, és ezek metszéspontját megkeresve megsejthető a megoldás helye.

A nemlineáris egyenletrendszerek megoldására általában fixpont-iterációt szokás alkalmazni. Azaz átírjuk az egyenletrendszert ekvivalens módon $\mathbf{F}(\overline{\mathbf{x}}) = \overline{\mathbf{x}}$ alakra és, ha az $\mathbf{F}$ függvényre teljesülnek a Banch-féle fixponttétel feltételei (természetesen ennek ellenőrzése általában nem

egyszerű), akkor annak segítségével a fixpont, vagyis az eredeti feladat megoldása meghatározható.

Néha nem szükséges a Banach-féle fixponttétel feltételeinek ellenőrzése. Egyszerűen elindítjuk egy $\overline{\mathbf{x}}^{(0)} \in H$ pontból az

$$\overline{\mathbf{x}}^{(k+1)} = \mathbf{F}(\overline{\mathbf{x}}^{(k)})$$

iterációt, és ha ez a sorozat konvergál egy $\overline{\mathbf{x}}^\star$ vektorhoz, akkor nyilvánvalóan $\mathbf{F}(\overline{\mathbf{x}}^\star) = \overline{\mathbf{x}}^\star$ miatt $\bar{\mathbf{f}}(\overline{\mathbf{x}}^\star) = \mathbf{0}$ is teljesülni fog ($\mathbf{F}$ folytonos).

A fixpont-iterációk egy speciális formája az egyváltozós esetben megismert Newton-módszer többdimenziós változata. Fejtsük sorba az $\bar{\mathbf{f}}$ függvényt az $\overline{\mathbf{x}}^{(k)}$ pontban, majd alkalmazzuk a sorfejtést az $\overline{\mathbf{x}}^\star$ fixpontra:

$$\mathbf{0} = \bar{\mathbf{f}}(\overline{\mathbf{x}}^\star) \approx \bar{\mathbf{f}}(\overline{\mathbf{x}}^{(k)}) + \mathbf{J}(\overline{\mathbf{x}}^{(k)})(\overline{\mathbf{x}}^\star - \overline{\mathbf{x}}^{(k)}),$$

ahol

$$\mathbf{J}(\overline{\mathbf{x}}^{(k)}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\overline{\mathbf{x}}^{(k)}) & \dots & \frac{\partial f_1}{\partial x_n}(\overline{\mathbf{x}}^{(k)}) \\ \vdots & \dots & \vdots \\ \frac{\partial f_n}{\partial x_1}(\overline{\mathbf{x}}^{(k)}) & \dots & \frac{\partial f_n}{\partial x_n}(\overline{\mathbf{x}}^{(k)}) \end{bmatrix}$$

az ún. Jacobi-mátrix. Innét kifejezhetjük $\overline{\mathbf{x}}^\star$-t:

$$\overline{\mathbf{x}}^\star \approx \overline{\mathbf{x}}^{(k)} - (\mathbf{J}(\overline{\mathbf{x}}^{(k)}))^{-1}\bar{\mathbf{f}}(\overline{\mathbf{x}}^{(k)}).$$

Mivel ez csak egy közelítés, ez nem adja meg $\overline{\mathbf{x}}^\star$-t, de alkalmas arra, hogy egy iterációt definiáljunk az

$$\overline{\mathbf{x}}^{(k+1)} \approx \overline{\mathbf{x}}^{(k)} - (\mathbf{J}(\overline{\mathbf{x}}^{(k)}))^{-1}\bar{\mathbf{f}}(\overline{\mathbf{x}}^{(k)})$$

alakban.

Vegyük észre, hogy a fenti iteráció valóban a Newton-módszer általánosítása, ezért ezt többváltozós Newton-módszernek nevezzük. Egyszerűen meg is jegyezhető, hiszen a deriválttal való osztás helyett a Jacobi-mátrix (deriváltak mátrixa) inverzével való szorzás szerepel. A módszer végrehajtása során műveletigényes, hogy a Jacobi-mátrixot minden lépésben ki kell számolni és meg kell oldani egy lineáris egyenletrendszert.

Ha az $\overline{\mathbf{x}}^\star$ megoldás egy környezetében a Jacobi-mátrixok inverzei korlátosak, és a Jacobi-mátrix Lipschitz-folytonos (megfelelő indukált normákat használva), akkor $\overline{\mathbf{x}}^\star$ egy elegendően kicsi környezetéből indítva az iterációt, az az egyenletrendszer megoldásához fog tartani, és a konvergencia másodrendű lesz.

5.4.1. megjegyzés. Láttuk, hogy a többváltozós Newton-módszernél minden lépésben meg kell oldanunk egy lineáris egyenletrendszert, melynek mátrixa az előző iterációs lépésnek megfelelő pontban kiszámolt Jacobi-mátrix. Ha nagymértű feladatot kell megoldnunk, akkor módosíthatjuk úgy az iterációt, hogy minden lépésben a kezdőpontban kiszámolt Jacobi-mátrixszal oldjuk meg az egyenletrendszert:

$$\overline{\mathbf{x}}^{(k+1)} = \overline{\mathbf{x}}^{(k)} - (\mathbf{J}(\overline{\mathbf{x}}^{(0)}))^{-1}\bar{\mathbf{f}}(\overline{\mathbf{x}}^{(k)}).$$

Ezt az eljárást módosított Newton-módszernek hívjuk. Ez a módszer gyorsabb, de már csak elsőrendű konvergenciát biztosít. ◇

5.5. Feladatok

Nemlineáris egyenletek megoldása

5.5.1. feladat. A $\sqrt{1}+\sqrt{2}+\ldots+\sqrt{100}$ összeget úgy számítjuk ki, hogy mindegyik tagot két tizedesjegyre kerekítjük. Hány helyes tizedesjegyre számíthatunk legalább a végeredményben?

5.5.2. feladat. 1225-ben Leonardo Pisano (alias Leonardo Bonacci, vagy Leonardo Fibonacci, vagy egyszerűen Fibonacci) meghatározta a

$$p(x) = x^3 + 2x^2 + 10x - 20$$

polinom $x = 1.368808107$ zérushelyét. Azt, hogy hogyan sikerült erre az eredményre jutnia, nem tudjuk. Határozzunk meg egy olyan intervallumot, amelyből a fenti polinom zérushelyei kikerülhetnek! Van-e más valós zérushelye a polinomnak?

5.5.3. feladat. Tegyük fel, hogy tudjuk, hogy az előző feladatban szereplő polinom zérushelye az [1,2] intervallumban van. Határozzuk meg a zérushelyet az intervallumfelezési-, a húr-, a szelő-, a Newton- és a fixpont-iterációs módszer segítségével (írjunk egyszerű MATLAB programokat)! Utóbbinál alkalmazzuk az $x = 20/(x^2+2x+10)$ egyenlőséget az iteráció konstrukciójában (igazoljuk a konvergenciát). Hasonlítsuk össze a módszereket a konvergenciasebesség szerint!

5.5.4. feladat. Alkalmazzuk az Aitken-gyorsítást az előző feladatbeli fixpont-iteráció $x^{(10)}, x^{(11)}$ és $x^{(12)}$ elemével! Mit tapasztalunk?

5.5.5. feladat. Oldjuk meg az alábbi egyenleteket: $2x - \ln x - 4 = 0$, $2^x - 4x = 0$, $x^5 - x - 2 = 0$, $x \ln x - 14 = 0$!

5.5.6. feladat. Adjunk olyan eljárást, amely osztás nélkül határozza meg egy a pozitív szám reciprokát! Kísérletezzünk az $1 - ax = 0$ és az $a - 1/x = 0$ egyenletek Newton-módszeres megoldásaival!

5.5.7. feladat. Igazoljuk, hogy ha egy α szám egy $p(x)$ polinom zérushelye, akkor az α számmal végrehajtva a Horner-módszert, a keletkező számok éppen a $p(x)/(x-\alpha)$ polinom együtthatóit adják meg. Ha tehát egy polinomnak megtaláltuk egy zérushelyét, akkor a fokszáma csökkenthető a fenti eljárással (defláció). Határozzuk meg az $x^4 - 16x^3 + 72x^2 - 96x + 24$ polinom összes zérushelyét a Newton-módszer és a deflációs eljárás segítségével!

5.5.8. feladat. Hány megoldása van a $\sin x = 1 - x$ egyenletnek a $[0,1]$ intervallumon? Határozzuk meg a gyököt intervallumfelezéssel! Hány lépés kell ahhoz, hogy a gyököt 4 tizedesjegy pontossággal meghatározhassuk. Határozzuk meg a gyököt fixpont-iterációval! Itt is becsüljük meg a szükséges lépések számát!

5.5.9. feladat. Egy harmadfokú $p(x)$ polinomnak három valós zérushelye van: z_1, z_2 és z_3. Mutassuk meg, hogy ha a Newton-módszert a $(z_1+z_2)/2$ pontból indítjuk, akkor a módszer egy lépésben megadja a z_3 gyököt!

5.5.10. feladat. Javasoljunk legalább két fixpont-iterációt az $e^{-x} - \sin x = 0$ egyenlet $\alpha \approx 0.5885$ gyökének megkeresésére. Vizsgáljuk meg a konvergenciát!

5.5.11. feladat. A Steffensen-módszer az alábbi fixpont-iterációval oldja meg az $f(x) = 0$ egyenletet:

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{\phi(x^{(k)})}, \quad \phi(x^{(k)}) = \frac{f(x^{(k)} + f(x^{(k)})) - f(x^{(k)})}{f(x^{(k)})}.$$

Igazoljuk, hogy a módszer másodrendű, és oldjuk meg vele az előző feladat egyenletét!

5.5.12. feladat. Az $f(x) = (2x^2 - 3x - 2)/(x-1)$ függvény zérushelyeinek megkereséséhez használjuk az alábbi fixpont-iterációkat:

$$x^{(k+1)} = \frac{3(x^{(k)})^2 - 4x^{(k)} - 2}{x^{(k)} - 1}, \quad x^{(k+1)} = x^{(k)} - 2 + \frac{x^{(k)}}{x^{(k)} - 1}.$$

Vizsgáljuk meg a módszerek rendjét! Milyen kezdőpontból fog 2-höz konvergálni a második módszer?

5.5.13. feladat. Oldjuk meg az $f(x) = e^x - 1 - x = 0$ egyenletet a Newton-módszerrel! Mekkora lesz a konvergenciarend? Próbáljuk ki az

$$x^{(k+1)} = x^{(k)} - \frac{2f(x^{(k)})}{f'(x^{(k)})}$$

iterációt a megoldásra!

5.5.14. feladat. A Newton-módszer is tulajdonképpen egy fixpont-iteráció. Határozzuk meg a konvergenciarendjét a fixpont-iterációk konvergenciarendjére vonatkozó tétel alapján!

5.5.15. feladat. Alkalmazzuk a Newton-módszert az $f(x) = x^3 - 2x + 2$ függvényre a 0 értéket választva kezdőpontnak, és a $g(x) = \sqrt[3]{x}$ függvényre!

Nemlineáris egyenletrendszerek megoldása

5.5.16. feladat. Oldjuk meg a Newton-módszer segítségével az alábbi egyenletrendszert!

$$1 - 4x + 2x^2 - 2y^3 = 0$$
$$-4 + x^4 + 4y + 4y^4 = 0$$

6. Interpolációs feladatok

Ebben a fejezetben arra a kérdésre válaszolunk, hogy néhány, a derékszögű koordinátarendszerben adott pontot hogyan tudunk összekötni polinomok ill. trigonometrikus polinomok grafikonjaival. Az összekötő polinomokat interpolációs polinomoknak hívjuk. Tárgyalni fogjuk az interpolációs polinomok Lagrange- és Newton-féle alakját. Megvizsgáljuk, hogy az alappontok megválasztása hogyan befolyásolja az interpolációs hibát. Így jutunk el a Csebisev-alappontok fogalmához. A trigonometrikus interpolációs feladatnál szó lesz a diszkrét Fourier-transzformációról és a gyors Fourier-transzformációról.

Tegyük fel, hogy egy $f : \mathbb{R} \to \mathbb{R}$ függvénynek $n+1$ pontban ismerjük az értékét. Legyenek ezek a pontok adottak az (x_i, f_i) $(i = 0, \ldots, n)$ alakban, ahol feltesszük hogy $x_i \neq x_j$, ha $i \neq j$. Az x_i $(i = 0, \ldots, n)$ pontokat alappontoknak nevezzük. Vezessük be az $I = [x_{\min}, x_{\max}]$ jelölést, ahol

$$x_{\min} = \min\{x_0, \ldots, x_n\}, \quad x_{\max} = \max\{x_0, \ldots, x_n\}.$$

Hogyan adhatnánk becslést az f függvény értékére az I intervallumbeli alappontoktól különböző pontban? Hogyan becsülhetnénk az f függvény deriváltját egy adott pontban vagy az integrálját egy adott intervallumon? Hogyan adhatnánk becslést arra, hogy hol van a függvénynek szélsőértéke?

A fenti feladatokat általánosan úgy oldhatjuk meg, hogy rögzítjük az I intervallumon definiált adott tulajdonságú függvények halmazát (pl. polinomok, trigonometrikus polinomok vagy szakaszonként polinomfüggvények), és ennek elemei közül kiválasztjuk azt, amelyik grafikonja átmegy az összes adott ponton. Az így nyert függvényt az adott pontokhoz tartozó interpolációs függvénynek hívjuk. Ez általában egyértelműen meghatározott a függvényhalmazon belül. Az I intervallum alappontoktól eltérő pontjaiban az interpolációs függvény értékeivel közelítjük az ismeretlen függvény értékeit. Ezt az eljárást *interpolációnak* nevezzük. Ezek után a kérdésekre az interpolációs függvény segítségével válaszolunk: annak az értékeit számoljuk ki, azt deriváljuk vagy integráljuk, vagy annak határozzuk meg a szélsőértékeit.

Az interpolációs feladatok számtalanszor előfordulnak gyakorlati problémák megoldásánál. A GPS műholdak csak bizonyos időközönként küldenek jelet arról, hogy hol vannak. Interpolációt használunk akkor, ha szeretnénk a köztes időpontokban is megbecsülni a műhold helyét, a sebességét, vagy a gyorsulását. A kőolajmezőket modellező egyenletekben szükségünk van a nyomás értékére. A költséges fúrások miatt a nyomásfüggvényt csak néhány fúrásnál elvégzett mérés alapján, interpolációval tudjuk közelíteni. Több esetben egy parciális differenciálegyenlet peremfeltételét és kezdeti feltételét is csak több mérési adatot interpolálva tudjuk közelíteni.

6.1. Globális polinominterpoláció

Tekintsük először azt az esetet, amikor az adott pontokat egyetlen polinom segítségével szeretnénk interpolálni a teljes I intervallumon. Ezt az eljárást *globális polinominterpolációnak* nevezzük és az interpolációs függvényt (globális) interpolációs polinomnak hívjuk. Tegyük fel, hogy az adott pontokat interpolálja egy p polinom. Ekkor a $p(x_i) = f_i$ $(i = 0, \ldots, n)$ feltételnek nyilvánvalóan teljesülnie kell. Mivel ez $n+1$ feltétel, így várható, hogy p egy legfeljebb n-edfokú polinom lesz, hiszen ezeknek $n+1$ szabadon választható együtthatója van.

6.1.1. tétel.
Minden rögzített $n+1$ darab ponthoz pontosan egy olyan legfeljebb n-edfokú L_n polinom van, melyre $L_n(x_i) = f_i$ $(i = 0, \dots, n)$.

Bizonyítás. Legyen a keresett polinom $L_n(x) = \sum_{k=0}^{n} a_k x^k$. Ekkor az interpolációhoz az alábbi egyenlőségeknek kell teljesülniük:

$$L_n(x_i) = \sum_{k=0}^{n} a_k x_i^k = f_i \ (i = 0, \dots, n).$$

Ez egy olyan lineáris egyenletrendszer, melynek együtthatómátrixa egy Vandermonde-mátrix. Mivel $x_i \neq x_j$, ha $i \neq j$, ezért a mátrix determinánsa nem nulla. Ebből következik, hogy az egyenletrendszer megoldása, azaz a polinom együtthatói, egyértelműen meghatározottak. ■

6.1.2. megjegyzés. A fenti tételben az együtthatómátrix invertálhatósága más módszerrel is igazolható. Ez a módszer általánosabb esetekben is alkalmazható lesz majd. Az együtthatókra felírt egyenletrendszer az alábbi alakú

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}.$$

Ha feltennénk indirekt, hogy az együtthatómátrix nem invertálható, akkor létezne a mátrixhoz egy olyan $\overline{\mathbf{0}} \neq \overline{\mathbf{y}} = [y_0, \dots, y_n]^T \in \mathbb{R}^{n+1}$ vektor, mellyel

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Ez viszont azt jelentené, hogy a $\sum_{k=0}^{n} y_k x^k$ legfeljebb n-edfokú, a konstans nulla polinomtól különböző polinomnak $n+1$ zérushelye lenne, ami ellentmondana az algebra alaptételének. ◇

6.1.1. Az interpolációs polinom Lagrange-féle előállítása

Most bemutatjuk az interpolációs polinom Lagrange[1]-tól származó előállítási módját.

[1] Joseph-Louis Lagrange, 1736–1813, olasz származású francia matematikus. Eredeti olasz neve Giuseppe Luigi Lagrangia. Fontos eredményeket ért el az analízis, a számelmélet és az elméleti mechanika területén. Bővebb életrajz: http://www-history.mcs.st-and.ac.uk/Mathematicians/Lagrange.html.

6.1.3. definíció.
Adott alappontok esetén az

$$l_k(x) = \frac{(x-x_0)\dots(x-x_{k-1})(x-x_{k+1})\dots(x-x_n)}{(x_k-x_0)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_n)}$$

$(k=0,\dots,n)$ polinomot a k-adik (alapponthoz tartozó) *Lagrange-féle alappolinomnak* nevezzük.

6.1.4. megjegyzés. Vezessük be a $w_{n+1}(x) = (x-x_0)\dots(x-x_n)$ jelölést. A $w_{n+1}(x)$ pontosan $n+1$-edfokú polinomot *alappontpolinomnak* hívjuk. Az alappontpolinom deriváltja

$$\begin{aligned} w'_{n+1}(x) = &(x-x_1)(x-x_2)(x-x_3)\dots(x-x_n)\\ &+(x-x_0)(x-x_2)(x-x_3)\dots(x-x_n)\\ &+(x-x_0)(x-x_1)(x-x_3)\dots(x-x_n)\\ &+\dots+(x-x_0)(x-x_1)(x-x_2)\dots(x-x_{n-1}), \end{aligned}$$

melybe az x_k alappontot helyettesítve a

$$w'_{n+1}(x_k) = (x_k-x_0)(x_k-x_1)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_n)$$

kifejezést kapjuk. Innét látható, hogy az alappontpolinom segítségével az l_k Lagrange-féle alappolinom az

$$l_k(x) = \frac{w_{n+1}(x)}{(x-x_k)w'_{n+1}(x_k)} \tag{6.1.1}$$

alakban is felírható $(k=0,\dots,n)$. $\diamond$

Példaként a 6.1.1. ábrán megrajzoltuk az $x_0=0$, $x_1=1$, $x_2=2$, $x_3=3$ és $x_4=4$ alappontokhoz tartozó Lagrange-féle alappolinomok grafikonjait.

6.1.5. tétel.
Az interpolációs polinom $n+1$ alappont esetén az

$$L_n(x) = \sum_{k=0}^{n} f_k l_k(x)$$

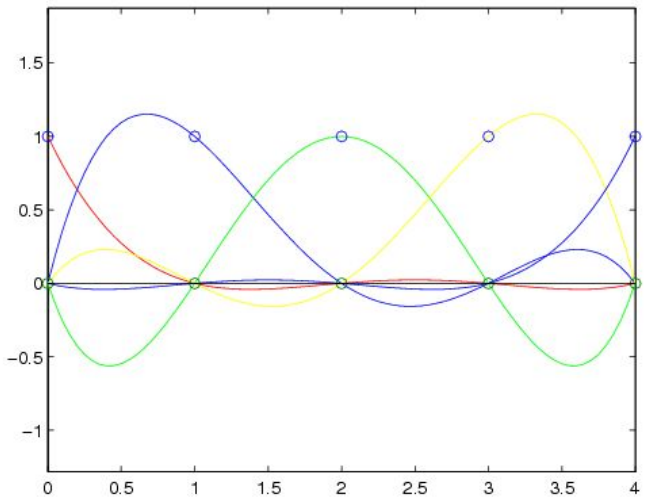
alakban írható.

Bizonyítás. Mivel a Lagrange-féle alappolinomok pontosan n-edfokúak, ezért az $L_n(x)$ polinom legfeljebb n-edfokú. A Lagrange-féle alappolinomokra nyilvánvalóan igaz az

$$l_k(x_i) = \begin{cases} 1, & \text{ha } i=k,\\ 0, & \text{ha } i\neq k \end{cases}$$

egyenlőség, ezért teljesül az $L_n(x_k)=f_k$ egyenlőség is $(k=0,\dots,n)$. Innét az interpolációs polinom egyértelműségéből következik az állítás. ■

6.1.6. megjegyzés. A 6.1.1. tételt igazolhatjuk az interpolációs polinom Lagrange-féle előállításának segítségével is. Az előállítás mutatja, hogy valóban van egy legfeljebb n-edfokú interpolációs



6.1.1. ábra: Az $x = 0, 1, 2, 3, 4$ pontokhoz tartozó Lagrange-féle alappolinomok.

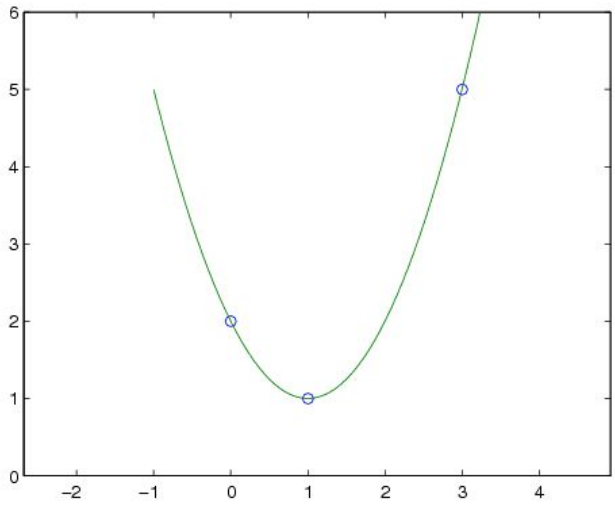
polinom, amely a feltételeknek megfelel. Az egyértelműsége pedig indirekt módon úgy igazolható, hogy ha lenne két különböző legfeljebb n-edfokú interpolációs polinom, akkor a különbségpolinomnak legalább $n+1$ zérushelye lenne (nevezetesen az alappontok), ami ellentmond az algebra alaptételének. $\diamond$

Az interpolációs feladatoknál nem az a cél, hogy az interpolációs polinomot x-hatványonként rendezett alakra hozzuk, hiszen általában nem a polinom előállítása, hanem a polinom különböző helyeken vett helyettesítési értékeinek kiszámolása a cél. Így egyszerűbb a polinom Lagrange-féle alakjába helyettesíteni, mint előállítani a polinomot x-hatványonként rendezett alakban (a polinom-együtthatók kiszámításának műveletigénye n-nel exponenciálisan növekszik) és utána ebbe helyettesíteni.

Könnyen ellenőrizhető, hogy az interpolációs polinom Lagrange-féle előállításával az interpolációs polinom egy adott helyen vett helyettesítési értéke $4n^2 + \mathcal{O}(n)$ flop művelettel számolható. Az előállítás előnye, hogy ha valamelyik alappontban a függvényértéket meg kell változtatnunk, akkor az új polinom helyettesítési értékeit könnyen újraszámolhatjuk a régi értékekből. Hátrány viszont, hogy egy új alappont felvétele esetén minden korábbi tagot újra kell számolnunk.

6.1.7. példa. Példaként határozzuk meg a (0,2), (1,1) és (3,5) pontokhoz tartozó interpolációs polinomot. Először a Lagrange-féle alappolinomokat határozzuk meg:

$$l_0(x) = \frac{(x-1)(x-3)}{(0-1)(0-3)} = \frac{1}{3}x^2 - \frac{4}{3}x + 1,$$
$$l_1(x) = \frac{(x-0)(x-3)}{(1-0)(1-3)} = -\frac{1}{2}x^2 + \frac{3}{2}x,$$
$$l_2(x) = \frac{(x-0)(x-1)}{(3-0)(3-1)} = \frac{1}{6}x^2 - \frac{1}{6}x,$$



6.1.2. ábra: A 6.1.7. példában szereplő pontok és a hozzájuk tartozó interpolációs polinom grafikonja.

majd ezek segítségével felírjuk a keresett interpolációs polinomot:

$$L_2(x) = 2\frac{(x-1)(x-3)}{(0-1)(0-3)} + 1\frac{(x-0)(x-3)}{(1-0)(1-3)} + 5\frac{(x-0)(x-1)}{(3-0)(3-1)}.$$

A polinomot x-hatványonként rendezett alakba írva

$$L_2(x) = 2\left(\frac{1}{3}x^2 - \frac{4}{3}x + 1\right) + 1\left(-\frac{1}{2}x^2 + \frac{3}{2}x\right) + 5\left(\frac{1}{6}x^2 - \frac{1}{6}x\right) = x^2 - 2x + 2.$$

Az adott pontokat és az interpolációs polinom grafikonját a 6.1.2. ábrán láthatjuk. ⋄

6.1.2. A baricentrikus interpolációs formula

Az interpolációs polinom Lagrange-féle előállításánál láttuk, hogy azt újra kell számolni, ha új alappontot veszünk fel. A formulát egy kicsit átalakítva viszont egy sokkal praktikusabb előállítását kapjuk az interpolációs polinomnak: a baricentrikus formulát. Egyszerűsége ellenére a formula tárgyalása gyakran kimarad a numerikus módszereket tárgyaló könyvekből. Ezt a hiányt pótoljuk most a formula ismertetésével [4].

Vezessük be a

$$q_k = \frac{1}{(x_k - x_0)\dots(x_k - x_{k-1})(x_k - x_{k+1})\dots(x_k - x_n)} \equiv \frac{1}{w'_{n+1}(x_k)}$$

$(k = 0, \dots, n)$ ún. baricentrikus súlyokat. Ezek a számok a Lagrange-féle alappolinomok főegyütt-

hatói. Így az interpolációs polinom a Lagrange-féle előállításból kiindulva

$$L_n(x) = \sum_{k=0}^{n} f_k q_k \frac{w_{n+1}(x)}{x - x_k} = w_{n+1}(x) \sum_{k=0}^{n} \frac{f_k q_k}{x - x_k} \tag{6.1.2}$$

alakra hozható. Ez a formula tovább egyszerűsíthető, ha észrevesszük, hogy

$$w_{n+1}(x) \sum_{k=0}^{n} \frac{q_k}{x - x_k} \equiv 1,$$

hiszen a bal oldal az $(x_k, 1)$ $(k = 0, \ldots, n)$ pontokhoz tartozó interpolációs polinomot adja, ami az interpolációs polinom egyértelműsége miatt csak a konstans 1 polinom lehet. Ennek alapján a (6.1.2) formulát átírhatjuk az alábbi módon:

$$L_n(x) = \frac{L_n(x)}{1} = \frac{w_{n+1}(x) \sum_{k=0}^{n} (f_k q_k)/(x - x_k)}{w_{n+1}(x) \sum_{k=0}^{n} q_k/(x - x_k)} = \frac{\sum_{k=0}^{n} q_k f_k/(x - x_k)}{\sum_{k=0}^{n} q_k/(x - x_k)}.$$

Az így nyert formulát *baricentrikus interpolációs formulának* nevezzük[2]. Ezzel az alakkal már lehet is műveletszámot spórolni. A súlyok kiszámítása egyenként $2n$ flop-ba kerül, azaz összesen $2n^2 + 2n$ flop. Ha a súlyokat már ismerjük, akkor a helyettesítési érték számolása már csak $5n + 5$ flop. Új osztópont felvétele sem költséges, hiszen az új $n + 2$-edik súly kiszámítása $2(n + 1)$ flopot igényel, a korábbi súlyok módosítása pedig egyenként 2 flop, azaz összesen $2(n + 1)$ flop. A súlymódosítás tehát összesen $4n + 4$ flop művelet.

6.1.8. példa. Tekintsük a 6.1.7. példában szereplő pontokat, és határozzuk meg az interpolációs polinomot a baricentrikus formula segítségével! Számítsuk ki a baricentrikus súlyokat:

$$q_0 = \frac{1}{(0-1)(0-3)} = \frac{1}{3},\ q_1 = \frac{1}{(1-0)(1-3)} = -\frac{1}{2},\ q_3 = \frac{1}{(3-0)(3-1)} = \frac{1}{6}.$$

Az interpolációs polinom tehát

$$L_2(x) = \frac{2 \cdot 1/(3(x-0)) - 1 \cdot 1/(2(x-1)) + 5 \cdot 1/(6(x-3))}{1/(3(x-0)) - 1/(2(x-1)) + 1/(6(x-3))}.$$

Ezt x-hatványonként rendezve természetesen ismét a 6.1.7. példában szereplő $x^2 - 2x + 2$ polinomot kapjunk. ◇

6.1.3. Az interpolációs polinom előállítása Newton-féle osztott differenciákkal

Keressük az interpolációs polinomot az (x_i, f_i) $(i = 0, \ldots, n)$ pontokhoz az alábbi alakban:

$$L_n(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \ldots + c_n(x - x_0)\ldots(x - x_{n-1}). \tag{6.1.3}$$

Mivel az egyes tagokban szereplő polinomok lineárisan függetlenek, így minden legalább n-edfokú polinom egyértelműen felírható a fenti alakban. Tehát egyértelműen léteznek olyan $c_0, \ldots, c_n$ konstansok, hogy a fenti polinom az interpolációs polinomot adja.

[2] A baricentrikus, azaz súlyponti, elnevezés arra utal, hogy a képlet hasonló a tömegpontrendszer súlypontját megadó képlethez.

Mivel $f_0 = L_n(x_0) = c_0$, így c_0 értékét már meghatároztuk. Helyettesítsünk most x_1-et a (6.1.3) polinomba:

$$f_1 = L_n(x_1) = c_0 + c_1(x_1 - x_0) = f_0 + c_1(x_1 - x_0),$$

ahonnét

$$c_1 = \frac{f_1 - f_0}{x_1 - x_0}.$$

Hasonlóan eljárva megkaphatjuk a többi együtthatót is, de a képletek egyre komplikáltabbak és nehezen átláthatók lesznek. Most megmutatjuk, hogy hogyan lehet az együtthatókat szisztematikus módon előállítani.

6.1.9. definíció.
Legyenek adva az (y_j, f_j), $j = 1, \dots, s$ pontok ($y_{j_1} \neq y_{j_2}$, ha $j_1 \neq j_2$). Ekkor az

$$[y_1, \dots, y_s]f = \sum_{j=1}^{s} \frac{f_j}{(y_j - y_1)\dots(y_j - y_{j-1})(y_j - y_{j+1})\dots(y_j - y_s)}$$

számot az $(y_1, f_1), \dots, (y_s, f_s)$ *pontokhoz tartozó* $(s-1)$*-edrendű (Newton-féle) osztott differenciának* nevezzük. A nulladrendű osztott differenciát (amennyiben csak egy adott pont szerepel) $[y_j]f = f_j$ módon értelmezzük.

6.1.10. példa. A (0,2), (1,1) és (3,5) pontokhoz tartozó másodrendű osztott differencia a definíció szerint

$$[0,1,3]f = \frac{2}{(0-1)(0-3)} + \frac{1}{(1-0)(1-3)} + \frac{5}{(3-0)(3-1)} = 1.$$

◇

6.1.11. megjegyzés. Vegyük észre, hogy az osztott differencia rendjét úgy definiáltuk, hogy az a felhasznált pontok számánál eggyel kisebb. ◇

6.1.12. tétel.
Az osztott differencia értéke független az y_j $(j = 1, \dots, s)$ pontok sorrendjétől, továbbá igaz az

$$[y_1, \dots, y_s]f = \frac{[y_2, \dots, y_s]f - [y_1, \dots, y_{s-1}]f}{y_s - y_1}$$

összefüggés.

Bizonyítás. Az első állítás a definíció közvetlen következménye. A második részhez igazoljuk, hogy

$$[y_1, \dots, y_s]f \cdot (y_s - y_1) = [y_2, \dots, y_s]f - [y_1, \dots, y_{s-1}]f.$$

Induljunk ki a jobb oldalból:

$$
\begin{aligned}
&[y_2,\dots,y_s]f-[y_1,\dots,y_{s-1}]f\\
&=\sum_{j=2}^{s}\underbrace{\frac{f_j}{(y_j-y_2)\dots(y_j-y_s)}}_{(y_j-y_j)\text{ hiányzik}}-\sum_{j=1}^{s-1}\underbrace{\frac{f_j}{(y_j-y_1)\dots(y_j-y_{s-1})}}_{(y_j-y_j)\text{ hiányzik}}\\
&=\sum_{j=2}^{s}\underbrace{\frac{f_j(y_j-y_1)}{(y_j-y_1)(y_j-y_2)\dots(y_j-y_s)}}_{(y_j-y_j)\text{ hiányzik}}-\sum_{j=1}^{s-1}\underbrace{\frac{f_j(y_j-y_s)}{(y_j-y_1)\dots(y_j-y_{s-1})(y_j-y_s)}}_{(y_j-y_j)\text{ hiányzik}}\\
&=\frac{f_s(y_s-y_1)}{(y_s-y_1)\dots(y_s-y_{s-1})}-\frac{f_1(y_1-y_s)}{(y_1-y_2)\dots(y_1-y_s)}+\sum_{j=2}^{s-1}\underbrace{\frac{f_j(y_s-y_1)}{(y_j-y_1)\dots(y_j-y_s)}}_{(y_j-y_j)\text{ hiányzik}}\\
&=[y_1,\dots,y_s]f\cdot(y_s-y_1).
\end{aligned}
$$

Ezt akartuk megmutatni. ■

6.1.13. tétel.
Az interpolációs polinom (6.1.3) alakjában szereplő c_l $(l=0,\dots,n)$ együtthatók a

$$c_l=[x_0,\dots,x_l]f$$

képlettel számíthatók ki.

Bizonyítás. Legyen az $l\in\{0,\dots,n\}$ index rögzítve, és definiáljuk a $p_l(x)$ polinomot az alábbi módon:

$$p_l(x)=c_0+c_1(x-x_0)+c_2(x-x_0)(x-x_1)+\dots+c_l(x-x_0)\dots(x-x_{l-1}).$$

Nyilvánvaló, hogy $p_l(x_0)=L_n(x_0),\dots,p_l(x_l)=L_n(x_l)$, azaz $p_l(x)$ pontosan az $(x_0,f_0),\dots,(x_l,f_l)$ pontokat interpoláló legfeljebb l-edfokú interpolációs polinom, c_l pedig ennek a főegyütthatója. Számítsuk most ki ezt a főegyütthatót! Ehhez írjuk fel a $p_l(x)$ interpolációs polinomot a Lagrange-módszerrel.

$$p_l(x)=\sum_{k=0}^{l}f_k\frac{(x-x_0)\dots(x-x_{k-1})(x-x_{k+1})\dots(x-x_l)}{(x_k-x_0)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_l)},$$

melynek főegyütthatója

$$c_l=\sum_{k=0}^{l}\frac{f_k}{(x_k-x_0)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_l)},$$

ami a Newton-féle osztott differencia definíciója szerint $[x_0,\dots,x_l]f$. Ezt szerettük volna megmutatni. ■

A c_l együtthatók ezek után rekurzióval az alábbi módon állíthatók elő. A definíció alapján: $[x_i]f=f_i$ $(i=0,\dots,n)$. Ezek a nulladrendű differenciák. A 6.1.12. tétel figyelembevételével az elsőrendű differenciákat az

$$[x_0,x_1]f=\frac{[x_1]f-[x_0]f}{x_1-x_0},\quad [x_1,x_2]f=\frac{[x_2]f-[x_1]f}{x_2-x_1}$$

formulákkal kapjuk, míg a másodrendűek az

$$[x_0, x_1, x_2]f = \frac{[x_1, x_2]f - [x_0, x_1]f}{x_2 - x_0}$$

módon nyerhetők. Az eljárást addig folytatjuk, amíg meg nem kapjuk az utolsó, n-edrendű osztott differenciát is.

6.1.14. példa. Keressük meg a (0,2), (1,1) és (3,5) pontokhoz tartozó interpolációs polinomot az osztott differenciák módszerével! A c_l együtthatók kiszámításához először táblázatba rendezzük az osztott differenciákat. Ebben a táblázatban áttekinthető módon feltüntetjük az osztott differenciákat számoló rekurzió egyes lépéseit. Az egyes oszlopok felső elemei adják a keresett c_l együtthatókat növekvő indexekkel.

x_i	$f_i = [x_i]f$	$[.,.]f$	$[.,.,.]f$
0	$2 = c_0$		
		$\frac{1-2}{1-0} = -1 = c_1$	
1	1		$\frac{2-(-1)}{3-0} = 1 = c_2$
		$\frac{5-1}{3-1} = 2$	
3	5		

Ezután a (6.1.3) képlet alapján kapjuk a keresett $2 + (-1)(x-0) + 1(x-0)(x-1)$ interpolációs polinomot, amelyet a Horner-sémához hasonló $(1(x-1)-1)(x-0)+2$ formában szoktak megadni és a helyettesítési értékeit kiszámolni. A polinomot rendezve a 6.1.7. példából ismert $x^2 - 2x + 2$ polinomhoz jutunk. ⋄

6.2. Az interpolációs hiba

Tegyük fel, hogy adottak az $(x_0, f_0), \ldots, (x_n, f_n)$ pontok, és hogy ezek a pontok egy $f : I \to \mathbb{R}$ függvény grafikonjáról származnak, azaz $f_k = f(x_k)$. Legyen továbbá $I = [x_{min}, x_{max}]$. Jelölje $(L_n f) : I \to \mathbb{R}$ az adott pontokra illesztett interpolációs polinomot. Ilyenkor azt mondjuk, hogy az f függvényt interpoláljuk az $x_0, \ldots, x_n$ pontokban, és L_n az f függvény adott alappontokban vett interpolációs polinomja.

6.2.1. definíció.
Az $E_n : I \to \mathbb{R}$, $E_n(x) = (L_n f)(x) - f(x)$ függvényt az f függvény $(x_0, f_0), \ldots, (x_n, f_n)$ pontokhoz tartozó interpolációs hibafüggvényének nevezzük. $E_n(x)$ az x pontbeli interpolációs hiba.

Természetesen, ha az f függvény D_f értelmezési tartománya tágabb az I halmaznál, akkor az interpolációs hibafüggvényt a D_f halmazon is értelmezhetjük.

Vajon mit mondhatunk az interpolációs hibáról? Tudunk-e rá felső becslést adni? Hogy változik az értéke, ha n értékét növeljük? Tart-e ebben az esetben pontonként az interpolációs polinomok sorozata az eredeti f függvényhez? Ha igen, akkor milyen feltételek mellett egyenletes a konvergencia? Ezekre a kérdésekre keressük a választ ebben a fejezetben.

Rögtön látható, hogy ha az f függvény simaságára nem teszünk feltételt, akkor a hibáról semmit sem tudunk mondani, hiszen az alappontokon kívül a függvényértékek tetszőlegesek lehetnek. Tegyük fel tehát, hogy az f függvény folytonos. Most megemlítünk néhány olyan eredményt, amelyek arra hívják fel a figyelmet, hogy általában nem várható el a pontonkénti konvergencia. Egy

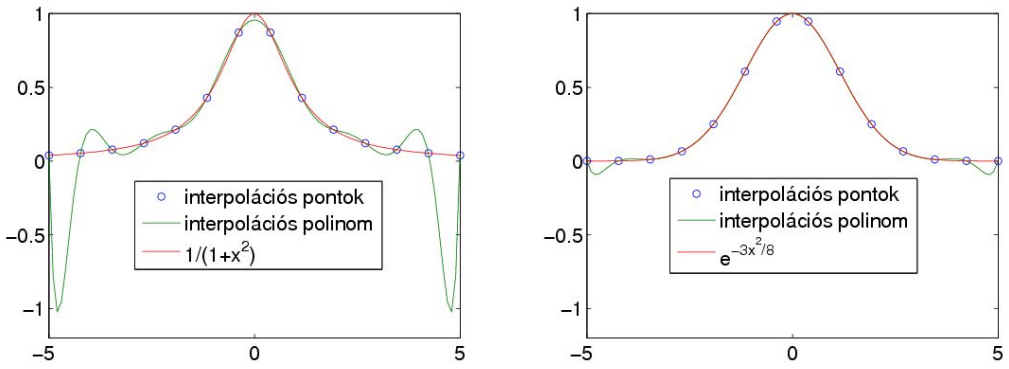
$[a,b]$ intervallumbeli $\{x_0^{(n)},\dots,x_n^{(n)}\}_{n=1}^{\infty}$ alappontsorozaton a továbbiakban egy olyan sorozatot értünk, melynek n-edik eleme az $[a,b]$ intervallum $n+1$ darab páronként különböző pontja.

6.2.2. tétel. (Georg Faber[3], 1914)

Bármely $\{x_0^{(n)},\dots,x_n^{(n)}\}_{n=1}^{\infty}$ $[a,b]$-beli alappontsorozathoz van olyan $f\in C[a,b]$ folytonos függvény, amelyre L_nf nem tart egyenletesen az f függvényhez az $[a,b]$ intervallumon.

1980-ban Erdős és Vértesi [11] megmutatták, hogy bármely alappontsorozathoz olyan f folytonos függvény is létezik, melyre $L_nf(x)\not\to f(x)$ majdnem mindenütt $[a,b]$-ben. Bernstein[4] egy korábbi, 1918-es példája [3] a $[-1,1]$ intervallum ekvidisztáns felosztássorozatához megadott már egy ilyen függvényt: az $f(x)=|x|$ függvény interpolációs polinomjai csak az $x=-1,0$ és 1 pontokban konvergálnak az $f(x)$ függvény megfelelő értékeihez.

Itt gondolhatnánk, hogy a rossz interpolációs viselkedést az okozza, hogy az $f(x)=|x|$ függvény nem deriválható az $x=0$ pontban. Runge[5] 1901-ben észrevette [28], hogy ha az $f(x)=1/(1+x^2)$ függvényt interpoláljuk az ekvidisztáns felosztású $[-5,5]$ intervallumon, akkor az így nyert polinomsorozat csak az $|x|<3.63$ (kerekített érték) feltételnek megfelelő x pontokban fog konvergálni az eredeti függvényhez. A polinomsorozat az intervallumon kívül divergál. Azt, hogy csak egy bizonyos intervallumon belül van konvergencia, az okozza, hogy az $f(z)=1/(1+z^2)$ – most már komplex változósnak tekintett – függvény szinguláris helyei ($\pm i$) közel helyezkednek el az interpolációs intervallumhoz. A 6.2.1. ábrán az $1/(1+x^2)$ Runge-féle függvény és az alakjában hasonló $e^{-3x^2/8}$ függvény 16 ekvidisztáns alappontos interpolációs polinomjának grafikonja látható. Vegyük észre, hogy a második függvényt sokkal jobban közelíti az interpolációs polinom! (Figyeljük meg, hogy az utóbbi függvénynek nincs szinguláris pontja!)



6.2.1. ábra: Az $1/(1+x^2)$ és az $e^{-3x^2/8}$ függvények interpolációs polinomjainak grafikonjai a $[-5,5]$ intervallumból ekvidisztánsan választott 16 alapponton.

A következő tétel azt mutatja, hogy az alappontok megfelelő megválasztásával az egyenletes konvergencia mindig elérhető. Megjegyezzük azonban, hogy nem mindig van lehetőség az alappontok szabad megválasztására.

[3] Georg Faber (1877 – 1966) német matematikus.
[4] Sergei Natanovich Bernstein (1880–1968) orosz matematikus.
[5] Carl David Tolmé Runge (1856–1927), német matematikus.

6.2.3. tétel. (Marcinkiewicz[6] [24], 1936)

Minden, az $[a,b]$ intervallumon folytonos f függvényhez létezik olyan $[a,b]$-beli $\{x_0^{(n)},\dots,x_n^{(n)}\}_{n=1}^{\infty}$ alappontsorozat, amelyen az L_nf interpolációs polinomok sorozata egyenletesen tart az f függvényhez az $[a,b]$ intervallumon.

Megemlítünk egy olyan eredményt, amely azt mutatja, hogy folytonos függvények tetszőlegesen megközelíthetők polinomok segítségével, ha nem követeljük meg az interpolációs feltételt.

6.2.4. tétel. (Weierstrass[7]-féle approximációs tétel [39], 1885)

Legyen $f\in C[a,b]$ egy tetszőleges folytonos függvény. Ekkor tetszőleges $\varepsilon>0$ számhoz van olyan p polinom, melyre $|f(x)-p(x)|<\varepsilon$ minden $x\in[a,b]$ esetén

Megfelelő simaságú függvények esetén az interpolációs hibát az alábbi, Cauchytól származó tétel segítségével számolhatjuk ki.

6.2.5. tétel.

Tegyük fel, hogy az $f\in C^{n+1}(I)$ függvényt interpoláljuk az $x_0,\dots,x_n$ alappontokban, ahol $I=[x_{min},x_{max}]$. Ekkor egy tetszőleges $x\in I$ pontban az interpolációs hiba az

$$E_n(x)=-\frac{f^{(n+1)}(\xi_x)}{(n+1)!}w_{n+1}(x)$$

alakban írható, ahol ξ_x, egy az I intervallum belsejébe eső megfelelő konstans (az x index arra utal, hogy az érték függ az x pont megválasztásától).

Bizonyítás. Ha x egybeesik valamelyik alapponttal, akkor az állítás triviális. Egyébként definiáljuk a $G:I\to\mathbb{R}$,

$$G(t)=E_n(t)-\frac{w_{n+1}(t)}{w_{n+1}(x)}E_n(x)$$

függvényt, amely legalább $n+1$-szer folytonosan deriválható I belsejében, és legalább $n+2$ zérushelye van: $x_0,\dots,x_n$ és x. Ekkor a Rolle-középértéktételt alkalmazva látható, hogy a $G'(t)$ függvénynek legalább $n+1$ zérushelye van. Így haladva mindig az eggyel kisebb deriváltak irányába azt kapjuk, hogy a $G^{(n+1)}(t)$ függvénynek legalább egy zérushelye van. Legyen ez a zérushely ξ_x. Számítsuk ki a $G^{(n+1)}(t)$ deriváltat, felhasználva, hogy egy legfeljebb n-edfokú polinom $(n+1)$-edik deriváltja nulla, és hogy $w_{n+1}^{(n+1)}(x)=(n+1)!$.

$$G^{(n+1)}(t)=-f^{(n+1)}(t)-\frac{(n+1)!}{w_{n+1}(x)}E_n(x),$$

azaz

$$G^{(n+1)}(\xi_x)=-f^{(n+1)}(\xi_x)-\frac{(n+1)!}{w_{n+1}(x)}E_n(x)=0,$$

tehát

$$E_n(x)=-\frac{f^{(n+1)}(\xi_x)}{(n+1)!}w_{n+1}(x).$$

Ezt akartuk bizonyítani. ■

[6] József Marcinkiewicz (1910–1940), lengyel matematikus.
[7] Karl Theodor Wilhelm Weierstrass (1815–1897), német matematikus.

A tétel segítségével elégséges feltételt tudunk adni az egyenletes konvergenciára.

6.2.6. tétel.
Ha $f \in C^\infty[a,b]$ és az $x_0^{(n)}, \dots, x_n^{(n)}$ alappontok mindig az $[a,b]$ intervallumból kerülnek ki $(n = 1, 2, \dots)$, továbbá ha létezik $M > 0$ úgy, hogy $\max_{x\in[a,b]}\{|f^{(n)}(x)|\} \le M^n$, akkor $\|f - L_n f\|_{C[a,b]} \to 0$, ha $n \to \infty$.

Bizonyítás. Az $[a,b]$ intervallum egy tetszőleges x pontjára

$$|E_n(x)| = \frac{|f^{(n+1)}(\xi_x)|}{(n+1)!}|w_{n+1}(x)| \le \frac{M^{n+1}}{(n+1)!}(b-a)^{n+1} \to 0$$

(x-től függetlenül), ha $n \to \infty$. ■

Az interpolációs hiba képletében szerepel a w_{n+1} alappontpolinom. Ennek abszolút értékére ad becslést az alábbi tétel. Jelöljük h-val a szomszédos alappontok közti legnagyobb távolságot, azaz $h := \max_{i=1,\dots,n}\{x_i - x_{i-1}\}$.

6.2.7. tétel.
A $w_{n+1}(x)$ alappontpolinomra érvényes a

$$|w_{n+1}(x)| \le \frac{n!}{4}h^{n+1}$$

becslés, ahol $x \in I$.

Bizonyítás. Ha x alappont, akkor az állítás nyilvánvalóan igaz.
Legyen $x \in (x_0, x_1)$. Ekkor $|(x-x_0)(x-x_1)|$ az $x = (x_0+x_1)/2$ választás esetén a legnagyobb, ami azt jelenti, hogy teljesül az

$$|(x-x_0)(x-x_1)| \le \left|\frac{x_1-x_0}{2}\right| \cdot \left|\frac{x_0-x_1}{2}\right| \le \frac{h^2}{4}$$

becslés. Tehát

$$|w_{n+1}(x)| \le \frac{h^2}{4} \cdot 2h \cdot 3h \cdot \ldots \cdot nh = \frac{h^{n+1}}{4}n!.$$

Legyen most $x \in (x_1, x_2)$. Ekkor $|(x-x_1)(x-x_2)|$ az $x = (x_1+x_2)/2$ választás esetén a legnagyobb, ami azt jelenti, hogy

$$|(x-x_1)(x-x_2)| \le \left|\frac{x_1-x_2}{2}\right| \cdot \left|\frac{x_2-x_1}{2}\right| \le \frac{h^2}{4}.$$

Tehát

$$|w_{n+1}(x)| \le 2h \cdot \frac{h^2}{4} \cdot 2h \cdot 3h \cdot \ldots \cdot (n-1)h = \frac{h^{n+1}}{4}\frac{2n!}{n} \le \frac{h^{n+1}}{4}n!.$$

Hasonlóan járhatunk el a többi intervallumnál is. A becslésekből látható, hogy az I intervallum szélén lévő osztóintervallumokon a felső becslés nagyobb, mint a beljebb lévő intervallumokon. Így a két szélső osztóintervallumon adott becslés lesz érvényes az egész I intervallumra. ■

6.2.8. megjegyzés. Az, hogy az I intervallum szélén lévő osztóintervallumokon az alappolinom értékére adott felső becslés nagyobb, mint a beljebb lévő intervallumokon, azt sugallja, hogy általában az I intervallum szélei közelében nagyobb interpolációs hiba várható, mint beljebb.

Ezen megfigyelés alapján úgy érdemes választani az osztópontokat, hogy azokat az intervallum széleinek közelében sűrűbben vesszük fel. Az alappontok egy lehetséges megválasztásáról szól a következő fejezet. ⋄

6.3. Interpoláció Csebisev-alappontokon

Az interpolációs hiba nagysága függ az alappontpolinom értékétől (6.2.5. tétel). Az alappontpolinom 1 főegyüttható (n + 1)-ed fokú polinom, melynek pontosan (n + 1) darab zérushelye van: az (n + 1) darab alappont. Így a fenti tulajdonságú polinomok és az alappontok között kölcsönösen egyértelmű megfeleltetés van. Most megvizsgáljuk, hogy hogyan kellene megválasztani az alappontpolinomot ahhoz, hogy annak maximumnormája a lehető legkisebb legyen.

6.3.1. tétel.

Legyen $[a,b]$ egy rögzített intervallum. Egy $p^{(1)}$ 1 főegyüttható, $(n+1)$-ed fokú polinomra pontosan akkor igaz, hogy $\|p^{(1)}\|_{C[a,b]} \le \|q^{(1)}\|_{C[a,b]}$ minden $q^{(1)}$ 1 főegyüttható $(n+1)$-ed fokú polinom esetén, ha a $p^{(1)}$ polinom az $[a,b]$ intervallumban rendelkezik legalább $n+2$ különböző abszolút szélsőértékhellyel, a szélsőértékhelyeken a függvényértékek abszolút értékben megegyeznek, és előjelük váltakozik.

Bizonyítás. Tegyük fel, hogy $p^{(1)}$ az $[a,b]$ intervallumon legalább $n+2$ különböző abszolút szélsőértékhellyel rendelkezik, a szélsőértékhelyeken a függvényértékek abszolút értékben megegyeznek, és előjelük váltakozik. Tegyük fel továbbá, hogy egy $q^{(1)} \neq p^{(1)}$ 1 főegyüttható, $(n+1)$-ed fokú polinomra $\|p^{(1)}\|_{C[a,b]} > \|q^{(1)}\|_{C[a,b]}$. Ekkor a $p^{(1)} - q^{(1)}$ legfeljebb n-ed fokú polinomnak $p^{(1)}$ szélsőértékhelyeinél (legalább $(n+2)$ darab) ugyanazok az előjelei, mint a $p^{(1)}$ polinomnak. Ebből következik, hogy a $p^{(1)} - q^{(1)}$ polinomnak legalább $n+1$ zérushelyének kellene lenni, ami ellentmond az algebra alaptételének.

A megfordítás igazolásához tegyük fel indirekt, hogy $p^{(1)}$ olyan 1 főegyüttható $(n+1)$-ed fokú polinom, melyre $\|p^{(1)}\|_{C[a,b]} \le \|q^{(1)}\|_{C[a,b]}$ minden $q^{(1)}$ 1 főegyüttható $(n+1)$-ed fokú polinom esetén. Tegyük fel indirekt, hogy a $p^{(1)}$ polinomnak maximum $n+1$ helyen van abszolút szélsőértéke úgy, hogy ezeken a helyeken a függvényértékek abszolút értékben egyenlők és előjelük váltakozik. Ekkor viszont létezik egy olyan legfeljebb n-edfokú polinom, melynek előjele a szélsőértékhelyeken megegyezik $p^{(1)}$ előjelével. Bármely két szomszédos, ellenkező előjelű szélsőértékhely között választhatunk egy olyan pontot, ahol a $p^{(1)}$ polinomnak zérushelye van. Legyenek ezek a zérushelyek rendre: $x_1 < x_2 < \ldots < x_k$ $(k \le n)$. Ekkor az $s(x) = \pm(x-x_1)\ldots(x-x_k)$ polinom legfeljebb n-edfokú, és alkalmas előjelet választva az előjele a szélsőértékhelyeken megegyezik $p^{(1)}$ előjelével. Ezen polinom megfelelően kicsi $\varepsilon > 0$ számszorosát kivonva $p^{(1)}$-ből olyan 1 főegyüttható legfeljebb $(n+1)$-ed fokú polinom nyerhető, melyre $\|p^{(1)} - \varepsilon s\|_{C[a,b]} < \|p^{(1)}\|_{C[a,b]}$, azaz $p^{(1)}$ nem lehetett az optimálisan közelítő polinom. Ez cáfolja az indirekt feltételt. ■

6.3.2. tétel.

Csak egyetlen olyan $p^{(1)}$ 1 főegyüttható $(n+1)$-ed fokú polinom van az $[a,b]$ intervallumon, melyre $\|p^{(1)}\|_{C[a,b]} \le \|q^{(1)}\|_{C[a,b]}$ minden $q^{(1)}$ 1 főegyüttható $(n+1)$-ed fokú polinom esetén.

Bizonyítás. Tegyük fel indirekt, hogy $p_1^{(1)}$ és $p_2^{(1)}$ is megfelelne a tétel feltételeinek. Ekkor nyilván $\|p_1^{(1)}\|_{C[a,b]} = \|p_2^{(1)}\|_{C[a,b]} =: D$ kell legyen. Ekkor viszont

$$D \le \left\| \frac{p_1^{(1)} + p_2^{(1)}}{2} \right\|_{C[a,b]} \le \frac{\|p_1^{(1)}\|_{C[a,b]} + \|p_2^{(1)}\|_{C[a,b]}}{2} = D$$

miatt $\|(p_1^{(1)}+p_2^{(1)})/2\|_{C[a,b]} = D$, így a $(p_1^{(1)}+p_2^{(1)})/2$ polinom is optimális lenne. Emiatt a polinom legalább $n+2$ helyen venné fel az abszolút szélsőértékét ($\pm D$) váltakozó előjellel, csakúgy, mint $p_1^{(1)}$ és $p_2^{(1)}$ (6.3.1. tétel). Ez csak úgy lehet, ha $p_1^{(1)}$ és $p_2^{(1)}$ is ugyanezeken a helyeken veszi fel a szélsőértékét. De akkor $p_1^{(1)}$ és $p_2^{(1)}$ legalább $n+2$ helyen ugyanazt az értéket veszi fel, így szükségképpen azonos polinomokról van szó. Így ellentmondáshoz jutottunk. ■

Hogyan állíthatjuk elő az optimális polinomokat? Vizsgáljuk a kérdést a $[-1,1]$ intervallumon! Más intervallumra egyszerű változótranszformációval térhetünk át. Jelölje $\tilde{T}_n$ az optimális n-edfokú polinomot. Nyilvánvalóan $\tilde{T}_0(x) = 1$ és $\tilde{T}_1(x) = x$. Emlékezzünk rá, hogy az optimális polinom legalább $(n+2)$-szer vesz fel abszolút szélsőértéket váltakozó előjellel. Innét jön az az ötlet, hogy valamilyen periodikus függvény segítségével állítsuk elő az optimális polinomokat. Az x változó alkalmas ϕ választással $x = \cos\phi$ alakban írható ($\phi \in [0,\pi]$). Így tehát $\tilde{T}_0(x) = \cos(0\phi)$, $\tilde{T}_1(x) = \cos(1\phi)$. Ekkor $\cos(2\phi) = \cos^2\phi - \sin^2\phi = \cos^2\phi - (1-\cos^2\phi) = 2\cos^2\phi - 1 = 2x^2 - 1$. Ez a polinom a konstrukció miatt három váltakozó előjelű abszolút szélsőértékkel rendelkezik, azaz a $\tilde{T}_2(x) = x^2 - 1/2$ választás optimális másodfokú polinomot ad. A következő tétel ezek alapján explicit módon megadja az optimális polinomokat.

6.3.3. tétel.
A

$$\tilde{T}_n(x) = \frac{\cos(n\arccos x)}{2^{n-1}},\ n \geq 1$$

függvény a $[-1,1]$ intervallumon egy 1 főegyütthatós n-edfokú polinom, amelynek $n+1$ váltakozó előjelű abszolút szélsőértékhelye van.

Bizonyítás. Az $n=1$ és $n=2$ választás mellett nyilvánvalóan igaz az állítás. Az is nyilvánvaló, hogy tetszőleges n érték mellett a függvénynek $n+1$ váltakozó előjelű abszolút szélsőértékhelye van. Nevezetesen $\tilde{T}_n(x)$ szélsőértékhelyei a $t_k = \cos(k\pi/n)$ $(k = 0,\dots,n)$ pontokban vannak, hiszen $|\tilde{T}_n(x)| \leq 1/2^{n-1}$ $(|x| \leq 1)$ és $\tilde{T}_n(t_k) = \cos(n\arccos(t_k))/2^{n-1} = \cos(k\pi)/2^{n-1} = \pm 1/2^{n-1}$ (felváltva). Azt kell már csak megmutatnunk, hogy a $\tilde{T}_n(x)$ polinomok 1 főegyütthatósak. Tegyük fel most, hogy $n = k-1$-re és $n = k$-ra igaz az állítás. Mivel

$$\begin{aligned}
2x\cos(k\arccos x) - \cos((k-1)\arccos x) &= 2x\cos(k\arccos x)\\
-(\cos(k\arccos x)x + \sin(k\arccos x)\sin(\arccos x))&\\
= x\cos(k\arccos x) - \sin(k\arccos x)\sin(\arccos x)&\\
= \cos(\arccos x)\cos(k\arccos x) - \sin(k\arccos x)\sin(\arccos x)&\\
= \cos((k+1)\arccos x),&
\end{aligned}$$

ezért

$$\tilde{T}_{k+1}(x) = \frac{\cos((k+1)\arccos x)}{2^k} = x\tilde{T}_k - \frac{\tilde{T}_{k-1}(x)}{4}, \tag{6.3.1}$$

ami pedig egy 1 főegyütthatós $(k+1)$-ed fokú polinom. ■

A (6.3.1) formula egy iterációs képletet definiál arra, hogy hogyan kaphatjuk meg a legjobban közelítő polinomokat. Így pl. $\tilde{T}_3(x) = x\tilde{T}_2(x) - \tilde{T}_1(x)/4 = x(x^2 - 1/2) - x/4 = x^3 - 3x/4$.

Az $(n+1)$-ed fokú, 1 főegyütthatós polinomok közül tehát a $\tilde{T}_{n+1}$ polinom normája lesz a legkisebb a $[-1,1]$ intervallumon. Visszatérve az eredeti problémához ez azt jelenti, hogy az alappontokat úgy kell megválasztanunk, hogy az alappontpolinom éppen a $\tilde{T}_{n+1}$ polinom legyen.

Ez nyilvánvalóan úgy érhető el, ha $\tilde{T}_{n+1}$ zérushelyeit választjuk alappontoknak. A $\tilde{T}_{n+1}$ polinom zérushelyei a következő alakban adhatók meg:

$$x_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right),\ k = 0, \dots, n.$$

6.3.4. definíció.
A $T_0(x) = 1$ és $T_n(x) = 2^{n-1}\tilde{T}_n(x)$ $(n \geq 1)$ polinomokat Csebisev[8]-polinomoknak nevezzük.

6.3.5. tétel.
A Csebisev-polinomok iterációval az alábbi módon állíthatók elő: $T_0(x) = 1$, $T_1(x) = x$, és a többi Csebisev-polinom a $T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x)$ rekurzióval nyerhető. Explicit alakjuk:

$$T_n(x) = \cos(n \arccos x).$$

A Csebisev-polinomoknak a $[-1, 1]$ intervallumban -1 az abszolút minimuma és 1 az abszolút maximuma. $T_n(x)$ főegyütthatója 2^{n-1}.

Bizonyítás. Az első állítás a (6.3.1) rekurzió alakjából következik. A többi állítás a 6.3.3. tétel közvetlen következménye. ■

A fenti elnevezéssel mondhatjuk tehát, hogy az alappontpolinom normája akkor lesz a legkisebb, mégpedig $1/2^n$, ha alappontoknak a Csebisev-polinomok zérushelyeit választjuk. Ebben az esetben az alappontokat Csebisev-alappontoknak hívjuk.

A 6.3.1. ábrán a $T_0, \dots, T_4$ és T_{10} Csebisev-polinomok grafikonjai láthatók a $[-1, 1]$ intervallumon. Figyeljük meg az abszolút szélsőértékhelyek és a zérushelyek elhelyezkedését. Látható, hogy az intervallum szélei közelében sűrűbben vannak a zérushelyek.

6.3.6. megjegyzés. A Csebisev-alappontok nem csak a $[-1, 1]$ intervallumon adhatók meg, hanem az

$$\hat{x}_k = \frac{a+b}{2} + \frac{b-a}{2}x_k$$

transzformációval bármilyen $[a, b]$ intervallumon is. ◇

6.3.7. megjegyzés. Csebisev-alappontokon interpolálva az interpolációs hiba felső becslésére teljesül, hogy

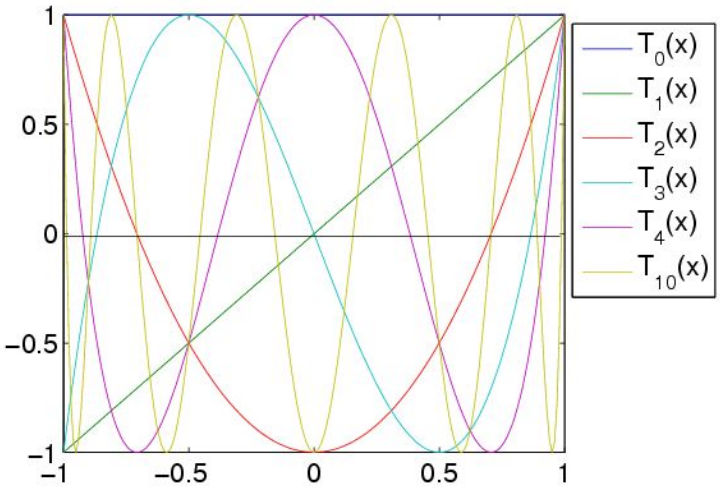
$$|E_n(x)| \leq \frac{M_{n+1}}{(n+1)!2^n},\ x \in I,$$

ahol $M_{n+1} = \max_{x \in I}\{|f^{(n+1)}(x)|\}$. ◇

Bizonyítás nélkül közöljük az alábbi tételt, amely a Csebisev-alappontokon való interpoláció konvergenciájáról szól.

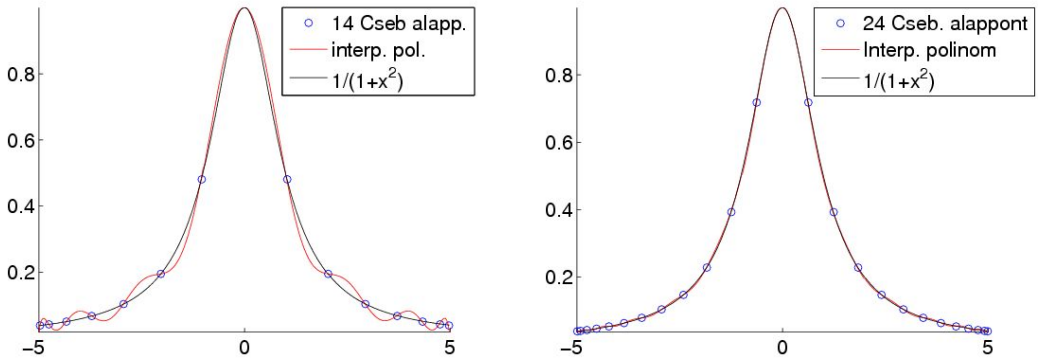
6.3.8. tétel.
A $[a, b]$ intervallumon abszolút folytonos[9] függvények Csebisev-alappontokon vett interpolációs polinomjainak sorozata $C[a, b]$ maximumnormájában tart az eredeti függvényhez, ha az alappontok száma tart a végtelenbe.

[8] Pafnutyij Lvovics Csebisev (1821–1894), orosz matematikus



6.3.1. ábra: A $T_0, \ldots, T_4$ és T_{10} Csebisev-polinomok grafikonjai a $[-1, 1]$ intervallumon.

A tétel következménye tehát pl., hogy az $|x|$ függvényt vagy a Runge-féle $1/(1+x^2)$ függvényt a Csebisev-alappontokon interpolálva az eredeti függvényhez konvergáló polinomsorozatot kapunk. A 6.3.2. ábrán a Runge-féle $1/(1+x^2)$ függvény interpolációs polinomjainak grafikonjait láthatjuk 14 ill. 24 Csebisev-alapponttot választva. Látható, hogy 24 alappont esetén az interpolációs polinom grafikonja már alig különböztethető meg az eredeti függvényétől. Érdemes az ábrát összevetni a 6.2.1. ábra bal oldali grafikonjával.



6.3.2. ábra: A Runge-féle $1/(1+x^2)$ függvény interpolációja 14 ill. 24 Csebisev-alapponton a [-5,5] intervallumon.

6.3.9. megjegyzés. Természetesen a Faber-tétel (6.2.2. tétel) miatt a Csebisev-alappontok esetén

[9] Egy $f : [a, b] \to \mathbb{R}$ függvényt abszolút folytonosnak hívunk, ha minden $\varepsilon > 0$ esetén van olyan $\delta > 0$, hogy ha $a \le a_1 < b_1 \le a_2 < b_2 \le \ldots \le a_m < b_m \le b$ és $\sum_{k=1}^{m} |b_i - a_i| < \delta$, akkor $\sum_{k=1}^{m} |f(a_i) - f(b_i)| < \varepsilon$. Ha egy függvény teljesíti a Lipschitz-tulajdonságot (van olyan $L \ge 0$ szám, hogy $|f(x) - f(y)| \le K|x - y|$ minden $x, y \in [a, b]$ esetén), akkor abszolút folytonos is.

is lesz olyan folytonos függvény, melynél nincs egyenletes konvergencia. ◇

6.4. Hermite-interpoláció

Az előző fejezetekben azt vizsgáltuk, hogy ha ismert egy függvény $n+1$ alapontbeli értéke, akkor hogyan állíthatunk elő egy olyan polinomot, amely ugyanazon alappontokban ugyanazon értékeket veszi fel. Most általánosításként vizsgáljuk azt az esetet, amikor az alappontokban ismerjük még az eredeti függvény deriváltjait is bizonyos rendig bezáróan. Azaz legyenek adottak az $x_0, \dots, x_n$ különböző alappontok, és az x_k $(k = 0, \dots, n)$ pontban legyen adott $m_k + 1$ darab számérték: $f_k^{(0)}, f_k^{(1)}, \dots, f_k^{(m_k)}$. Keressünk olyan H polinomot, melyre

$$H^{(i)}(x_k) = f_k^{(i)}, \ (k = 0, \dots n, \ i = 0, \dots, m_k).$$

Ezt az eljárást *Hermite-féle interpolációnak* nevezzük. Amennyiben speciálisan minden pontban csak a függvényértékek és az első deriváltak adottak, akkor *Hermite–Fejér-interpolációról*[10] beszélünk. Mivel összesen $N := m_0 + \dots + m_n + n + 1$ adat adott, így várható, hogy egy legfeljebb $(N-1)$-ed fokú polinom megfelel a feltételeknek.

6.4.1. tétel.
Egyértelműen létezik egy olyan H_{N-1} legfeljebb $(N-1)$-ed fokú polinom, amely teljesíti a

$$H_{N-1}^{(i)}(x_k) = f_k^{(i)}, \ k = 0, \dots, n, \ i = 0, \dots, m_k$$

feltételeket.

Bizonyítás. Legyen $H_{N-1}(x) = a_0 + a_1 x + \dots + a_{N-1}x^{N-1}$ alakú. Ekkor az együtthatók meghatározásához az alábbi egyenletrendszert kell megoldanunk:

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^{N-1} \\ 0 & 1 & 2x_0 & \dots & (N-1)x_0^{N-2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_1 & x_1^2 & \dots & x_1^{N-1} \\ 0 & 1 & 2x_1 & \dots & (N-1)x_1^{N-2} \\ \vdots & \vdots & \vdots & \dots & \vdots \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} f_0^{(0)} \\ f_0^{(1)} \\ \vdots \\ \vdots \\ f_1^{(0)} \\ f_1^{(1)} \\ \vdots \\ \vdots \end{bmatrix}.$$

Itt N egyenlet van és N ismeretlen, és az együtthatómátrix nemszinguláris. Ugyanis ha létezne olyan nemnulla vektor, mellyel a mátrixot szorozva nullvektort kapnánk, akkor a H_{N-1} polinomnak N zérushelye lenne, ami lehetetlen. Itt felhasználtuk azt, hogy ha egy polinomnak x k-szoros zérushelye, akkor a polinom első $k-1$ deriváltjának is zérushelye x. ■

A továbbiakban csak az Hermite–Fejér-interpolációval foglalkozunk, azaz azzal az esettel, amikor minden alappontban csak a függvényértéke és az első deriváltja adott a keresett polinomnak

[10] Fejér Lipót (1880 – 1959), magyar matematikus. Bővebb életrajz a http://www.omikk.bme.hu/archivum/magyarok/htm/fejerrov.htm címen található.

$(f_k^{(0)}, f_k^{(1)}, k = 0, \dots, n)$. Ha $n+1$ alappontunk van, akkor ez $2n+2$ adatot jelent, így egy legfeljebb $(2n+1)$-ed fokú polinom lesz az interpolációs polinom. Jelölje H_{2n+1} a keresett polinomot. Az Hermite–Fejér-interpolációs polinom előállítható úgy, hogy az előző tételben szereplő egyenletrendszert megoldjuk. Ezt a gyakorlatban általában nem alkalmazzuk, mert vannak módszerek, melyek műveletszáma lényegesen kisebb.

6.4.2. tétel.
Az Hermite–Fejér-interpolációs polinom a

$$H_{2n+1}(x) = \sum_{k=0}^{n} f_k^{(0)}(1 - 2(x - x_k)l_k'(x_k))l_k^2(x) + \sum_{k=0}^{n} f_k^{(1)}(x - x_k)l_k^2(x)$$

képlettel állítható elő, ahol l_k a k-adik alapponthoz tartozó Lagrange-féle alappolinom.

Bizonyítás. Hasonlóan az interpolációs polinom Lagrange-féle előállításához előállítunk olyan polinomokat, melyek $(2n+1)$-ed fokúak és vagy az értékük vagy a deriváltjuk 1 valamelyik alappontban, és a többi értékük és deriváltjuk pedig 0 a többi alappontban. Jelölje h_{k0} azt az alappolinomot, melyre

$$h_{k0}(x_l) = \delta_{kl}, \; h_{k0}'(x_l) = 0,$$

és h_{k1} azt az alappolinomot, melyre

$$h_{k1}'(x_l) = \delta_{kl}, \; h_{k1}(x_l) = 0,$$

ahol $k, l = 0, \dots, n$ és $i = 0, 1$. δ_{kl} a szokásos Kronecker-szimbólum. Keressük az alappolinomokat $h(x) = s(x)l_k^2(x)$ alakban, ahol $s(x)$ egy meghatározandó elsőfokú polinom. Mivel $h'(x) = s'(x)l_k^2(x) + 2s(x)l_k(x)l_k'(x)$, ezért ha h-tól azt várjuk el, hogy értéke és deriváltja minden x_k-tól különböző alappontban legyen nulla, x_k-ban az értéke legyen 0 és deriváltja 1, akkor az $s(x_k) = 0$ és $s'(x_k) = 1$ feltételeknek kell teljesülniük, azaz $s(x) = x - x_k$ megfelelő választás. Ebből következik, hogy $h_{k1}(x) = (x - x_k)l_k^2(x)$ $(k = 0, \dots, n)$. Amennyiben h-tól azt várjuk el, hogy értéke és deriváltja minden x_k-tól különböző alappontban legyen nulla, x_k-ban az értéke legyen 1 és deriváltja 0, akkor az $s(x_k) = 1$ és $s'(x_k) = -2l_k'(x_k)$ feltételeknek kell teljesülniük. Tehát az $s(x) = 1 - 2(x - x_k)l_k'(x_k)$ jó választás. Így $h_{k0}(x) = (1 - 2(x - x_k)l_k'(x_k))l_k^2(x)$. A keresett interpolációs polinom tehát a

$$H_{2n+1}(x) = \sum_{k=0}^{n} f_k^{(0)}h_{k0}(x) + \sum_{k=0}^{n} f_k^{(1)}h_{k1}(x)$$

alakban írható. Ezt akartuk megmutatni. ■

Az interpolációs hibánál ismertetett módon igazolható az alábbi tétel az Hermite–Fejér-interpoláció hibájával kapcsolatban.

6.4.3. tétel.
Tegyük fel, hogy az $f \in C^{2n+2}(I)$ függvényt interpoláljuk a Hermite–Fejér-interpolációval az $x_0, \ldots, x_n$ alappontokban, ahol $I = [x_{\min}, x_{\max}]$. Ekkor egy tetszőleges $x \in I$ pontban az interpolációs hiba az

$$E_n(x) = H_{2n+1}(x) - f(x) = -\frac{f^{(2n+2)}(\xi_x)}{(2n+2)!} w_{n+1}^2(x)$$

alakban írható, ahol ξ_x egy, az I intervallum belsejébe eső megfelelő konstans (az x index arra utal, hogy értéke függ az x pont megválasztásától).

6.5. Szakaszonként polinomiális interpoláció

Ha az interpoláció során az alappontok adottak (pl. mérési eredményekből származnak), akkor nem lehet a Csebisev-alappontokat használni. Ez nagy interpolációs hibához vezethet. Az is látható, hogy az interpolációs polinom általában nem úgy köti össze az adott pontokat, ahogy azokat szabad kézzel összekötnénk: pl. lehet, hogy egymás melletti pontokhoz ugyanaz az ordináta tartozik, az interpolációs polinom értéke mégis nagyon eltérhet ettől az értéktől ezen a szakaszon. Ha a pontok monoton növő ordinátával rendelkeznek egy szakaszon, ez nem vonja maga után, hogy az interpolációs polinom is monoton növő lesz a megfelelő szakaszon. Ezeket a hibákat küszöbölhetjük ki, ha szakaszonként polinomiális interpolációt alkalmazunk. Ekkor a szomszédos alappontok közti szakaszokon, egy-egy alacsony fokszámú polinommal interpolálunk. Ezt az interpolációs módot *spline*[11]*-interpolációnak* nevezzük.

6.5.1. Szakaszonként lineáris interpoláció

A legegyszerűbb spline-interpoláció a szakaszonként lineáris interpoláció. Ilyenkor a szomszédos pontokat egyenes szakaszokkal kötjük össze. Jelöljük most is az alappontokat növekvő sorrendben az $x_0 < x_1 < \ldots < x_n$ módon, az egyes pontokbeli függvényértékeket jelölje $f_0, \ldots, f_n$, és az $[x_{k-1}, x_k]$ szakaszon az interpolációs spline-polinom legyen s_k $(k = 1, \ldots, n)$. Szakaszonként lineáris spline esetén nyilvánvalóan

$$s_k(x) = f_{k-1}\frac{x - x_k}{x_{k-1} - x_k} + f_k\frac{x - x_{k-1}}{x_k - x_{k-1}}, \ x \in [x_{k-1}, x_k].$$

A teljes intervallumon értelmezett interpolációs függvény ekkor

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

alakban adható meg. Ez egy folytonos függvény, hiszen $s_k(x_k) = s_{k+1}(x_k) = f_k$.

[11] Angol szó, vékony, lapos, hajlítható fa vagy fém csík, melyet görbék rajzolására használunk.

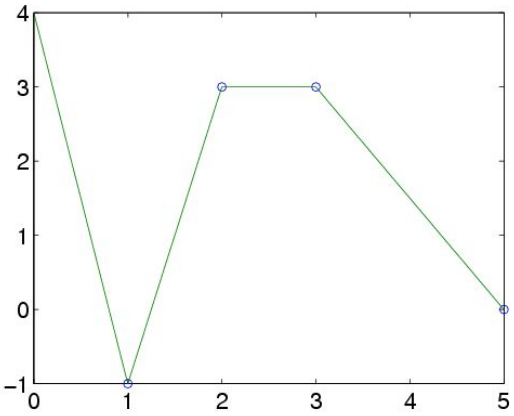
6.5.1. tétel.
Legyen $f \in C^2(I)$. Ekkor a lineáris spline-interpolációs függvény hibája

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

ahol M_2 egy felső korlát f második deriváltjára az I intervallumon, és h a szomszédos alappontok közötti maximális távolság.

Bizonyítás. A tétel a 6.2.5. és a 6.2.7. tételek szakaszonkénti alkalmazásából adódik. ■

A lineáris spline-interpoláció megvalósítja azt a követelményt, hogy monoton koordináták esetén az interpolációs függvény is monoton. Hátránya viszont, hogy a teljes $[x_0, x_n]$ intervallumra nyert interpolációs függvény nem lesz deriválható (lásd a 6.5.1. ábrát).



6.5.1. ábra: A (0,4), (1,-1) (2,3) (3,3) (5,0) pontokhoz tartozó lineáris spline-interpolációs függvények grafikonjai.

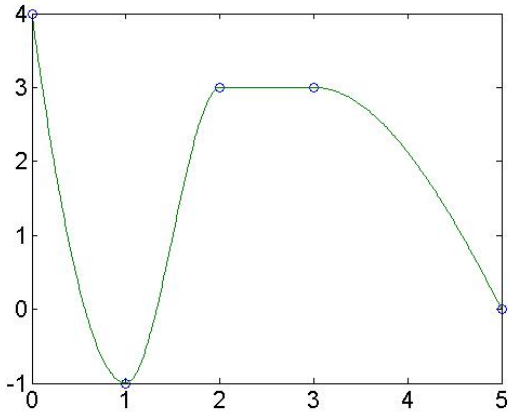
6.5.2. Szakaszonként kvadratikus interpoláció

A szakaszonként lineáris interpoláció egyik hátránya, hogy az így nyert s interpolációs függvény ugyan folytonos, de a deriváltja már nem. Ezt küszöbölhetjük ki másodfokú polinomok alkalmazásával. Válasszuk meg az s interpolációs függvény deriváltjának értékét az x_0 pontban. Legyen ez d_0. Ekkor az első $[x_0, x_1]$ szakaszon Hermite-interpolációt végrehajtva egyértelműen meghatározott egy olyan s_1 legfeljebb másodfokú polinom, melyre $s_1(x_0) = f_0$, $s_1'(x_0) = d_0$ és $s_1(x_1) = f_1$. Válasszuk d_1-nek az $s_1'(x_1)$ értéket. Ezután az $[x_1, x_2]$ szakaszon Hermite-interpolációt végrehajtva egyértelműen meghatározott egy olyan s_2 legfeljebb másodfokú polinom, melyre $s_2(x_1) = f_1$, $s_2'(x_1) = d_1$ és $s_2(x_2) = f_2$. A többi intervallumon hasonlóan eljárva az eredő s függvény és deriváltja is folytonos lesz az egész intervallumon.

6.5.3. Szakaszonként harmadfokú interpoláció

Tekintsük most azt az esetet, amikor minden intervallumon egy legfeljebb harmadfokú polinomot alkalmazunk az interpolációra. Ekkor, ha megadjuk az $x_0, \ldots, x_n$ pontokban a $d_0, \ldots, d_n$ értékeket, akkor minden részintervallum mindkét végpontjában adott egy-egy függvényérték és egy-egy

deriváltérték. Ezek egyértelműen meghatároznak egy legfeljebb harmadfokú polinomot (Hermite–Fejér-interpoláció). Mivel a deriváltakat tetszőlegesen választhatjuk, ezért elérhető, hogy az interpolációs függvény monoton legyen, ha a pontok koordinátái monotonok. Ilyen interpolációra láthatunk példát a 6.5.2. ábrán.



6.5.2. ábra: A (0,4), (1,-1) (2,3) (3,3) (5,0) pontokhoz tartozó, szakaszonként harmadfokú, monotonitást megőrző interpoláció.

Ez az eljárás ismét csak folytonosan deriválható interpolációs függvényt ad. Látni fogjuk, hogy a $d_0, \dots, d_n$ deriváltértékek alkalmas megválasztásával elérhető az is, hogy a másodrendű deriváltak is folytonosak legyenek.

Legyen s_k az $[x_{k-1}, x_k]$ intervallumon adott legfeljebb harmadfokú polinom ($k = 0, \dots, n$). Ekkor adott d_{k-1} és d_k deriváltértékekkel s_k az Hermite–Fejér-interpoláció segítségével adható meg.

alappontok	$f_i = [.]f$	$[.,.]f$	$[.,.,.]f$	$[.,.,.,.]f$
x_{k-1}	$f_{k-1} =: c_{k0}$			
		$d_{k-1} =: c_{k1}$		
x_{k-1}	f_{k-1}		$\frac{\frac{f_k-f_{k-1}}{x_k-x_{k-1}}-d_{k-1}}{x_k-x_{k-1}} =: c_{k2}$	
		$\frac{f_k-f_{k-1}}{x_k-x_{k-1}}$		$\frac{\frac{d_k-\frac{f_k-f_{k-1}}{x_k-x_{k-1}}}{x_k-x_{k-1}}-\frac{\frac{f_k-f_{k-1}}{x_k-x_{k-1}}-d_{k-1}}{x_k-x_{k-1}}}{x_k-x_{k-1}} =: c_{k3}$
x_k	f_k		$\frac{d_k-\frac{f_k-f_{k-1}}{x_k-x_{k-1}}}{x_k-x_{k-1}}$	
		d_k		
x_k	f_k			

A táblázatban szereplő jelölésekkel s_k az

$$s_k(x) = c_{k0} + c_{k1}(x - x_{k-1}) + c_{k2}(x - x_{k-1})^2 + c_{k3}(x - x_{k-1})^2(x - x_k)$$

alakban írható. Itt a c_{ki} együtthatók a k. intervallumon definiált s_k polinom megfelelő együtthatóit jelentik. Hasonlóan kaphatjuk az s_{k+1} polinomot az $[x_k, x_{k+1}]$ intervallumon

$$s_{k+1}(x) = c_{k+1,0} + c_{k+1,1}(x - x_{k-1}) + c_{k+1,2}(x - x_{k-1})^2 + c_{k+1,3}(x - x_{k-1})^2(x - x_k)$$

alakban. Mivel

$$s_k''(x) = 2c_{k2} + 2c_{k3}(x - x_k) + 4c_{k3}(x - x_{k-1})$$

és

$$s_{k+1}''(x) = 2c_{k+1,2} + 2c_{k+1,3}(x - x_{k+1}) + 4c_{k+1,3}(x - x_k),$$

ahhoz, hogy a második derivált is folytonos legyen az x_k pontban, a

$$2c_{k2} + 4c_{k3}(x_k - x_{k-1}) = 2c_{k+1,2} + 2c_{k+1,3}(x_k - x_{k+1})$$

egyenlőségnek kell teljesülni. Az együtthatók behelyettesítése után a

$$\frac{2}{x_k - x_{k-1}} d_{k-1} + 4\left(\frac{1}{x_k - x_{k-1}} + \frac{1}{x_{k+1} - x_k}\right) d_k + \frac{2}{x_{k+1} - x_k} d_{k+1} \tag{6.5.1}$$

$$= 6\left(\frac{f_k - f_{k-1}}{(x_k - x_{k-1})^2} + \frac{f_{k+1} - f_k}{(x_{k+1} - x_k)^2}\right)$$

egyenlethez jutunk ($k = 1, \ldots, n-1$). Amennyiben a $d_0, \ldots, d_n$ deriváltértékek teljesítik a (6.5.1) feltételeket, akkor a második derivált folytonos lesz a teljes intervallumon. Mivel $n-1$ egyenletünk van és $n+1$ ismeretlen deriváltérték, így várható, hogy még két plusz feltételt is előírhatunk. Legyenek ezek az intervallum két szélén előírt második deriváltak. Jelölje őket D_0 és D_n.

Ha $k = 1$, akkor

$$s_1''(x_0) = 2c_{12} + 2c_{13}(x_0 - x_1) = D_0,$$

ahonnét a

$$\frac{2}{x_1 - x_0} d_0 + \frac{1}{x_1 - x_0} d_1 = 3\frac{f_1 - f_0}{(x_1 - x_0)^2} - \frac{D_0}{2} \tag{6.5.2}$$

egyenlőséget kapjuk.

Ha $k = n$, akkor

$$s_n''(x_n) = 2c_{n2} + 4c_{n3}(x_n - x_{n-1}) = D_n.$$

Így az

$$\frac{1}{x_n - x_{n-1}} d_{n-1} + \frac{2}{x_n - x_{n-1}} d_n = 3\frac{f_n - f_{n-1}}{(x_n - x_{n-1})^2} + \frac{D_n}{2} \tag{6.5.3}$$

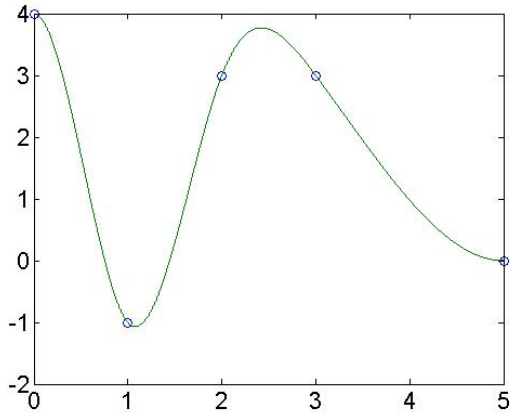
egyenlethez jutunk.

A (6.5.1), (6.5.2), (6.5.3) egyenletekből álló egyenletrendszert megoldva megkaphatjuk a $d_0, \ldots, d_n$ deriváltértékeket, melyek biztosítják a második derivált folytonosságát és az x_0 és x_n pontokban a második derivált D_0 ill. D_n értékét.

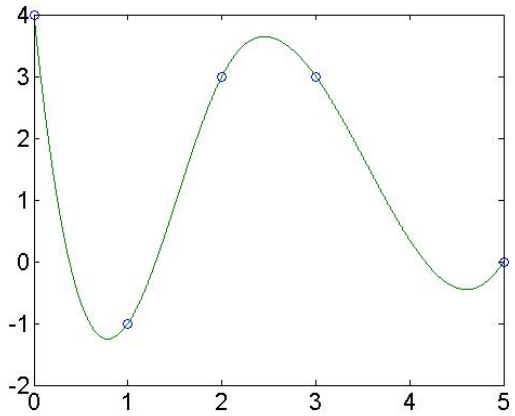
Természetesen mivel $n+1$ ismeretlen deriváltértékünk van, és a második derivált folytonossága csak $n-1$ egyenletet igényel, így az egyértelműséghez szükséges plusz két feltétel nem csak úgy biztosítható, hogy előírjuk a második deriváltat a végpontokban. Szokás az is, hogy magát a deriváltértéket írjuk elő ezekben a pontokban.

A (0,4), (1,-1) (2,3) (3,3) (5,0) pontokhoz tartozó legfeljebb harmadfokú spline-interpolációs függvény grafikonját a 6.5.3. ábrán láthatjuk. Ebben az esetben azt írtuk elő a végpontokban, hogy a derivált legyen nulla. A 6.5.4. ábrán pedig szintén egy legfeljebb harmadfokú spline-interpoláció látható, de itt a végpontokban azt írtuk elő, hogy a második derivált legyen nulla.

Amennyiben az alappontok ekvidisztáns módon helyezkednek el, azaz $x_{k+1} - x_k = h$ minden $k = 0, \ldots, n-1$ esetén, akkor a (6.5.1), (6.5.2), (6.5.3) egyenletekből álló egyenletrendszer az



6.5.3. ábra: A (0,4), (1,-1) (2,3) (3,3) (5,0) pontokhoz tartozó harmadfokú spline-interpolációs függvény grafikonja. A derivált a végpontokban nulla.



6.5.4. ábra: A (0,4), (1,-1) (2,3) (3,3) (5,0) pontokhoz tartozó harmadfokú spline interpolációs függvény grafikonja. A második derivált a végpontokban nulla.

alábbi áttekinthető formában írható:

$$\frac{h}{3}\begin{bmatrix} 2 & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & & & & & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_n \end{bmatrix} = \begin{bmatrix} f_1 - f_0 - D_0 h^2/6 \\ f_2 - f_0 \\ f_3 - f_1 \\ \vdots \\ f_n - f_{n-1} + D_n h^2/6 \end{bmatrix}.$$

Ha a két szélső pontban a deriváltak adottak, akkor az egyenletrendszer

$$\frac{h}{3}\begin{bmatrix} 4 & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & & & & & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 4 \end{bmatrix}\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \end{bmatrix} = \begin{bmatrix} -d_0 h/3 \\ f_2 - f_0 \\ f_3 - f_1 \\ \vdots \\ -d_n h/3 \end{bmatrix}$$

alakú lesz. Mindkét esetben az egyenletrendszer mátrixa invertálható, így egyértelmű megoldást kapunk a deriváltértékekre, amik pedig egyértelműen meghatározzák az egyes szakaszokon a legfeljebb harmadfokú polinomokat.

A következő tétel a szakaszonként harmadfokú spline-interpolációs függvény egy extremális tulajdonságáról szól.

6.5.2. tétel.

Legyenek adottak az (x_i, f_i), $(i = 0, \dots, n)$ pontok, és legyen s a hozzájuk tartozó szakaszonként legfeljebb harmadfokú spline-függvény, amelynek a végpontokban vett második deriváltjai (D_0 és D_n) nullák. Ekkor

$$\int_I (s''(x))^2 \,\mathrm{d}x \le \int_I (g''(x))^2 \,\mathrm{d}x$$

minden olyan $g \in C^2(I)$ függvény esetén, amelyek interpolálják az adott pontokat.

Bizonyítás. Nyilvánvalóan $s \in C^2(I)$. Legyen $g \in C^2(I)$ tetszőleges olyan függvény, amely interpolálja az adott pontokat. Ekkor azt kell megmutatnunk, hogy

$$\int_{x_0}^{x_n} (g''(x))^2 - (s''(x))^2 \,\mathrm{d}x \ge 0.$$

Könnyen ellenőrizhető, hogy

$$\int_{x_0}^{x_n} (g''(x))^2 - (s''(x))^2 \,\mathrm{d}x = \int_{x_0}^{x_n} (g''(x) - s''(x))^2 \,\mathrm{d}x - 2\int_{x_0}^{x_n} s''(x)(s''(x) - g''(x)) \,\mathrm{d}x.$$

A jobb oldali első tag nyilván nemnegatív. A második tagról pedig megmutatjuk, hogy nulla az értéke. Ez igazolja az állításunkat. Számoljuk ki az integrált részintervallumonkénti parciális integrálással, kihasználva, hogy a részintervallumokon s végtelen sokszor deriválhat!

$$\int_{x_0}^{x_n} s''(x)(s''(x) - g''(x)) \,\mathrm{d}x = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} s''(x)(s''(x) - g''(x)) \,\mathrm{d}x$$

$$= \sum_{k=1}^{n} \left([s''(x)(s'(x) - g'(x))]_{x_{k-1}}^{x_k} - \int_{x_{k-1}}^{x_k} s'''(x)(s'(x) - g'(x)) \,\mathrm{d}x \right)$$

$$= \sum_{k=1}^{n} \left([s''(x)(s'(x) - g'(x))]_{x_{k-1}}^{x_k} - \left([s'''(x)(s(x) - g(x))]_{x_{k-1}}^{x_k} - \underbrace{\int_{x_{k-1}}^{x_k} s''''(x)(s(x) - g(x)) \,\mathrm{d}x}_{=0} \right) \right)$$

$$= \sum_{k=1}^{n} \Big([s''(x)(s'(x) - g'(x))]_{x_{k-1}}^{x_k} - \Big([s'''(x)(s(x) - g(x))]_{x_{k-1}}^{x_k} \Big) \Big)$$
$$= \underbrace{s''(x_n)}_{=D_0=0}(s'(x_n) - g'(x_n)) - \underbrace{s''(x_0)}_{=D_n=0}(s'(x_0) - g'(x_0)) = 0.$$

Az utolsó sor egyrészt abból következik, hogy az utolsó előtti sor második tagjában szereplő $s(x) - g(x)$ tényező minden alappontban nulla, hiszen mindkét függvény grafikonja átmegy az adott pontokon. Másrészt az első tagbeli összegből csak az utolsó és első tagok különbsége marad meg, hiszen a függvények második deriváltja folytonos az intervallumon. Ezzel igazoltuk a tétel állítását. ■

6.6. Trigonometrikus interpoláció

Az alkalmazásokban (pl. digitális jelfeldolgozás, képfeldolgozás) gyakran találkozunk azzal a feladattal, hogy egy adott periodikus függvényt trigonometrikus összetevőkre kell bontanunk. Egy 2π szerint periodikus $f(x)$ függvénynek kereshetjük pl. az

$$f(x) = \alpha_0 + \sum_{j=1}^{\infty} (\alpha_j \cos(jx) + \beta_j \sin(jx))$$

alakú előállítását, ahol $\alpha_0, \alpha_j, \beta_j$ $(j = 1, 2, \ldots)$ megfelelő konstansok. A fenti előállítást az $f(x)$ periodikus függvény *trigonometrikus sorának* nevezzük. Ismert hogy ahhoz, hogy a Fourier-sor az $f(x)$ függvényhez konvergáljon, az együtthatókat az

$$\begin{aligned} \alpha_0 &= \frac{1}{2\pi} \int_0^{2\pi} f(x)\,\mathrm{d}x, \\ \alpha_j &= \frac{1}{\pi} \int_0^{2\pi} f(x)\cos(jx)\,\mathrm{d}x, \\ \beta_j &= \frac{1}{\pi} \int_0^{2\pi} f(x)\sin(jx)\,\mathrm{d}x \end{aligned} \tag{6.6.1}$$

módon kell megválasztanunk, így nyerjük az ún. Fourier-sort. A fenti integrálokat csak nagyon speciális $f(x)$ függvények esetén tudjuk kiszámolni. További nehézség, hogy gyakran az $f(x)$ függvényt sem ismerjük pontosan, hanem csak néhány (leggyakrabban ekvidisztáns) alappontban ismerjük az értékét.

A fenti problémákat úgy hidalhatjuk át, hogy először az $f(x)$ függvény adott alappontokban vett értékeire illesztett trigonometrikus polinomot határozzuk meg – ezt az eljárást *trigonometrikus interpolációnak* nevezzük, majd ezután ezzel közelítjük az $f(x)$ függvény Fourier-sorát. Interpolációs feladatok során akkor is érdemes lehet trigonometrikus interpolációt használni, ha tudjuk, hogy az adatok periodikus függvényről származnak.

Tegyük fel tehát, hogy egy 2π periódusú függvénynek ismerjük az $f_k = f(x_k)$ értékeit az $x_k = 2\pi k/(n+1) \in [0, 2\pi)$ pontokban $(k = 0, \ldots, n)$, ahol n egy pozitív egész szám. Keressük azt a

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

alakú ún. m-edfokú (ha $|a_m| + |b_m| \neq 0$) trigonometrikus polinomot, amelyre $t_m(x_k) = f_k$ $(k = 0, \ldots, n)$. Az ismeretlen $a_0, a_1, \ldots, a_m, b_1, \ldots, b_m$ együtthatókat *diszkrét Fourier-együtthatóknak* nevezzük. Összesen $n+1$ egyenletet kell kielégíteni $2m+1$ diszkrét Fourier-együtthatóval. Ha

n páros, akkor várható, hogy $m = n/2$ fokú polinom megfelelő lesz. Ha n páratlan, akkor $m = (n+1)/2$-ed fokú polinomra lesz várhatóan szükségünk. Ekkor viszont $n+2$ együtthatónk és $n+1$ egyenletünk lesz, azaz a rendszer alulhatározott. Vegyük észre, hogy

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

azaz a b_m együtthatóhoz tartozó $\sin(mx)$ függény az alappontokban nullát vesz fel, így b_m értéke nem befolyásolja az interpolációt. Így b_m értékét választhatjuk nullának. Az ilyen trigonometrikus polinomokat (páratlan n esetén) *kiegyensúlyozott trigonometrikus polinomoknak* nevezzük. Azt, hogy a fent kitalált fokszámú polinomokkal meg is valósítható az interpoláció, az alábbi tételek mutatják.

6.6.1. tétel.

Tegyük fel, hogy az $x_k = 2\pi k/(n+1)$ alappontokban adottak az $f_k \in \mathbb{R}$ értékek $(k = 0, \dots, n)$. Tegyük fel, hogy n páratlan. Ekkor egyértelműen létezik egy olyan $m = (n+1)/2$-ed fokú kiegyensúlyozott t_m trigonometrikus polinom, melyre $t_m(x_k) = f_k$ $(k = 0, \dots, n)$. A valós diszkrét Fourier együtthatók az alábbi módon számolhatók:

$$a_0 = \frac{1}{n+1}\sum_{k=0}^{n} f_k, \quad a_m = \frac{1}{n+1}\sum_{k=0}^{n} f_k \cos(mx_k),$$

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

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

Bizonyítás. A tételt úgy igazoljuk, hogy előállítjuk explicit módon az interpolációs polinomot. Közben látni fogjuk, hogy az előállítás egyértelmű. A kompex számok Euler-alakját használva

$$e^{\mathtt{i}jx} = \cos(jx) + \mathtt{i}\sin(jx),$$
$$e^{-\mathtt{i}jx} = \cos(jx) - \mathtt{i}\sin(jx),$$

ahonnét azt kapjuk, hogy

$$\cos(jx) = \frac{e^{\mathtt{i}jx} + e^{-\mathtt{i}jx}}{2} \quad \text{és} \quad \sin(jx) = \frac{e^{\mathtt{i}jx} - e^{-\mathtt{i}jx}}{2\mathtt{i}}.$$

Az így kapott kifejezéseket helyettesítsük vissza az eredeti t_m polinomba.

$$\begin{aligned} t_m(x) &= a_0 + \sum_{j=1}^{m}\left(a_j\frac{e^{\mathtt{i}jx}+e^{-\mathtt{i}jx}}{2} + b_j\frac{e^{\mathtt{i}jx}-e^{-\mathtt{i}jx}}{2\mathtt{i}}\right) \\ &= \overbrace{a_0}^{=:c_0} + \sum_{j=1}^{m-1}\left(\overbrace{\frac{a_j - b_j\mathtt{i}}{2}}^{=:c_j} e^{\mathtt{i}jx} + \overbrace{\frac{a_j + b_j\mathtt{i}}{2}}^{=:c_{-j}} e^{-\mathtt{i}jx}\right) + \overbrace{\frac{a_m}{2}}^{=:c_m/2} e^{\mathtt{i}mx} + \overbrace{\frac{a_m}{2}}^{=:c_{-m}/2} e^{-\mathtt{i}mx} \\ &= \sum_{j=-(m-1)}^{m-1} c_j e^{\mathtt{i}jx} + \frac{c_m}{2}e^{\mathtt{i}mx} + \frac{c_{-m}}{2}e^{-\mathtt{i}mx}. \end{aligned}$$

Vezessük be a fenti képletekben szereplő egyes együtthatókra a

$$\begin{aligned} c_0 &= a_0, \\ c_j &= \frac{a_j - b_j \mathtt{i}}{2}, \\ c_{-j} &= \frac{a_j + b_j \mathtt{i}}{2}, \\ c_m &= c_{-m} = a_m \end{aligned} \tag{6.6.2}$$

jelöléseket. Ezeket a c_j $(j = -m, \dots, m)$ együtthatókat komplex Fourier-együtthatóknak nevezzük. Az eredeti valós együtthatók is könnyen előállíthatók a c_j együtthatókkal:

$$\begin{aligned} a_0 &= c_0, \\ a_m &= c_m, \\ a_j &= c_j + c_{-j}, \ (j = 1, \dots, m-1), \\ b_j &= (c_j - c_{-j})\mathtt{i} \ (j = 1, \dots, m-1). \end{aligned} \tag{6.6.3}$$

A trigonometrikus interpolációs polinom ezen átalakítása után térjünk át a $t_m(x_k) = f_k$, azaz

$$\sum_{j=-(m-1)}^{m-1} c_j e^{\mathtt{i}jx_k} + \frac{c_m}{2} e^{\mathtt{i}mx_k} + \frac{c_{-m}}{2} e^{-\mathtt{i}mx_k} = f_k \tag{6.6.4}$$

interpolációs követelmények teljesítésére $(k = 0, \dots, n)$!

Vezessük be a

$$w = e^{\mathtt{i}2\pi/(n+1)}$$

jelölést az $(n+1)$-edik komplex egységgyökre $(w^{n+1} = 1)$. Ezzel a jelöléssel

$$e^{\mathtt{i}jx_k} = w^{jk}$$

és

$$w^{mk} = w^{(n+1)k/2} = (e^{\mathtt{i}\pi})^k = (-1)^k.$$

Tehát (6.6.4) bal oldalának utolsó két tagja összevonható

$$\sum_{j=-(m-1)}^{m-1} c_j w^{jk} + \underbrace{\frac{c_m}{2}(-1)^k + \frac{c_{-m}}{2}(-1)^k}_{c_m w^{mk}} = f_k$$

módon, azaz a megoldandó egyenletrendszer a

$$\sum_{j=-(m-1)}^{m} c_j w^{jk} = f_k \quad (k = 0, \dots, n) \tag{6.6.5}$$

alakot ölti. Ez az egyenletrendszer mátrixos alakban

$$\underbrace{\begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ w^{-(m-1)} & w^{-(m-2)} & \dots & w^{(m-1)} & w^m \\ w^{-2(m-1)} & w^{-2(m-2)} & \dots & w^{2(m-1)} & w^{2m} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ w^{-n(m-1)} & w^{-n(m-2)} & \dots & w^{n(m-1)} & w^{nm} \end{bmatrix}}_{\mathbf{G}_{n+1}} \underbrace{\begin{bmatrix} c_{-(m-1)} \\ c_{-(m-2)} \\ c_{-(m-3)} \\ \vdots \\ c_m \end{bmatrix}}_{\bar{\mathbf{c}}_{n+1}} = \underbrace{\begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}}_{\bar{\mathbf{f}}_{n+1}}$$

módon írható. Bevezetve a fenti jelöléseket a

$$\mathbf{G}_{n+1}\overline{\mathbf{c}}_{n+1} = \overline{\mathbf{f}}_{n+1} \tag{6.6.6}$$

lineáris egyenletrendszert kell megoldanunk $\overline{\mathbf{c}}_{n+1}$-re, ahol az alsó indexek a mátrixok és vektorok méretére utalnak. Most megmutatjuk, hogy az $(1/\sqrt{n+1})\mathbf{G}_{n+1}$ mátrix unitér (inverze a transzponált konjugáltja), így a keresett komplex Fourier együtthatók vektora a

$$\overline{\mathbf{c}}_{n+1} = \frac{1}{n+1}\mathbf{G}_{n+1}^H\overline{\mathbf{f}}_{n+1}, \tag{6.6.7}$$

vagy koordinátánként

$$c_j = \frac{1}{n+1}\sum_{k=0}^{n} f_k w^{-jk}, \quad j = -(m-1),\dots,m \tag{6.6.8}$$

módon, egyértelműen állítható elő.

Ehhez elég megmutatni, hogy

$$\mathbf{G}_{n+1}\mathbf{G}_{n+1}^H = (n+1)\mathbf{E}.$$

Ez látható, ha kiszámoljuk a mátrix k indexű sorának j indexű elemét ($k = 0,\dots,n$, $j = -(m-1),\dots,m$)

$$\begin{aligned}(\mathbf{G}_{n+1}\mathbf{G}_{n+1}^H)_{kj} &= \sum_{s=-(m-1)}^{m} w^{ks}w^{-js} = \sum_{s=-(m-1)}^{m} w^{s(k-j)} \\ &= \begin{cases} n+1, & \text{ha } k = j, \\ w^{-(m-1)(k-1)}\frac{(w^{k-j})^{n+1}-1}{w^{k-j}-1} = 0, & \text{ha } k \neq j. \end{cases}\end{aligned}$$

Mivel az f_k függvényértékek valósak, ezért $c_{-j} = \overline{c}_j$ $(j = 1,\dots,m-1)$, azaz ezek az együtthatók egymás konjugáltjai; $a_0 = c_0$ és $a_m = c_m$ valós értékek. Így tehát $a_j = 2\mathrm{Re}(c_j)$ és $b_j = -2\mathrm{Im}(c_j)$. Innét kapjuk a valós Fourier-együtthatók tételbeli alakját. ■

A bizonyításban szereplő (6.6.5) vagy (6.6.6) képletet a *Fourier-szintézis* vagy más néven *inverz diszkrét Fourier-transzformáció (IDFT)* képletének hívjuk, hiszen a komplex Fourier-együtthatókból ez a képlet állítja elő az alappontokbeli függvényértékeket. A (6.6.7) vagy (6.6.8) képletet pedig a *Fourier-analízis* vagy más néven *diszkrét Fourier-transzformáció (DFT)* képletének nevezzük. Ezen képlet segítségével lehet az alappontokbeli függvényértékekből a komplex Fourier-együtthatókat meghatározni.

Az előző tételhez hasonló igaz páros n esetén is. A tételt bizonyítás nélkül közöljük. (Bizonyítása az előző tételnek megfelelően könnyen megadható.)

6.6.2. tétel.
Tegyük fel, hogy az $x_k = 2\pi k/(n+1)$ alappontokban adottak az $f_k \in \mathbb{R}$ értékek ($k = 0, \dots, n$). Tegyük fel, hogy n páros. Ekkor egyértelműen létezik egy olyan $m = n/2$-ed fokú t_m trigonometrikus polinom, melyre $t_m(x_k) = f_k$ ($k = 0, \dots, n$). A valós diszkrét Fourier-együtthatók az alábbi módon számolhatók:

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

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

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

6.6.3. megjegyzés. Vegyük észre, hogy a fenti tételekben szereplő diszkrét Fourier-együtthatók tulajdonképpen a (6.6.1) képletben szereplő folytonos Fourier-együtthatók numerikus közelítései (lásd a numerikus integrálásról szóló 8. fejezetet). ◇

6.6.4. példa. Határozzuk meg a $(0,3)$, $(\pi/2,5)$, $(\pi,0)$, $(3\pi/2,1)$ pontokra illeszkedő legalacsonyabbfokú trigonometrikus interpolációs polinomot!

Mivel $n = 3$, így csak az a_0, a_1, a_2, b_1 diszkrét Fourier-együtthatókat kell meghatároznunk. A számolásokat célszerű áttekinthető formába táblázatba rendezni.

x_k	0	$\pi/2$	π	$3\pi/2$		
f_k	3	5	0	1	9	$9/4 = a_0$
$\cos x_k$	1	0	-1	0	3	$6/4 = 3/2 = a_1$
$\sin x_k$	0	1	0	-1	4	$8/4 = 2 = b_1$
$\cos(2x_k)$	1	-1	1	-1	-3	$-3/4 = a_2$

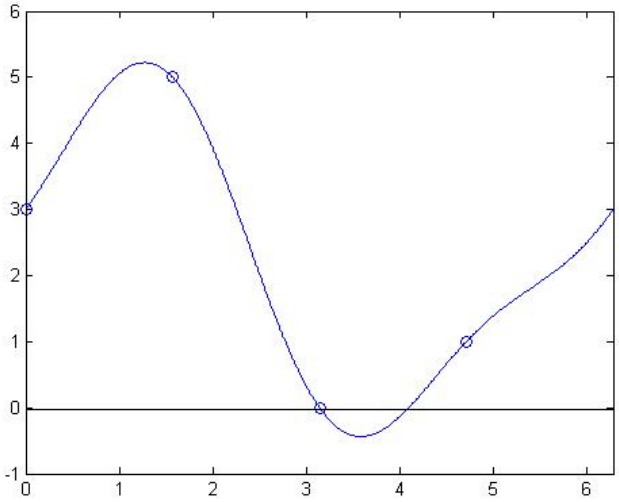
Itt az utolsó előtti oszlopban a függvényértékek és a trigonometrikus értékek vektorának skaláris szorzatai szerepelnek. Az utolsó oszlopban pedig ezen értékek vannak beszorozva $2/(n+1) = 2/4 = 1/2$-del, a_0 és a_2 esetén csak ennek felével. Így a keresett interpolációs polinom (6.6.1. ábra)

$$t_2(x) = \frac{9}{4} + \frac{3}{2}\cos x + 2\sin x - \frac{3}{4}\cos(2x).$$

◇

6.7. Gyors Fourier-transzformáció

A diszkrét Fourier-transzformáció esetén a $\overline{\mathbf{c}}_{n+1}$ együtthatóvektor kiszámításához be kell szoroznunk az $\overline{\mathbf{f}}_{n+1}$ függvényértékvektort a $\mathbf{G}_{n+1}^H$ mátrixszal. Ez alapesetben $(n+1)^2$ komplex szorzást jelent, ha már az $(n+1)$-edik egységgyököket előre kiszámítottuk. (Itt az egyszerűség kedvéért csak a szorzásokat számoljuk. Ezt azzal indokoljuk, hogy egy komplex összeadás csak 2 flop,



6.6.1. ábra: A $(0,3)$, $(\pi/2,5)$, $(\pi,0)$, $(3\pi/2,1)$ pontokat interpoláló trigonometrikus polinom.

míg egy komplex szorzás 6 flop.) Vajon megvalósítható-e ez a szorzás kevesebb művelettel is? Első ránézésre ez egyáltalán nem világos, hiszen a $\mathbf{G}_{n+1}^H$ mátrix egyik eleme sem nulla, azaz egy teli mátrixszal van dolgunk. Ugyanakkor az is látható, hogy a mátrix csak $n+1$ különböző elemet tartalmazhat, hiszen ennyi különböző $n+1$-edik egységgyök van. Ez reményt adhat arra, hogy egyszerűsíthető a szorzás. A továbbiakban fel fogjuk tenni, hogy $n+1$ páros. Ekkor egy $m=(n+1)/2$-ed fokú kiegyensúlyozott interpolációs polinomot keresünk. Az $m=(n+1)/2$ egyenlőség miatt hol m-et, hol $(n+1)/2$-öt fogunk használni. Mindig azt, amelyik kényelmesebb, vagy jobban mutatja a lényeget.

A most ismertetendő módszert, amit *gyors Fourier-transzformációnak* (angolul: fast Fourier transform (FFT)) neveznek, már Gauss is leírta az 1800-as évek elején, de munkája feledésbe merült. Az eljárást a számítógépek megjelenésével újra felfedezték. Az első alkalmazás 1965-ben James W. Cooley (IBM) és John W. Tukey (Princeton) nevéhez fűződik.

A diszkrét Fourier-transzformáció végrehajtásához ki kell számolnunk a (6.6.7) mátrixszorzást, ami koordinátánként kiírva

$$\begin{bmatrix} c_{-(m-1)} \\ c_{-(m-2)} \\ c_{-(m-3)} \\ \vdots \\ c_m \end{bmatrix} = \frac{1}{n+1} \begin{bmatrix} 1 & w^{(m-1)} & w^{2(m-1)} & \dots & w^{n(m-1)} \\ 1 & w^{(m-2)} & w^{2(m-2)} & \dots & w^{n(m-2)} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & w^{-(m-1)} & w^{-2(m-1)} & \dots & w^{-n(m-1)} \\ 1 & w^{-m} & w^{-2m} & \dots & w^{-nm} \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}$$

alakú. Sorcserékkel (a $\overline{\mathbf{c}}_{n+1}$ vektor elemsorrendje is megváltozik) és kihasználva, hogy w $(n+1)$-

edik egységgyök (azaz $w^s = w^{s+2m}$), az alábbi alakra írható ez az egyenletrendszer

$$\begin{bmatrix} c_0 \\ c_{-1} \\ c_{-2} \\ \vdots \\ c_{-(m-1)} \\ c_m \\ c_{m-1} \\ \vdots \\ c_1 \end{bmatrix} = \frac{1}{n+1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & w & w^2 & \dots & w^n \\ 1 & w^2 & w^4 & \dots & w^{2n} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & w^{m-1} & w^{2(m-1)} & \dots & w^{(m-1)n} \\ 1 & w^m & w^{2m} & \dots & w^{mn} \\ 1 & w^{(m+1)} & w^{2(m+1)} & \dots & w^{(m+1)n} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & w^n & w^{2n} & \dots & w^{n^2} \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}.$$

Ezután oszlopcserét hajtunk végre úgy, hogy előre cseréljük a mátrix páratlan sorszámú oszlopait. Ekkor az $\bar{\mathbf{f}}_{n+1}$ vektor elemsorrendje is megváltozik.

$$\begin{bmatrix} c_0 \\ c_{-1} \\ c_{-2} \\ \vdots \\ c_{-(m-1)} \\ c_m \\ c_{m-1} \\ \vdots \\ c_1 \end{bmatrix} = \frac{1}{n+1} \underbrace{\left[\begin{array}{cccc|cccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & w^2 & \dots & w^{n-1} & w & w^3 & \dots & w^n \\ 1 & w^4 & \dots & w^{2(n-1)} & w^2 & w^6 & \dots & w^{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & w^{2(m-1)} & \dots & w^{(m-1)(n-1)} & w^{m-1} & w^{3(m-1)} & \dots & w^{(m-1)n} \\ \hline 1 & w^{2m} & \dots & w^{m(n-1)} & w^m & w^{3m} & \dots & w^{mn} \\ 1 & w^{2(m+1)} & \dots & w^{(m+1)(n-1)} & w^{(m+1)} & w^{3(m+1)} & \dots & w^{(m+1)n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & w^{2n} & \dots & w^{n(n-1)} & w^n & w^{3n} & \dots & w^{n^2} \end{array}\right]}_{\mathbf{F}_{n+1}} \left[\begin{array}{c} f_0 \\ f_2 \\ f_4 \\ \vdots \\ f_{n-1} \\ \hline f_1 \\ f_3 \\ \vdots \\ f_n \end{array}\right].$$

Vezessük be a fenti egyenletrendszer mátrixára az $\mathbf{F}_{n+1}$ jelölést, továbbá az $\bar{\mathbf{f}}_{n+1}$ vektort is osszuk fel két $\tilde{\mathbf{f}}_1 = [f_0, f_2, f_4, \dots, f_{n-1}]^T \in \mathbb{R}^m$ és $\tilde{\mathbf{f}}_2 = [f_1, f_3, \dots, f_n]^T \in \mathbb{R}^m$ vektorra. Vegyük észre, hogy az $\mathbf{F}_{n+1}$ mátrixot már az $n+1$ számérték maga meghatározza. A DFT felgyorsításához az

$$\mathbf{F}_{n+1} \begin{bmatrix} \tilde{\mathbf{f}}_1 \\ \tilde{\mathbf{f}}_2 \end{bmatrix}$$

szorzást kellene $(n+1)^2$ komplex szorzás helyett kevesebbel végrehajtani. Ez a szorzás az $\mathbf{F}_{n+1}$ mátrix ügyes felírásával az

$$\mathbf{F}_{n+1} \begin{bmatrix} \tilde{\mathbf{f}}_1 \\ \tilde{\mathbf{f}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{E}_{(n+1)/2} & \mathbf{D}_{(n+1)/2} \\ \mathbf{E}_{(n+1)/2} & -\mathbf{D}_{(n+1)/2} \end{bmatrix} \begin{bmatrix} \mathbf{F}_{(n+1)/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_{(n+1)/2} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{f}}_1 \\ \tilde{\mathbf{f}}_2 \end{bmatrix}$$

alakban írható (lásd még a mátrix partícióját az előző képletben), ahol

$$\mathbf{D}_{(n+1)/2} = \operatorname{diag}(1, w, \dots, w^{m-1}) \in \mathbb{R}^{m \times m}$$

diagonálmátrix, $\mathbf{F}_{(n+1)/2}$ az $(n+1)/2$-edik egységgyökökkel $\mathbf{F}_{n+1}$-hez hasonlóan képzett mátrix, $\mathbf{E}_{(n+1)/2}$ pedig az $m \times m$-es egységmátrix. Számoljuk ki a komplex szorzások számát a fenti mátrixszorzás során! Először is az $\mathbf{F}_{(n+1)/2}\tilde{\mathbf{f}}_1$ szorzás $((n+1)/2)^2$ komplex szorzást jelent, hasonlóan az $\mathbf{F}_{(n+1)/2}\tilde{\mathbf{f}}_2$ szorzáshoz (mindkettő egy fele akkora méretű diszkrét Fourier-transzformációnak felel meg, mint az eredeti feladat). Ezután még a $\mathbf{D}_{(n+1)/2}$ diagonálmátrixszal kell beszoroznunk az $\mathbf{F}_{(n+1)/2}\tilde{\mathbf{f}}_2$ szorzat $(n+1)/2$ méretű eredményvektorát. Ez további $(n+1)/2$ komplex szorzást

jelent. Több komplex szorzásra nincs szükség. Így összesen az eredeti $(n+1)^2$ komplex szorzás helyett csak

$$2\left(\frac{n+1}{2}\right)^2+\frac{n+1}{2}$$

szorzást kell végrehajtanunk.

Még jelentősebb műveletszám-csökkenés érhető el, ha észrevesszük, hogy az $\mathbf{F}_{(n+1)/2}\tilde{\mathbf{f}}_1$ szorzás során egy ugyanolyan típusú, csak kisebb méretű mátrixszal kell szoroznunk egy vektort, mint az eredeti feladat esetén. Ha tehát $(n+1)/2$ maga is páros, akkor az ismertetett eljárás újra alkalmazható az $\mathbf{F}_{(n+1)/2}\tilde{\mathbf{f}}_1$ és az $\mathbf{F}_{(n+1)/2}\tilde{\mathbf{f}}_2$ szorzás esetén is. Az ideális eset az, amikor $n+1$ kettőhatvány. Vizsgáljuk meg ezt az esetet külön az elvégzendő komplex szorzások szempontjából.

Jelölje Q_l a 2^l osztóponttal rendelkező FFT komplex szorzásainak számát. Ekkor nyilván

$$Q_l = 2Q_{l-1}+2^{l-1},$$

hiszen két fele akkora méretű mátrixszorzást és egy fele akkora méretű vektor elemeinek egységgyökhatványokkal való beszorzását kell elvégeznünk. Figyelembe véve, hogy $Q_1=1$, teljes indukcióval kapjuk, hogy

$$Q_l = l2^{l-1}=\frac{1}{2}(n+1)\log_2(n+1).$$

Ahogy ezt az alábbi táblázat mutatja, ez jelentős műveletszám-csökkenés a DFT-hoz képest. A táblázatban az osztópontok $n+1$ számának függvényében adtuk meg a DFT-hez és a FFT-hez szükséges komplex szorzások számát.

$n+1$	DFT	FFT
$2^5=32$	1024	80
$2^{10}=1024$	1048576	5120
$2^{20}=1048576$	1099511627776	10485760

6.8. Közelítés legkisebb négyzetek értelemben

A statisztikában gyakori az a feladat, hogy a koordinátarendszerben elhelyezkedő (x_i,f_i) $(i=1,\dots,n)$ pontokhoz "legközelebb haladó" adott típusú függvényt kell meghatároznunk. Ennek segítségével lehet ugyanis megállapítani az x_i és f_i értékeket szolgáltató valószínűségi változók közti kapcsolatot. Jelölje $\mathcal{F}$ azon függvények halmazát, melyek közül szeretnénk kiválasztani az adott pontokhoz "legközelebb haladó" függvényt. Mielőtt tisztázzuk, hogy mit értünk "legközelebb haladó" függvényen, vezessük be a következő jelöléseket. Az $\overline{\mathbf{x}}$ vektor fogja jelölni az alappontok vektorát, ahol az alappontokat monoton növő sorrendben indexeljük: $\overline{\mathbf{x}}=[x_1,\dots,x_n]^T$. Egy adott $\phi\in\mathcal{F}$ függvény esetén $\phi(\overline{\mathbf{x}})$ fogja jelölni a $[\phi(x_1),\dots,\phi(x_n)]^T$ vektort (ez a jelölés a MATLAB-ban is szokásos). Jelölje továbbá $\bar{\mathbf{f}}$ az $[f_1,\dots,f_n]^T$ vektort.

6.8.1. definíció.

Azt mondjuk, hogy a $\phi^\star\in\mathcal{F}$ függvény *legkisebb négyzetek értelemben legjobb közelítése* az (x_i,f_i) $(i=1,\dots,n)$ pontoknak, ha $\|\phi^\star(\overline{\mathbf{x}})-\bar{\mathbf{f}}\|_2^2\le\|\phi(\overline{\mathbf{x}})-\bar{\mathbf{f}}\|_2^2$ minden $\phi\in\mathcal{F}$ függvény esetén.

A legkisebb négyzetek értelemben legjobb közelítést két speciális esetben fogjuk megadni. Az egyik az az eset lesz, amikor $\mathcal{F}$ a legfeljebb k-adfokú polinomok halmaza, a másik pedig az, amikor $\mathcal{F}$ bizonyos, az alappontokon ortonormált függvények összes lineáris kombinációja.

Kezdjük tehát azzal az esettel, amikor $\mathcal{F}=P_k$, azaz a legfeljebb k-adfokú polinomok halmaza. Nyilvánvalóan feltehetjük, hogy $k\le n_{\text{kül.}}-1$, ahol $n_{\text{kül.}}$ jelöli az $x_1,\dots,x_n$ értékek között a különböző értékek számát.

6.8.2. tétel.
Az (x_i, f_i) $(i = 1, \dots, n)$ pontokat legkisebb négyzetek értelemben legjobban közelítő, legfeljebb k-adfokú $(k \leq n_{\text{kül.}} - 1)$ $a_k x^k + \ldots + a_1 x + a_0$ polinom együtthatóit az

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^k \\ 1 & x_2 & x_2^2 & \dots & x_2^k \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^k \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}}_{\bar{\mathbf{a}}} = \underbrace{\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_n \end{bmatrix}}_{\bar{\mathbf{f}}} \tag{6.8.1}$$

túlhatározott lineáris egyenletrendszer legkisebb négyzetek értelemben legjobb $\bar{\mathbf{a}}_{LS}$ megoldása adja.

Bizonyítás. A (6.8.1) egyenletrendszer nyilvánvalóan túlhatározott, hiszen $n \geq n_{\text{kül.}} \geq k+1$, és teljes rangú. Ha a mátrix nem lenne teljes rangú, akkor léteznének olyan $a_0, \dots, a_k$ együtthatók, melyek közül legalább az egyik különbözik nullától, és

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^k \\ 1 & x_2 & x_2^2 & \dots & x_2^k \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^k \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Ekkor viszont az $a_k x^k + \ldots + a_1 x + a_0$ legfeljebb k-ad fokú polinomnak $n_{\text{kül.}} \geq k+1$ zérushelye lenne, ami ellentmond az algebra alaptételének.

A túlhatározott egyenletrendszer legkisebb négyzetek értelemben legjobb $\bar{\mathbf{a}}_{LS}$ megoldására (a tételbeli jelölésekkel) $\|\mathbf{A}\bar{\mathbf{a}}_{LS} - \bar{\mathbf{f}}\|_2^2 \leq \|\mathbf{A}\bar{\mathbf{a}} - \bar{\mathbf{f}}\|_2^2$ minden $\bar{\mathbf{a}} \in \mathbb{R}^{k+1}$ esetén. Mivel $(\mathbf{A}\bar{\mathbf{a}})_i = a_k x_i^k + \ldots + a_1 x_i + a_0$ $(i = 1, \dots, n)$, ezért az $\bar{\mathbf{a}}_{LS}$ vektor éppen a legkisebb négyzetek értelemben legjobban közelítő polinom együtthatóit tartalmazó vektor lesz. ■

6.8.3. példa. Keressük meg a $(-1, 1), (0, 2), (1, 2), (2, 4)$ pontokat legjobban közelítő legfeljebb elsőfokú polinomot! Először felírjuk az

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix}.$$

normálegyenetet, amit tömörebben az

$$\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n f_i \\ \sum_{i=1}^n (x_i f_i) \end{bmatrix}$$

alakban írhatunk. Az egyenletrendszer megoldása $a_0 = 9/5$, $a_1 = 9/10$, így a legjobban közelítő polinom a $p(x) = 9x/10 + 9/5$ polinom lesz. ◇

Most térjünk át arra az esetre, amikor $\mathcal{F}$ bizonyos ortonormált függvények összes lineáris kombinációját tartalmazza! Tegyük fel most, hogy $n = n_{\text{kül.}}$, azaz, hogy nincs két egyforma abszcisszájú pont az adott pontok között.

6.8.4. definíció.

Azt mondjuk, hogy a ϕ_1 és ϕ_2 függvények ortogonálisak az $x_1, \ldots, x_n$ alappontokon, ha $\phi_1^T(\overline{\mathbf{x}})\phi_2(\overline{\mathbf{x}}) = 0$. Ha $\phi^T(\overline{\mathbf{x}})\phi(\overline{\mathbf{x}}) = 1$, akkor azt mondjuk, hogy a ϕ függvény normált az alappontokon.

6.8.5. tétel.

Legyenek $\phi_1, \ldots, \phi_k$ páronként ortogonálisak és normáltak az $x_1, \ldots, x_n$ alappontokon, és legyen $\mathcal{F} = \operatorname{lin}\{\phi_1, \ldots, \phi_k\}$, azaz a ϕ_i függvények összes lineáris kombinációja. Az (x_i, f_i) $(i = 1, \ldots, n)$ (különböző abszcisszájú) pontokat legkisebb négyzetek értelemben legjobban közelítő $\phi^\star$ függvény az $\mathcal{F}$ halmazból a

$$\phi^\star(x) = \sum_{i=1}^{k} (\phi_i^T(\overline{\mathbf{x}})\overline{\mathbf{f}})\phi_i(x)$$

alakban írható.

Bizonyítás. Először vegyük észre, hogy nyilvánvalóan $k \leq n$ kell legyen, hiszen n-elemű vektorból csak maximum n páronként ortogonális vektor adható meg. Keressük a legkisebb négyzetek értelemben legjobban közelítő függvényt $\mathcal{F}$-ből a $\phi^\star(x) = \sum_{i=1}^{k} \alpha_i\phi_i(x)$ alakban!

Ekkor mivel tetszőleges $\phi \in \mathcal{F}$, $\phi(x) = \alpha_1\phi_1(x) + \ldots + \alpha_k\phi_k(x)$ függvény esetén

$$\phi(\overline{\mathbf{x}}) = \alpha_1\phi_1(\overline{\mathbf{x}}) + \ldots + \alpha_k\phi_k(\overline{\mathbf{x}}) = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \ldots & \phi_k(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \ldots & \phi_k(x_2) \\ \vdots & \vdots & \ldots & \vdots \\ \phi_1(x_n) & \phi_2(x_n) & \ldots & \phi_k(x_n) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix},$$

így az a $\phi^\star \in \mathcal{F}$ függvény adja a legjobb közelítést, melyre az $\alpha_1, \ldots, \alpha_n$ értékek éppen az

$$\underbrace{\begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \ldots & \phi_k(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \ldots & \phi_k(x_2) \\ \vdots & \vdots & \ldots & \vdots \\ \phi_1(x_n) & \phi_2(x_n) & \ldots & \phi_k(x_n) \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}}_{\overline{\boldsymbol{\alpha}}_{LS}} = \underbrace{\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_n \end{bmatrix}}_{\overline{\mathbf{f}}}$$

túlhatározott lineáris egyenletrendszer legkisebb négyzetek értelemben legjobb megoldását adják. Az egyenletrendszer az $n \geq k$ egyenlőség miatt nyilván túlhatározott és az oszlopvektorok ortogonalitása miatt teljes rangú is. A 3.9. fejezet alapján az egyenletrendszer legkisebb négyzetek értelemben legjobb $\overline{\boldsymbol{\alpha}}_{LS} = [\alpha_1, \ldots, \alpha_k]^T$ megoldása a fenti jelölésekkel az

$$\overline{\boldsymbol{\alpha}}_{LS} = (\underbrace{\mathbf{A}^T\mathbf{A}}_{=\mathbf{E}})^{-1}\mathbf{A}^T\overline{\mathbf{f}} = \mathbf{A}^T\overline{\mathbf{f}}$$

alakban írható, azaz $\alpha_i = (\mathbf{A}^T\overline{\mathbf{f}})_i = \sum_{j=1}^{k} \phi_i(x_j)f_j = \phi_i^T(\overline{\mathbf{x}})\overline{\mathbf{f}}$. Ezt akartuk megmutatni. ■

Ha a fenti tétel segítségével szeretnénk a legkisebb négyzetek értelemben legjobban közelítő legfeljebb k-adfokú ($k \leq n-1$) polinomot megadni, akkor a következő módon járhatunk el. Vegyük észre, hogy a P_k halmazon a $\langle p, q\rangle = p^T(\overline{\mathbf{x}})q(\overline{\mathbf{x}})$ függvény skaláris szorzatot definiál. Ezzel a skaláris szorzattal a Gram–Schmidt-ortogonalizációval előállítunk az $1, x, \dots x^k$ lineárisan független polinomokból egy ortonormált bázist: $q_0, q_1, \dots, q_k$, ahol az alsó index a polinom fokszámát jelöli. Ezután a legkisebb négyzetek értelemben legjobban közelítő polinom a

$$p(x) = \sum_{i=0}^{k} (q_i^T(\overline{\mathbf{x}})\overline{\mathbf{f}})q_i(x)$$

alakban írható.

6.8.6. példa. Keressük meg a $(-1,1), (0,2), (1,2), (2,4)$ pontokhoz legjobban illeszkedő legfeljebb elsőfokú polinomot az alappontokon ortonormált polinomok segítségével!

Az 1 és x polinomokat ortonormálva a megfelelő skaláris szorzatra nézve azt kapjuk, hogy $q_0(x) = 1/2$ és $q_1(x) = (x - 1/2)/\sqrt{5}$. Innét

$$p(x) = \frac{9}{2}\frac{1}{2} + \frac{9}{2\sqrt{5}}\frac{1}{\sqrt{5}}(x - 1/2) = \frac{9}{10}x + \frac{9}{5}.$$

◇

Az $x_k = 2\pi k/(n+1)$ $(k = 0, \dots, n)$ alappontokon az

$$1, \sin(jx), \cos(jx), \ (j = 1, 2, \dots)$$

polinomok ortogonális rendszert alkotnak (6.10.15. feladat). Ezt a tényt felhasználva a legkisebb négyzetek értelemben legjobban közelítő legfeljebb k-adfokú trigonometrikus polinom hasonlóan határozható meg, mint a polinomok esetén. Eredményül azt kapjuk, hogy az alacsonyabb fokszámú négyzetösszeg értelemben legjobban közelítő trigonometrikus polinomokat az interpoláló trigonometrikus polinom megfelelő csonkolásaival állíthatjuk elő.

6.8.7. megjegyzés. Tegyük fel, hogy már meghatároztuk az adott pontokat legjobban közelítő legfeljebb k-adfokú polinomot. Ekkor, ha a legfeljebb $k+1$ fokú hasonló polinomra vagyunk kíváncsiak, és a normálegyenlettel határoztuk meg a legfeljebb k-adfokú polinomot, akkor azt most újra fel kell írni, és meg kell oldani. Ha ellenben ortonormált polinomokat használtunk, akkor cak meg kell határozni a $(k+1)$-edik ortonormált polinomot, és ezt megfelelő együtthatóval hozzáadni a korábban kiszámolt k-adfokú polinomhoz. ◇

6.9. Interpolációs feladatok megoldása a MATLAB-ban

A MATLAB-ban többek között az alábbi parancsokat használhatjuk interpolációs feladatok megoldására.

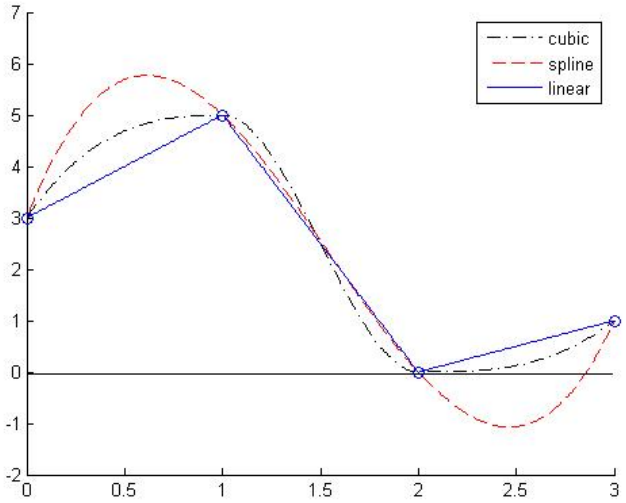
- `z=polyfit(x,y,n)`: Az interpolációs pontok x-koordinátáit az `x` vektor, míg y-koordinátáit az ugyanolyan méretű `y` vektor tartalmazza. Az `n` érték az illesztendő polinom fokszámát adja meg. Ha nincs a feltételnek megfelelő polinom, akkor a legkisebb négyzetek értelmében legjobban közelítő polinomot kapjuk vissza. Az interpolációs polinom együtthatói a `z` vektorba kerülnek a legmagasabb fokú taggal kezdődően.

- `yi=polyval(z,xi)`: A `z` vektorral adott polinom értékét számolja ki a parancs az `xi` vektorban meghatározott helyeken. A függvényértékek rendre az `yi` vektorban jelennek meg.
- `yi = interp1(x,y,xi,'módszer')`: A parancs az `x` és `y` vektorok által meghatározott pontokra illeszt a 'módszer' eljárásnak megfelelően egy interpolációs függvényt. Ezen függvény `xi` vektorban adott helyeken vett függvényértékeit tartalmazza az `yi` vektor. A 'módszer' lehet pl.

  'linear': szakaszonként lineáris,

  'spline': szakaszonként legfeljebb harmadfokú spline-interpoláció,

  'cubic': szakaszonként legfeljebb harmadfokú polinom, mely monotonitástartó. Azaz monoton y értékek esetén az interpolációs függvény is ugyanolyan monotonitású, konstans értékek között a függvény is konstans (6.9.1. ábra).
- `fft`, `ifft`: gyors Fourier-transzformáció és inverzének parancsa.



6.9.1. ábra: A 'cubic', 'spline' és 'linear' interpoláció megvalósításának szemléltetése.

Lássunk most néhány példát a fenti parancsok alkalmazására!

```
>> x=[0,1,2,3], y=[0,1,4,9] % A pontok x- és y-koordinátáinak megadása.

x =

     0     1     2     3


y =
```

```
     0     1     4     9

>> z=polyfit(x,y,3) % Legfeljebb harmadfokú interpolációs polinom.

z =

    0.0000    1.0000    0.0000    0.0000        % Ez az x^3 függvény.

>> yi=polyval(z,[0.3,0.6]) % A polinom helyettesítési értéke 0.3-nél
                           % és 0.6-nél.

yi =

    0.0900    0.3600

>> yi=interp(x,y,[0.5,1.5],'linear'); % Szakaszonként lineáris
                         % interpolációs függvény 0.5-nél és 1.5-nél.

yi =

    0.5000    2.5000
```

A MATLAB `fft` parancsa a komplex Fourier-együtthatókat határozza meg. Az alábbi program megadja a MATLAB által adott komplex együtthatókból a valós együtthatókat.

```
function dftvalos(n,fx)
m=(n+1)/2;
h = 2*pi/(n+1); x=[0:h:2*pi*(n/(n+1))]; w=exp(i*h); %fx=eval(f);
X=fft(fx);
a=2*real(X(2:m+1))/(n+1);
b=-2*imag(X(2:m+1))/(n+1);
display('A valós diszkrét Fourier-együtthatók:')
a0=X(1)/(n+1)
if mod(n,2) == 1
  a(m)=a(2)/2;
  end
a
b
```

A legkisebb négyzetek értelemben vett közelítéseket jól szemlélteti pl. a http://www.chem.uoa.gr/applets/AppletPoly/Appl_Poly2.html oldalon található alkalmazás.

6.10. Feladatok

Polinominterpoláció

6.10.1. feladat. Számítógép segítségével határozzuk meg az

$$\int_0^1 e^{-x^2}\,\mathrm{d}x$$

integrál értékét úgy, hogy a [0,1] intervallumot öt egyenlő részre osztjuk, meghatározzuk az interpolációs polinomot, és azt integráljuk az adott intervallumon! Az integrál "pontos" értéke 0.7468241330. Közelítsük az e^{-x^2} függvény deriváltját az interpolációs polinom deriváltja segítségével az $x = 0.5$ pontban! Hasonlítsuk össze az eredményt a pontos derivált értékével!

6.10.2. feladat. A $\sin x$ függvény értékeit ismerjük a $[0, \pi]$ intervallumon a $k\pi/6$ pontokban ($k = 0, 1, 2, 3, 4, 5, 6$). Illesszünk ezekre a pontokra egy interpolációs polinomot, és annak segítségével határozzuk meg $\sin 1$ közelítő értékét! Becsüljük meg a számítás előtt, hogy mekkora hibára számíthatunk!

6.10.3. feladat. Számítógép használata nélkül határozzuk meg a $(-1, 6), (0, 3)$ és $(1, 2)$ pontokra illeszkedő interpolációs polinomot a Lagrange- és Newton-módszerrel is!

6.10.4. feladat. Számítógép használata nélkül határozzuk meg az $f(x) = \sqrt[4]{x} + x - 2$ függvény értékeit a 0.5, 80 és 25 pontokban a 0, 1, 16 és 81 pontokhoz tartozó interpolációs polinom segítségével!

6.10.5. feladat. Hogyan egyszerűsíthető az interpolációs polinom meghatározása a Newton-módszerrel, ha az alappontok egyforma távol vannak egymástól? Határozzuk meg a módszerrel a (4,1), (6,3), (8,8) és (10,20) pontokhoz tartozó interpolációs polinomot!

6.10.6. feladat. Igazoljuk a 6.2.7. tételt teljes indukció segítségével!

6.10.7. feladat. Számítógép nélkül határozzuk meg azt a legalacsonyabb fokú p polinomot, melyre $p(1) = 2$, $p'(1) = 1$, $p(3) = 1$ és $p'(3) = 2$!

6.10.8. feladat. Az előző feladat adatait módosítsuk úgy, hogy még ismert a $p(2) = 2$ és $p'(2) = 1$ feltétel is! Mi lesz ekkor az interpolációs polinom?

6.10.9. feladat. Illesszünk szakaszonként harmadfokú spline-függvényt az (1,2), (2,1), (3,1) alappontokra!

6.10.10. feladat. Osszuk fel a $[0, \pi]$ intervallumot három egyenlő hosszúságú intervallumra és tekintsük az osztópontokban a $\sin x$ függvény értékeit. Az így nyert pontokra illesszünk szakaszonként harmadfokú spline-függvényt! Becsüljük meg az eredeti függvény és a spline-függvény maximális eltérését!

6.10.11. feladat. Adjuk meg, hogy hogyan egyszerűsödik a q_k súlyok kiszámítása a baricentrikus interpolációs formulánál abban az esetben, ha a Csebisev-alappontokat alkalmazzuk az interpolációra!

Trigonometrikus interpoláció

6.10.12. feladat. Tekintsük az $f(x) = |x|$ függvényt, és válasszuk alappontoknak az $x_k = 2\pi k/(3+1)$ ($k = 0, \ldots, 3$) pontokat. Határozzuk meg az f függvényt az alappontokon interpoláló legalacsonyabb fokszámú (kiegyensúlyozott) trigonometrikus polinomot számítógép alkalmazása nélkül!

6.10.13. feladat. Hogyan lehet a DFT-t felgyorsítani abban az esetben, ha az interpolációs alappontok száma nem kettőhatvány, hanem két egész szám szorzatakét áll elő?

6.10.14. feladat. Alkalmazzuk az előző feladatban konstruált FFT módszert a Fourier-együtthatók

előállítására az alábbi alappontok esetén! Az FFT eljárást kézi számolással hajtsuk végre!

x_k	0	$2\pi/6$	$4\pi/6$	$6\pi/6$	$8\pi/6$	$10\pi/6$
f_k	0	1	1	0	-1	-1

Legkisebb négyzetek értelemben legjobb közelítések

6.10.15. feladat. Igazoljuk, hogy az $1, \cos(jx), \sin(jx)$ $(j = 1, 2, \ldots)$ függvények ortogonálisak az $x_k = 2\pi k/(n+1)$ $(k = 0, \ldots, n)$ alappontokon! Határozzuk meg ez alapján a trigonometrikus interpolációs polinom együtthatóit abban az esetben, ha n páros érték!

6.10.16. feladat. Határozzuk meg az alábbi pontokat legkisebb négyzetek értelemben legjobban közelítő legfeljebb első és másodfokú polinomokat! (Oldjuk meg a feladatot a normálegyenlet felírásával ill. ortogonális polinomok használatával!) Oldjuk meg a feladatot ellenőrzésképpen a MATLAB `polyfit` parancsával is!

x_k	-1	0	1	2	4
f_k	3	1	1	0	-1

Ellenőrző kérdések

1. Adjuk meg a globális polinominterpoláció alapfeladatát! Mit mondhatunk az interpolációs polinom létezéséről és unicitásáról?
2. Hogyan állíthatjuk elő az interpolációs polinomot Lagrange módszerével?
3. Definiáljuk az osztott-differenciák fogalmát és adjuk meg tulajdonságaikat!
4. Hogyan állíthatjuk elő az interpolációs polinomot Newton módszerével?
5. Mit ad meg az interpolációs hiba és hogyan tudjuk megadni az értékét?
6. Mit szemléltet Runge példája?
7. Milyen extremális tulajdonságai vannak a Csebisev-polinomoknak?
8. Mit értünk Hermite–Fejér interpoláción? Hogy kell ezt az interpolációs polinomot meghatározni?
9. Mit értünk spline-interpoláción?
10. Milyen tulajdonsággal rendelkeznek a szakaszonként legfeljebb harmadfokú spline-függvények?
11. Mit jelentenek a Fourier-szintézis és Fourier-analízis fogalmak?
12. Milyen eljárást nevezünk gyors Fourier-transzformációnak?
13. Hogyan határozhatunk adott alappontokat legkisebb négyzetek értelemben legjobban közelítő függvényeket? Hasonlítsuk össze a módszereket!

7. Numerikus deriválás

Ebben a fejezetben azt vizsgáljuk, hogy hogyan lehet egy függvény deriváltjait közelíteni a függvényértékek segítségével. Bevezetjük a haladó-, retrográd- és központi differenciákat. A bemutatott képletek a differenciálegyenletek numerikus megoldásával foglalkozó fejezetben játszanak majd fontos szerepet.

7.1. A numerikus deriválás alapfeladata

Tegyük fel, hogy az $x_0, x_0 \pm h, x_0 \pm 2h, \dots, x_0 \pm kh \in \mathbb{R}$ $(h > 0, k \in \mathbb{N})$ pontokban ismertek egy f kellően sokszor differenciálható függvény függvényértékei. Jelölje ezeket rendre $f_0, f_{\pm 1}, f_{\pm 2}, \dots, f_{\pm k}$. Ebben a fejezetben azt vizsgáljuk meg, hogy hogyan közelíthetjük az f függvény deriváltjait az x_0 pontban (az egyszerűség kedvéért f_0', f_0'' stb. jelöli ezeket) a függvényértékek segítségével és hogy ezek a közelítések milyen tulajdonságokkal rendelkeznek.

A fenti típusú feladattal találkozunk amikor diszkrét időpontokban mért helykoordináták segítségével sebességet, diszkrét sebességértékek segítségével gyorsulást, diszkrét töltésmennyiség értékek esetén áramerősséget stb. szeretnénk becsülni. Pl. a GPS műholdak 15 percenként sugároznak adatokat a helyzetükről. Hogyan közelíthetnénk a műhold sebességét és gyorsulását az adatok alapján? A numerikus deriválás másik fontos alkalmazási területe a differenciálegyenletek numerikus megoldása. Pl. a véges differenciás módszerek esetén az ismeretlen megoldásfüggvény deriváltjait az egyelőre szintén ismeretlen diszkrét pontokbeli függvényértékek segítségével közelítjük. Ebből egy egyenletrendszert nyerünk a függvényértékekre, amit megoldva megkaphatjuk a megoldás közelítését.

7.1.1. definíció.

Jelölje a kellően sokszor deriválható f függvény egy tetszőleges deriváltját az x_0 pontban Df, és ennek egy közelítése legyen $\Delta f(h)$ (a h argumentum azt fejezi ki, hogy a közelítés függ az alappontok h távolságától). Azt mondjuk, hogy az x_0 pontban a $\Delta f(h)$ *közelítés rendje (legalább)* r, ha van olyan $K > 0$ szám, hogy

$$|Df - \Delta f(h)| \leq Kh^r,$$

azaz, ha $|Df - \Delta f(h)| = \mathcal{O}(h^r)$.

7.1.2. megjegyzés. A fenti definíciót már az 1.3.10. definícióban megadtuk általános közelítésekre. Ez a definíció csupán az általános eset megfogalmazása a deriváltak közelítésének problémájára. ◇

7.1.3. megjegyzés. Mi csak azt az esetet vizsgáljuk, mikor az alappontok elhelyezkedése ekvidisztáns. A nem ekvidisztáns esetre vonatkozó formulák hasonlóan nyerhetők (lásd 7.6.2. feladat). ◇

7.2. Az első derivált közelítése

Egy függvény deriváltját a differenciahányados határértékeként értelmeztük. Így kézenfekvő a deriváltat a differenciahányados értékeivel közelíteni. Két egyszerű eset erre vonatkozóan a

$$\Delta f_+ := \frac{f_1 - f_0}{h}$$

ún. haladó (angolul forward) differencia és a

$$\Delta f_- := \frac{f_0 - f_{-1}}{h}$$

ún. retrográd (angolul backward) differencia. Az elnevezés onnét származik, hogy a haladó differencia az x értékek növekedési irányába (a számegyenesen jobbra) eső alappontot használja, míg a retrográd a csökkenés irányába esőt.

7.2.1. tétel.
A haladó és a retrográd differencia is elsőrendű közelítése egy $f \in C^2$ függvény deriváltjának.

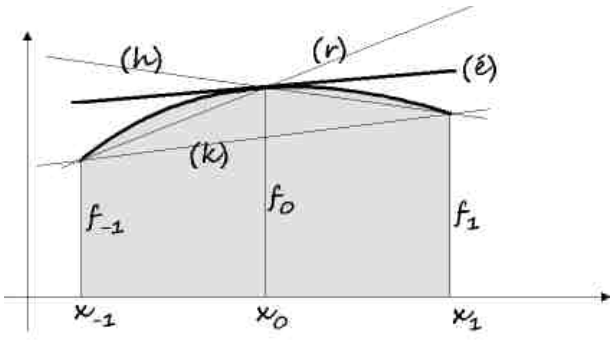
Bizonyítás. Alkalmazzuk a Taylor-sorfejtést a másodrendű tagnál a Lagrange-féle maradéktagot használva egy megfelelő ξ pontban.

$$\Delta f_+ = \frac{f_1 - f_0}{h} = \frac{(f_0 + f_0' h + f''(\xi)h^2/2) - f_0}{h} = f_0' + f''(\xi)h/2 = f_0' + \mathcal{O}(h).$$

$$\Delta f_- = \frac{f_0 - f_{-1}}{h} = \frac{f_0 - (f_0 - f_0' h + f''(\xi)h^2/2)}{h} = f_0' - f''(\xi)h/2 = f_0' + \mathcal{O}(h).$$

Azaz a közelítések valóban elsőrendűek. ■

A fenti bizonyítás szerepelt a konvergenciarendre vonatkozó példaként az 1.3.11. példában. Az elsőrendű közelítés szemléletesen azt jelenti, hogy felezve h értékét a hiba is körülbelül feleződik. Hogyan érhetnénk el magasabb rendű közelítéseket? Tekintsük a 7.2.1. ábrát! A haladó differencia megegyezik a (h)-val jelölt egyenes meredekségével, míg a retrográd differencia az (r)-rel jelölt egyenes meredekségét adja. Az érintőegyenest (é) jelöli. Az ábra jól szemlélteti, hogy az (x_{-1}, f_{-1}), (x_1, f_1) pontokat összekötő egyenes meredeksége közelebb állhat az érintő meredekségéhez, mint a haladó és retrográd differenciák. Ez a meredekség



7.2.1. ábra: Az elsőrendű derivált haladó, retrográd és központi közelítésének szemléltetése.

$$\frac{f_1 - f_{-1}}{2h} = \frac{\Delta f_+ + \Delta f_-}{2},$$

azaz megegyezik a haladó és retrográd differenciák átlagával. A

$$\Delta f_c := \frac{\Delta f_+ + \Delta f_-}{2}$$

értéket az f függvény x_0 pontbeli központi (angolul centered) differenciájának nevezzük. Vizsgáljuk meg, hogy ez a közelítés valóban jobb-e, mint a másik kettő!

7.2.2. tétel.
A központi differencia másodrendű közelítése egy $f \in C^3$ függvény első deriváltjának.

Bizonyítás. Megint a Taylor-sorfejtést alkalmazzuk. Most a harmadfokú tagnál szerepeltetjük a Lagrange-féle maradéktagot megfelelő ξ_1 és ξ_2 konstansokkal. Mivel f''' folytonos, így van olyan ξ érték, melyre $(f'''(\xi_1) + f'''(\xi_2))/2 = f'''(\xi)$.

$$\Delta f_c = \frac{f_1 - f_{-1}}{2h} = \frac{f_0 + f_0'h + f_0''h^2/2 + f'''(\xi_1)h^3/6}{2h} - \frac{f_0 - f_0'h + f_0''h^2/2 - f'''(\xi_2)h^3/6}{2h}$$
$$= f_0' + f'''(\xi)\frac{h^2}{6} = f_0' + \mathcal{O}(h^2).$$

Ez mutatja, hogy ez a közelítés másodrendű. ■

7.3. A második derivált közelítése

Mivel a második derivált az első derivált deriváltja, így kézenfekvőnek tűnik azt a

$$\Delta^2 f_c := \frac{\Delta f_+ - \Delta f_-}{h} = \frac{f_1 - 2f_0 + f_{-1}}{h^2}$$

módon közelíteni. Ezt a közelítést *másodrendű központi differenciának* nevezzük.

7.3.1. tétel.
A másodrendű központi differencia másodrendű közelítése egy $f \in C^4$ függvény második deriváltjának.

Bizonyítás. Taylor-sorfejtést használva a megfelelő ξ_1, ξ_2 és ξ konstansokkal

$$\Delta^2 f_c = \frac{f_0 + f_0'h + f_0''h^2/2 + f_0'''h^3/6 + f''''(\xi_1)h^4/24}{h^2} - \frac{2f_0}{h^2}$$

$$+\frac{f_0 - f_0'h + f_0''h^2/2 - f_0'''h^3/6 + f''''(\xi_2)h^4/24}{h^2} = f_0'' + f''''(\xi)\frac{h^2}{12}.$$

Ez mutatja, hogy ez a közelítés másodrendű. ■

7.4. A deriváltak másfajta közelítései

Természetesen a deriváltakat nem csak a differenciahányados segítségével közelíthetjük. Kézenfekvőnek tűnik az a megközelítés is, hogy az adott pontokra interpolációs polinomot illesztünk, és a deriváltakat az interpolációs polinom deriváltjaival közelítjük. Ezeket a közelítéseket interpolációs differenciálási formuláknak nevezzük. Az alábbi, bizonyítás nélkül közölt tételek azt mutatják, hogy az interpolációs differenciálási formulák pontosan ugyanazokat a deriváltközelítéseket adják, mint amiket az előző fejezetekben a differenciahányadosok segítségével nyertünk.

7.4.1. tétel.
Az $(x_0, f_0), (x_0 + h, f_1)$ pontokra illeszkedő interpolációs polinom (legfeljebb elsőfokú) deriváltja egybeesik a haladó differenciával. Az $(x_0 - h, f_{-1}), (x_0, f_0)$ pontokra illeszkedő interpolációs polinom (legfeljebb elsőfokú) deriváltja egybeesik a retrográd differenciával.

7.4.2. tétel.
Az $(x_0 - h, f_{-1}), (x_0, f_0), (x_0 + h, f_1)$ pontokra illeszkedő interpolációs polinom (legfeljebb másodfokú) deriváltja x_0-ban egybeesik a központi differenciával, második deriváltja pedig a másodrenű központi differenciával.

7.4.3. tétel.
Az $(x_0 - h, f_{-1}), (x_0, f_0), (x_0 + h, f_1)$ pontokra illeszkedő harmadfokú spline interpolációs függvény x_0-pontbeli deriváltja egybeesik a központi differenciával.

7.4.4. megjegyzés. A deriváltakat nem csak az $x_0, x_0 \pm h, x_0 \pm 2h, \ldots$ pontokban közelíthetjük, hanem az adott intervallum bármelyik pontjában. Ezt megtehetjük úgy, hogy a kiszámolt deriváltközelítésekre interpolációs polinomot illesztünk, és ennek értékeivel közelítjük a deriváltakat az adott pontok között. Másik lehetőség, hogy megfelelően módosítjuk a fejezetben megismert képleteket. Pl. az $(f_1 - f_0)/h$ képlettel az f függvény $x_0 + h/2$ pontbeli deriváltja közelíthető. ⋄

7.5. Lépéstávolság-dilemma

A deriváltak közelítése tart a pontos deriváltértékhez, ha a h lépéstávolság nullához tart. Elméletileg tehát minél kisebb h értéket választunk, a közelítés annál pontosabb lesz. Ahogy a következő példa mutatja, a gyakorlatban ez nem így van.

7.5.1. példa. Közelítsük a $\cos''(0.8) = -0.6967067093$ értéket a másodrendű központi differenciával! Tegyük fel, hogy mindent 9 tizedesjegy pontossággal számolunk. A pontos deriváltérték és a közelítés eltérését az alábbi táblázat mutatja.

h	$\lvert$közelítés $- \cos''(0.8)\rvert$
0.1	0.0005804093
0.01	**0.0000167093**
0.001	0.0007067093

A kerekítési hiba miatt (vagy általánosan a lebegőpontos számítás hibája miatt) tehát csökkenő h értékekkel a közelítés hibája csökken, majd újra növekedni kezd. $\diamond$

Vizsgáljuk meg a fenti példa jelenségét általánosan! Tegyük fel, hogy az f_{-1}, f_0, f_1 értékeket megváltoztatjuk ε-nál kisebb értékekkel. Azaz legyen

$$\tilde{f}_{-1} = f_{-1} + \varepsilon_{-1},\ \tilde{f}_0 = f_0 + \varepsilon_0,\ \tilde{f}_1 = f_1 + \varepsilon_1,$$

ahol $|\varepsilon_{-1}|, |\varepsilon_0|, |\varepsilon_1| \leq \varepsilon$. Ekkor

$$\begin{aligned}\frac{\tilde{f}_{-1} - 2\tilde{f}_0 + \tilde{f}_1}{h^2} &= \frac{f_{-1} - 2f_0 + f_1}{h^2} + \frac{\varepsilon_{-1} - 2\varepsilon_0 + \varepsilon_1}{h^2} \\ &= f_0'' + \frac{f''''(\xi)h^2}{12} + \frac{\varepsilon_{-1} - 2\varepsilon_0 + \varepsilon_1}{h^2}.\end{aligned}$$

Tehát

$$\left| f_0'' - \frac{\tilde{f}_{-1} - 2\tilde{f}_0 + \tilde{f}_1}{h^2} \right| \leq \frac{M_4 h^2}{12} + \frac{4\varepsilon}{h^2},$$

ahol M_4 az f függény abszolút értékben vett negyedik deriváltjának egy felső becslése. A képlet mutatja, hogy akkor lesz kicsi a hiba, ha h se nem túl nagy, se nem túl kicsi. Ezt a megfigyelést *lépéstávolság dilemmának* nevezzük. Egy közelítő optimális értéket úgy nyerhetünk, hogy a felső becslést minimalizáljuk h-ban. Az így nyert optimális érték a

$$h_{opt.} \approx \sqrt[4]{48\varepsilon/M_4}$$

képlettel számítható. A példában szereplő konkrét feladatra így az $M_4 = 1$ és $\varepsilon = 5 \cdot 10^{-10}$ választás mellett a $h_{opt.} = 0.012$ érték adódik, ami nagyjából meg is felel a numerikus kísérlet során nyert optimumnak.

Az 1.3. fejezetben tárgyaltuk a Richardson-extrapolációt, ami két alacsonyabb rendű közelítés megfelelő lineáris kombinációjának segítségével ad egy magasabbrendű közelítést és nem a h paraméter további csökkentésével (ami az előbbiek alapján a hiba növekedéséhez vezethet). Nézzük meg a módszer hatékonyságát az előző példán! Látjuk, hogy a $h = 0.01$ lépéstávolságnál kisebb értékek pontatlanabb közelítését adják a deriváltnak. A $h = 0.01$ lépéstávolság 0.69669 közelítést ad (a közelítés hibája a táblázatban található), míg a kétszer akkora $h = 0.02$-es 0.6966825-est (ennek hibája 0.00002420935). Mivel a közelítés másodrendű, ezért az (1.3.3) formulát $r = 2$-vel alkalmazva

$$\frac{2^2 0.69669 - 0.6966825}{2^2 - 1} = -0.6966925,$$

melynek hibája kisebb (0.00001420935), mint a $h = 0.01$-es vagy a $h = 0.001$ értékekkel nyert közelítéseké.

7.6. Feladatok

Numerikus deriválás

7.6.1. feladat. Tegyük fel, hogy adottak az $(x_0 - h_-, f_{-1})$, (x_0, f_0) és $(x_0 + h_+, f_1)$ pontok, ahol $h_- \neq h_+$. Adjunk meg egy másodrendű közelítést az x_0-beli első deriváltra!

7.6.2. feladat. Igazoljuk, hogy

$$\frac{-f_2 + 8f_1 - 8f_{-1} + f_{-2}}{12h}$$

az első derivált egy negyedrendű központi közelítése!

7.6.3. feladat. Igazoljuk, hogy

$$\frac{-3f + 4f_1 - f_2}{2h}$$

az első derivált egy másodrendű haladó közelítése!

7.6.4. feladat. Igazoljuk a 7.4.1., 7.4.2. és 7.4.3. tételeket!

7.6.5. feladat. Alkalmazzunk Richardson-extrapolációt az $f(x) = 1/x$ függvény esetén az $f'(0.05)$ érték meghatározására. Válasszunk h-nak 0.0016-ot és 0.0008-at!

7.6.6. feladat. Egy f függvény függvényértékeit tartalmazza az alábbi táblázat. Közelítsük $f'(1)$, $f''(1)$ és $f'''(1)$ értékeit!

x	1.00	1.05	1.10	1.15	1.20	1.25	1.30
$f(x)$	1.00000	1.02470	1.04881	1.07238	1.09544	1.11803	1.14017

Ellenőrző kérdések

1. Milyen módon közelíthetjük egy függvény első és másodrendű deriváltjait függvényértékeinek segítségével?
2. Mit jelent az, hogy egy deriváltközelítés elsőrendű vagy másodrendű?
3. Igaz-e, hogy a másodrendű közelítés minden esetben pontosabban közelíti a deriváltat mint az elsőrendű?
4. Hova tart a derivált közelítő értéke, ha a lépéstávolsággal nullához tartunk?
5. Mi az a lépéstávolság-dilemma?
6. Mire használható a Richardson-extrapoláció?

8. Numerikus integrálás

Ebben a fejezetben azt vizsgáljuk, hogy egy függvény néhány helyen vett függvényértékének segítségével hogyan lehet közelíteni a függvény határozott integrálját. Az ún. interpolációs módszerekkel fogunk foglalkozni, azaz azokkal, melyek a függvényértékekre illesztett interpolációs polinomok integráljával közelítik a tényleges integrálértéket. Részletesen vizsgáljuk a trapéz-, érintő- és Simpson-formulákat ill. a Gauss-féle integrálformulát.

8.1. A numerikus integrálás alapfeladata

Ismert, hogy ha egy $[a, b]$ intervallumon egy f függvény integrálható, és van ezen az intervallumon primitív függvénye ($F' = f$), akkor a függvény határozott integrálja a Newton–Leibniz-szabály szerint az

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

képlettel számolható. Mielőtt arra gondolnánk, hogy ez a képlet minden integrállal kapcsolatos kérdésünkre választ ad, felsorolunk néhány olyan esetet, amikor a képlet nem alkalmazható közvetlenül.

Nem alkalmazható a képlet akkor, ha nem ismerjük magát az f függvényt vagy a primitív függvényét zárt alakban. Az első eset fordul elő akkor, ha pl. az f függvény értékei mérési eredmények, a második pedig akkor, ha az f függvény primitív függvényét nem akarjuk, vagy nem tudjuk explicit módon meghatározni. Az utóbbira példák pl. a $(\sin x)/x$, $\sin x^2$ vagy az e^{-x^2} függvények, melyek primitív függvényeit - amelyek léteznek, hiszen a függvények folytonosak - nem tudjuk zárt alakban előállítani. Van olyan eset is, amikor nem is akarjuk az integrált pontosan kiszámítani, mert az sokkal komplikáltabb lenne, mint annak egy jó közelítését adni. Pl. egy számítógépes programnak gyakran egyszerűbb egy jó közelítést mondania egy adott f függvény $[a, b]$ intervallumon vett integráljára, mint valamilyen kompjuteralgebrai módszerrel szimbólikusan meghatározni f primitív függvényét és azzal alkalmazni a Newton–Leibniz-szabályt. Tipikus példa erre a differenciálegyenletek megoldása, ahol a megoldás formális előállítása általában nehéz feladat, míg annak numerikus közelítése egyszerűen előállítható.

Természetesen rögtön eszünkbe jut néhány lehetséges módszer a fenti problémák kiküszöbölésére. Ha nem lehet előállítani a primitív függvényt zárt alakban, akkor azért általában hatványsor formájában előállítható az. Helyettesítsük ilyenkor be a Newton–Leibniz-képletbe a hatványsor egy részletösszegét, és annak integrálásával adjuk meg az integrál közelítését! Másik lehetőség, hogy mivel az integrált a közelítő összegek határértékeként definiáltuk, egy alkalmasan választott közelítő összeg megfelelő közelítését adhatja a tényleges integrálnak. Az első módszer alkalmazására csak egy példát mutatunk, hiszen általában egy-egy konkrét integrál meghatározására alkalmazható csak a gyakorlatban is. A második módszer pedig az ún. interpolációs típusú integrálási módszerek közé tartozik, melyeket részletesen ismertetünk a továbbiakban.

8.1.1. példa. Határozzuk meg az $\int_0^1 e^{-x^2}\,\mathrm{d}x$ integrált 10^{-6}-nál kisebb hibával! Mivel

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots,$$

ezért

$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} \pm \dots$$

és a sor tagonként integrálható. Tehát

$$\int_0^1 e^{-x^2}\,\mathrm{d}x = \left[x - \frac{x^3}{3} + \frac{x^5}{5\cdot 2!} - \frac{x^7}{7\cdot 3!} \pm \dots\right]_0^1 = 1 - \frac{1}{3} + \frac{1}{10} - \frac{1}{42} \pm \dots = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)k!}.$$

Vizsgáljuk meg, hogy hányadik tagig kell elmennünk a sorfejtésben, hogy az adott pontossággal tudjuk meghatározni az integrál értékét. Mivel az integrál értékét egy Leibniz-sor adja, ha annak k-adik tagja kisebb 10^{-6}-nál, akkor az első $k-1$ tag összege már 10^{-6}-nál jobban megközelíti a sor összegét. A $k = 9$ választás esetén

$$\frac{1}{(2k+1)k!} = \frac{1}{6894720} < 10^{-6},$$

így

$$\sum_{k=0}^{8} \frac{(-1)^k}{(2k+1)k!} = \frac{1098032417}{1470268800} \approx 0.74682426573970691618.$$

megfelelő közelítése lesz az integrálnak. ⋄

8.1.2. megjegyzés. Tekintsük az következő egyszerű határozott integrált:

$$\int_0^1 \cos x\,\mathrm{d}x = [\sin x]_0^1 = \sin 1.$$

Ebben az esetben a Newton–Leibniz-tétel alkalmazása nem okoz nehézséget, de a végeredményt ebben az esetben is csak a szinusz függvény $x = 1$ helyen vett Taylor-sorának egy részletösszegével tudjuk közelíteni:

$$\sin 1 = 1 - \frac{1}{3!} + \frac{1}{5!} - \frac{1}{7!} \pm \dots.$$

Így sok esetben nincs lényeges különbség aközött, hogy az integrál kiszámításánál vagy a kiszámított érték számszerű felírásánál alkalmazunk közelítéseket. ⋄

Azt az eljárást, amikor egy függvény határozott integrálját nem a Newton–Leibniz-szabály segítségével adjuk meg, hanem azt valamilyen módszerrel közelítjük, *numerikus integrálásnak* hívjuk. A numerikus integrálás alapfeladata tehát a következő. Ismert egy $[a,b]$ intervallumon integrálható f függvény értéke néhány pontban. Legyenek ezek a pontok

$$a \le x_0 < x_1 < \dots < x_n \le b, \tag{8.1.1}$$

és az itteni függvényértékek rendre $f_0 := f(x_0), \dots, f_n := f(x_n)$. Adjunk becslést az $\int_a^b f(x)\,\mathrm{d}x$ integrál értékére! A becsléstől természetesen elvárjuk, hogy könnyen kiszámítható legyen. Azt is

elvárjuk, hogy ha az adott pontokat egyre sűrűbben vesszük fel (ha lehetséges), akkor a becsléseink "tartsanak" a pontos integrálértékhez és "jó tulajdonságú" függvényekre legyen gyors ez a konvergencia.

Jelölje egy f integrálható függvény $[a,b]$-n vett pontos integrálértékét $I(f)$, és állítsuk elő ennek egy közelítését az adott függvényértékek lineáris kombinációjaként az

$$I_n(f) = \sum_{k=0}^{n} a_k f_k \tag{8.1.2}$$

alakban. Az n alsó index arra utal, hogy a közelítés függ az osztópontok számától és elhelyezkedésétől. A (8.1.2) formulát numerikus integrálási vagy *kvadratúraformulának*[1] nevezzük, a benne szereplő a_k együtthatókat pedig súlyoknak hívjuk.

8.1.3. definíció.
A (8.1.2) kvadratúraformulát *zárt kvadratúraformulának nevezzük, ha szerepelnek benne az a és b-beli függvényértékek. Ha ezek nem szerepelnek benne, akkor* nyílt kvadratúraformuláról beszélünk.

Jelölje h a (8.1.1) felosztásban a legnagyobb távolságot a szomszédos osztópontok között.

8.1.4. definíció.
Azt mondjuk, hogy az $I_n(f)$ kvadratúraformula *konvergenciarendje (legalább)* $r \geq 1$, *ha* $|I_n(f) - I(f)| = O(h^r)$ *(az 1.3.10. definíció speciálisan kvadratúraformulákra megfogalmazva).*

8.1.5. definíció.
Azt mondjuk, hogy az $I_n(f)$ kvadratúraformula *pontossági rendje* $r \geq 1$, *ha minden* P_{r-1}*-beli polinomra* $I(p) = I_n(p)$, *de van olyan* $p \in P_r$, *melyre* $I(p) \neq I_n(p)$.

A kvadratúraformulák fenti kétféle rendje látszólag különböző dolog. Az első a konvergenciasebességet adja meg, ha h értéke tetszőlegesen kicsi lehet, a második pedig azt adja meg, hogy hanyadfokú a legalacsonyabb fokszámú polinom, amit már nem integrál ki pontosan a képlet. A későbbiekben látni fogjuk, hogy ez a két rend összefügg (megegyezik).

8.2. Newton–Cotes-féle kvadratúraformulák

Ebben a fejezetben egy gyakran alkalmazott kvadratúramódszerrel ismerkedünk meg: a Newton–Cotes[2]-formulákkal.

[1] Az ókori görögök úgy határozták meg a síkidomok területét, hogy konstruáltak egy, a síkidommal megegyező területű négyzetet és annak számították ki a területét. Ezt az eljárást kvadratúrának hívták. Innét kapta a numerikus integrálási eljárás is a kvadratúra nevet, hiszen ez az eljárás is egy síkidom (az x-tengely és a függvény grafikonja közé eső tartomány) területét számolja ki.

[2] Roger Cotes (1682–1716) angol matematikus volt. Ő szerkesztette Newton Principiájának második kiadását. Fontos eredményeket ért el a logaritmus, az integrálkalkulus és a numerikus módszerek, főleg az interpolációs módszerek területén. Bővebb életrajz a http://www-history.mcs.st-and.ac.uk/Biographies/Cotes.html oldalon található

8.2.1. definíció.

Egy kvadratúraformulát *interpolációs kvadratúraformulának nevezünk, ha az alappontokbeli függvényértékekre illesztett interpolációs polinom integrálját adja meg. Ha egy interpolációs kvadratúraformulában az alappontok egyforma távol vannak egymástól, akkor a formulát Newton–Cotes-formulának hívjuk.*

Vizsgáljuk meg, hogy hogyan állíthatók elő a zárt Newton–Cotes-formulák. Adott (x_i, f_i) $(i = 0, \dots, n)$ pontok esetén (zárt formulánál $x_0 = a$ és $x_n = b$) a kvadratúraformula értéke definíció szerint a pontokra illesztett interpolációs polinom integrálját adja, azaz az interpolációs feladatoknál megismert jelölésekkel

$$I_n(f) = \int_a^b L_n(x)\,\mathrm{d}x = \int_a^b \left(\sum_{k=0}^{n} f_k l_k(x)\right)\,\mathrm{d}x = \sum_{k=0}^{n} f_k \left(\overbrace{\int_a^b l_k(x)\,\mathrm{d}x}^{a_k}\right).$$

Innét látható, hogy a formula súlyai a Lagrange-féle alappolinomok $[a, b]$ intervallumon vett integráljai lesznek, azaz

$$a_k = \int_a^b l_k(x)\,\mathrm{d}x,\ \ k = 0, \dots, n.$$

Alkalmazzuk a fenti integrálokban az $x = a + t(b - a)$ helyettesítést, ahol $t \in [0, 1]$. Ekkor a helyettesítéses integrálás szabálya szerint

$$a_k = \int_a^b l_k(x)\,\mathrm{d}x = (b - a)\int_0^1 l_k(a + t(b - a))\,\mathrm{d}t =: (b - a)N_{\text{zárt}}^{n,k}.$$

A második integrál csak az n és k indexektől függ, hiszen értéke nem más, mint az $(i/n, \delta_{ki})$ (δ_{ki} a Kronecker-szimbólumot jelöli) $(i = 0, \dots, n)$ pontokra illesztett (egyértelműen meghatározott) interpolációs polinom integrálja a $[0, 1]$ intervallumon. Vezessük be ezen integrálok értékére az $N_{\text{zárt}}^{n,k}$ jelölést! Az $N_{\text{zárt}}^{n,k}$ értékeket zárt Newton–Cotes-együtthatóknak nevezzünk. Ezeket előre kiszámíthatjuk, táblázatba foglalhatjuk, és amikor szükségesek, felhasználhatjuk őket. Néhány zárt Newton–Cotes-együttható értékét az alábbi táblázatban közöljük.

$N_{\text{zárt}}^{n,k}$	$k = 0$	$k = 1$	$k = 2$	$k = 3$	
$n = 1$	1/2	1/2			← trapézformula együtthatói
$n = 2$	1/6	4/6	1/6		← Simpson-formula együtthatói
$n = 3$	1/8	3/8	3/8	1/8	

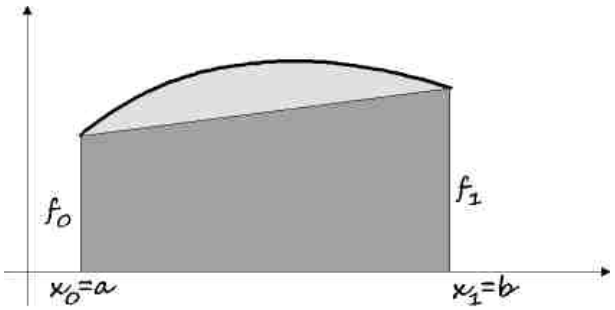
Az $n = 1$ és $n = 2$ esetet külön is érdemes megvizsgálni. Az $n = 1$ esetben csak az $x_0 = a$ és $x_1 = b$ végpontokbeli függvényértékekből számítjuk a közelítő integrált 1/2-1/2 súlyokkal

$$I_1(f) = (b - a)\frac{f_0 + f_1}{2}$$

alakban. Ez az érték pontosan a 8.2.1. ábrán látható trapéz területével egyezik meg, így a fenti képletet trapézformulának nevezzük.

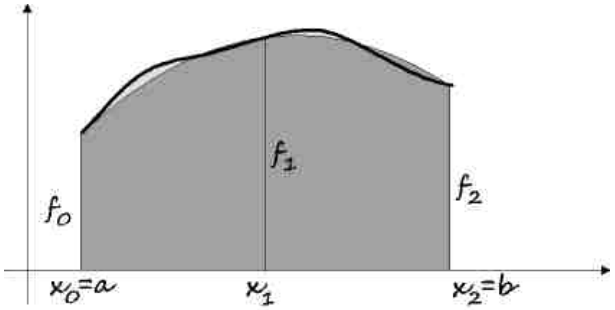
Az $n = 2$ esetben az $x_0 = a$, $x_1 = (a+b)/2$ és $x_2 = b$ pontokbeli függvényértékekből számítjuk a közelítő integrált 1/6-4/6-1/6 súlyokkal

$$I_2(f) = (b - a)\frac{f_0 + 4f_1 + f_2}{6}$$



8.2.1. ábra: A trapézformula az ábrán látható besötétített trapéz területével közelíti az integrál értékét.

alakban. Ezt a képletet Simpson-formulának[3] nevezzük. A közelítés a pontokra illesztett legfeljebb másodfokú polinom integrálját adja meg az $[a, b]$ intervallumon (8.2.2. ábra).



8.2.2. ábra: A Simpson-formula az ábrán látható besötétített, parabolával és egyenes szakaszokkal határolt tartomány területével közelíti az integrál értékét.

8.2.2. példa. Alkalmazzuk a trapézformulát az $f(x) = x^2 - 2x + 2$ függvény integráljának közelítésére az $[1, 3]$ intervallumon. Az osztópontok az $x_0 = 1$ és $x_1 = 3$ végpontok, ezen pontokban a függvényértékek pedig rendre $f_0 = f(1) = 1$ és $f_1 = f(3) = 5$. A súlyokat az előző táblázat $n = 1$ sorából olvashatjuk ki. Így az integrál közelítő értékére

$$I_1(f) = (3 - 1)(1 \cdot 1/2 + 5 \cdot 1/2) = 6$$

adódik. Az integrál pontos értéke $14/3$. ⋄

8.2.3. példa. Alkalmazzuk a Simpson-formulát a 8.2.2. példában szereplő integrál meghatározására! Az osztópontok most az $x_0 = 1$, $x_1 = 2$ és $x_2 = 3$ pontok, ezen pontokban a

[3] Thomas Simpson (1710–1761) angol matematikus. Főbb eredményei az interpoláció és mumerikus integrálás területéről kerültek ki. A róla elnevezett Simpson-formulát Newtonnak tulajdonítják. A nemlineáris egyenetek megoldására tanult Newton-módszert viszont Simpson írta le először a mai formájában. Bővebb életrajz található a http://www-history.mcs.st-and.ac.uk/Biographies/Simpson.html oldalon.

függvényértékek pedig rendre $f_0 = f(1) = 1$, $f_1 = f(2) = 2$ és $f_2 = f(3) = 5$. A súlyokat az előző táblázat $n = 2$ sorából olvashatjuk ki. Így az integrál közelítő értékére

$$I_2(f) = (3-1)(1 \cdot 1/6 + 2 \cdot 4/6 + 5 \cdot 1/6) = 14/3$$

adódik. Vegyük észre, hogy a kvadratúraformula ebben az esetben az integrál pontos értékét adja, hiszen a három alappontra illesztett interpolációs polinom azonos magával az f függvénnyel. ◇

Térjünk át a nyílt Newton–Cotes-formulák megadására. Ekkor a szomszédos osztópontok egyforma távol vannak egymástól, de a végpontokbeli függvényértékek nem szerepelnek a kvadratúraképletben. Vezessük be az $a = x_{-1}$ és $b = x_{n+1}$ jelöléseket. Így $x_i = a+(i+1)h$ $(i = -1, \dots, n+1)$, ahol $h = (b-a)/(n+2)$. Hasonlóan a zárt formulák megadásához, az (x_i, f_i) $(i = 0, \dots, n)$ pontokra illesztett interpolációs polinom integrálját kell kiszámolnunk.

$$I_n(f) = \sum_{k=0}^{n} f_k \left(\overbrace{\int_a^b l_k(x)\,\mathrm{d}x}^{a_k} \right),$$

ahol az a_k súlyok megint a k-adik Lagrange-féle alappolinom $[a, b]$ intervallumon vett integráljával egyenlőek. Helyettesítéses intergálással ez az integrál az

$$a_k = (b-a)N^{n,k}_{\text{nyílt}}$$

alakban írható, ahol $N^{n,k}_{\text{nyílt}}$ az $((i+1)/(n+2), \delta_{ik})$ $(i = 0, \dots, n)$ pontokra illesztett interpolációs polinom $[a, b]$ intervallumon vett integrálja.

Az alábbi táblázat tartalmazza néhány nyílt Newton–Cotes-együttható értékét.

$N^{n,k}_{\text{nyílt}}$	$k=0$	$k=1$	$k=2$	
$n=0$	1			← érintőformula együtthatói
$n=1$	1/2	1/2		
$n=2$	2/3	−1/3	2/3	

A nyílt Newton–Cotes-formulák közül az $n = 0$ esetet érdemes megvizsgálni részletesebben. Ekkor egyetlen egy pontban számítjuk csak ki a függvényértéket, nevezetesen az $x_0 = (a+b)/2$ pontban, és az integrált az

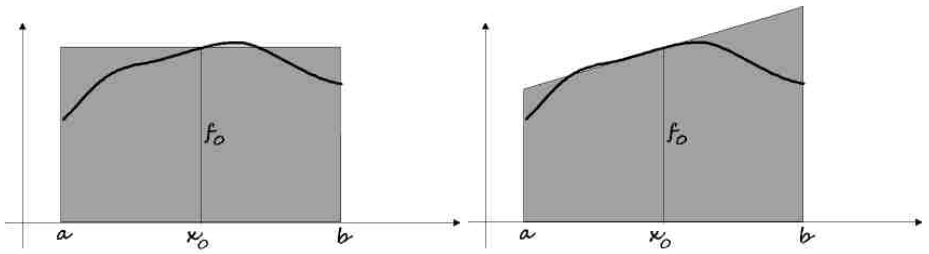
$$I_0(f) = (b-a)f_0$$

értékkel közelítjük. Ez az érték nem más, mint egy $b-a$ alapú és f_0 magasságú téglalap területe (lásd a besötétített területet a 8.2.3. ábra bal oldalán). Vegyük észre, hogy ez a terület ugyanakkora, mint az (x_0, f_0) pontban az integrálandó függvényhez húzott érintőegyenes (amennyiben létezik) integrálja az $[a, b]$ intervallumon (lásd a besötétített területet a 8.2.3. ábra jobb oldalán). Emiatt a képletet érintőformulának hívjuk.

8.2.4. példa. Közelítsük a 8.2.2. példában szereplő integrált az érintőformula segítségével! Ekkor $x_0 = 2$ és $f_0 = f(2) = 2$. A táblázatból kiolvasva az $n = 0$ értékhez tartozó együtthatót (1) kapjuk, hogy

$$I_0(f) = (3-1) \cdot 1 \cdot 2 = 4.$$

◇



8.2.3. ábra: Az érintőformula az intervallum felénél a grafikonhoz húzott érintő integráljával közelíti a pontos integrál értékét.

8.2.5. megjegyzés. A trapézformula által adott

$$\frac{(b-a)(f(a)+f(b))}{2}$$

értéket és az érintőformula által adott

$$(b-a)f((a+b)/2)$$

értéket 1:2 arányban súlyozva kapjuk a

$$\frac{(b-a)(f(a)+f(b))/2+2(b-a)f((a+b)/2)}{3}=(b-a)\frac{f(a)+4f((a+b)/2)+f(b)}{6}$$

értéket, ami pontosan a Simpson-formula által adott közelítés. ◇

Vizsgáljuk most meg az interpolációs kvadratúraformulák pontossági rendjét!

8.2.6. tétel.
Egy $n+1$ alappontos kvadratúraformula akkor és csak akkor pontos minden legfeljebb n-edfokú polinomra, ha interpolációs kvadratúraformula.

Bizonyítás. A feltétel elégségessége nyilvánvaló. A másik irány igazolásához induljunk ki abból, hogy a formulának pontosnak kell lennie minden n-edfokú polinomra is, így pontos az $l_j(x)$ $(j=0,\dots,n)$ Lagrange-féle alappolinomokra is. Azaz

$$\int_a^b l_j(x)\,\mathrm{d}x=\sum_{k=0}^{n}a_k l_j(x_k)=a_j.$$

Tehát a súlyok a Lagrange-féle alappolinomok integráljai – csakúgy mint az interpolációs kvadratúraformuláknál. Ezt akartuk megmutatni. ■

Jelentse $N^{n,k}$ a megfelelő zárt vagy nyílt Newton–Cotes-együtthatót. Ezen együtthatók néhány fontos tulajdonságát adja meg az alábbi tétel.

8.2.7. tétel.
A Newton–Cotes-együtthatókra igaz az alábbi két tulajdonság:

$$\sum_{k=0}^{n}N^{n,k}=1,\quad N^{n,k}=N^{n,n-k}\ (k=0,\dots,n).$$

Bizonyítás. Az előző tétel miatt

$$\int_a^b 1\,\mathrm{d}x = b-a = \sum_{k=0}^{n}(N^{n,k}(b-a)1) = (b-a)\sum_{k=0}^{n} N^{n,k}.$$

Ez igazolja az első állítást. A második pedig egyszerűen abból következik, hogy l_k és l_{n-k} grafikonjai egymás tükörképei az $x=(a+b)/2$ egyenesre, így integráljuk is megegyezik. ■

Nyilvánvaló, hogy a trapéz- és érintő-formulák pontossági rendje 2, hiszen lineáris függvényekre pontosak, és nyilván vannak olyan másodfokú polinomok, amelyekre pedig nem pontosak. Az is nyilvánvaló az előző tétel miatt, hogy a Simpson-formula pontossági rendje legalább 3. Valójában a pontossági rend 4. Ez egy általánosabb állítás következményeként könnyen adódik.

8.2.8. tétel.
Az $n+1$ alappontos Newton–Cotes-formulák pontossági rendje páratlan n esetén $n+1$, és páros n esetén $n+2$.

Bizonyítás. Az állítás első része a 8.2.6. tétel következménye. A második rész igazolásához tegyük fel, hogy n páros, és így $n+1$ páratlan alappontunk van. Azt kell megmutatnunk, hogy egy tetszőleges p_{n+1} $n+1$-edfokú polinom integrálját a formula pontosan számolja ki. Írjuk fel a p_{n+1} $n+1$-edfokú polinomot $(x-(a+b)/2)$ polinomjaként

$$p_{n+1}(x) = \alpha_{n+1}\left(x-\frac{a+b}{2}\right)^{n+1} + \underbrace{\alpha_n\left(x-\frac{a+b}{2}\right)^{n} + \ldots + \alpha_0}_{\text{Erre pontos a formula.}},$$

és használjuk ki, hogy a 8.2.6. tétel miatt a formula pontos minden legfeljebb n-edfokú polinomra. Így már csak azt kell megmutatnunk, hogy az első tagra pontos a kvadratúraformula. A Newton–Cotes-együtthatók szimmetriáját ($N^{n,k}=N^{n-k,k}$) és a következő képletben f-fel jelölt függvény grafikonjának $((a+b)/2,0)$ pontra való szimmetriáját ($f((a+b)/2-x) = -f((a+b)/2+x)$) kihasználva kapjuk, hogy

$$\int_a^b \underbrace{\alpha_{n+1}\left(x-\frac{a+b}{2}\right)^{n+1}}_{=:f(x)}\,\mathrm{d}x = (b-a)\sum_{k=0}^{n} \underbrace{N^{n,k}}_{N^{n,n-k}}\underbrace{f(x_k)}_{-f(x_{n-k})} = 0.$$

Tehát erre a polinomra is pontos értéket ad a formula, hiszen a függvény integrálja az f függvény grafikonjának $((a+b)/2,0)$ pontra való szimmetriája miatt valóban nulla. Ezt akartuk megmutatni. ■

Sok alappont esetén nem célszerű a Newton–Cotes-formulákat használni (ezért is csak kicsi n értékekre adtuk meg a korábbi táblázatokban az együtthatók értékét). Ennek oka egyrészt az, hogy növekvő n értékek esetén egyre több Newton–Cotes-együtthatót kellene kiszámolnunk, vagy valamilyen táblázatból kiolvasnunk, másrészt nagyobb n értékek esetén az $N^{n,k}$ értékek között megjelennek negatív értékek is, melyekkel pozitív értékek súlyozásánál fennáll a számítógépes számolások során a kiegyszerűsödés veszélye. Nagyobb n értékek esetén az ún. összetett kvadratúraformulákat használjuk. Ezeket mutatjuk be a következő fejezetben.

8.3. Összetett kvadratúraformulák

Sok alappont esetén a Newton–Cotes-formulák közvetlen alkalmazása helyett úgy járhatunk el, hogy az integrálási intervallumot részintervallumokra osztjuk, majd ezen intervallumokon az előző

fejezetben megismert trapéz-, Simpson- és érintőformulákat alkalmazzuk. Így kapjuk az ún. összetett trapéz-, összetett Simpson- és összetett érintőformulákat.

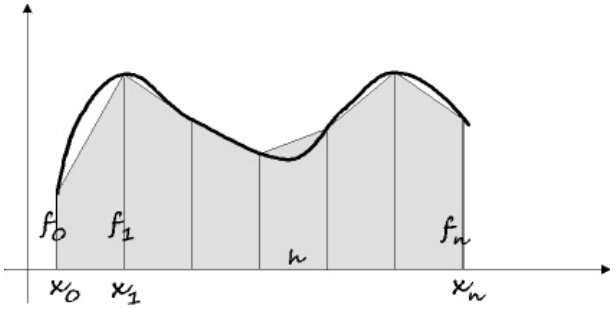
8.3.1. Összetett trapézformula

Legyen f egy, az $[a,b]$ intervallumon integrálható függvény. Osszuk fel az $[a,b]$ intervallumot n egyenlő részre az $a = x_0 < x_1 \ldots < x_n = b$ osztópontokkal, és jelöljük h-val az osztóintervallumok hosszait, azaz legyen $h = (b-a)/n$. Az f függvény határozott integrálját felírhatjuk az

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x$$

összegként. Ha most az összeg minden integrálját a trapézformulával közelítjük, akkor az ún. *összetett trapézformulához* jutunk. A formula tehát a 8.3.1. ábrán besötétített tartomány területével közelíti az integrál pontos értékét. Ismét használva az $f_i = f(x_i)$ $(i = 0, \ldots, n)$ jelölést az összetett trapézfomula képlete tehát

$$I_{n,\mathrm{trap}}(f) = \frac{h}{2} f_0 + h(f_1 + \ldots + f_{n-1}) + \frac{h}{2} f_n = h\left(\frac{1}{2} f_0 + f_1 + \ldots + f_{n-1} + \frac{1}{2} f_n\right).$$



8.3.1. ábra: Az összetett trapézformula a besötétített tartomány területével közelíti az integrál pontos értékét.

Látható, hogy az összetett trapézformula egy könnyen alkalmazható zárt kvadratúraformula. Az $s_n \le I_{n,\mathrm{trap}}(f) \le S_n$ egyenlőtlenség miatt, ahol s_n és S_n az adott felosztáshoz tartozó alsó és felső összeg, ha az f függvény Riemann-integrálható, akkor a felosztás finomításával a kvadratúraformula értéke az integrál pontos értékéhez tart.

Az összetett trapézformula pontossági rendje nyilvánvalóan 2, hiszen a trapézformula pontossági rendje is 2. Vizsgáljuk meg a konvergenciarendet először egy numerikus kísérlet segítségével. Számítsuk ki a formula segítségével az

$$\int_0^1 \sin x/x\,\mathrm{d}x \approx 0.9460830704$$

integrált úgy, hogy a $[0,1]$ intervallumot n egyenlő részre osztjuk. A pontos és a közelítő értékek eltérését az n osztóintervallumszám függvényében az alábbi táblázat mutatja.

n	$I_n(f)$	$\lvert I(f) - I_n(f)\rvert$
1	0.9207354924	0.25×10^{-1}
10	0.9458320719	0.25×10^{-3}
100	0.9460805606	0.25×10^{-5}
1000	0.9460830704	0.27×10^{-7}

A táblázat azt sejteti, hogy a módszer konvergenciarendje is 2. A konvergenciarend igazolásához szükségünk lesz az alábbi, analízisből ismert integrál-középértéktételre.

8.3.1. tétel.
Ha ϕ $[a,b]$-n integrálható nemnegatív függvény, és g folytonos függvény, akkor van olyan $\eta \in [a,b]$, hogy

$$\int_a^b \phi(x)g(x)\,\mathrm{d}x = g(\eta)\int_a^b \phi(x)\,\mathrm{d}x.$$

8.3.2. tétel.
Az összetett trapézformula hibája $f \in C^2[a,b]$ függvények esetén

$$I_{n,\mathrm{trap}}(f) - I(f) = \frac{(b-a)h^2}{12} f''(\eta),$$

ahol $\eta \in [a,b]$.

Bizonyítás. Legyen k egy rögzített index az $\{1,\dots,n\}$ halmazból. Az (x_{k-1}, f_{k-1}) és (x_k, f_k) pontokra illesztett $L_1(x)$ interpolációs polinom (ebben az esetben lineáris függvény) interpolációs hibája egy adott $x \in (x_{k-1}, x_k)$ pontban

$$L_1(x) - f(x) = -\frac{f''(\xi_x)}{2}(x - x_{k-1})(x - x_k), \tag{8.3.1}$$

ahol ξ_x egy az x ponttól függő megfelelő konstans (6.2.5. tétel). Definiáljuk a $g : [x_{k-1}, x_k] \to \mathbb{R}$ függvényt a $g(x) = f''(\xi_x)$ módon, ha $x \in (x_{k-1}, x_k)$, és legyen $g(x_{k-1}) = f''(x_{k-1})$ és $g(x_k) = f''(x_k)$. Ez a g függvény folytonos az $[x_{k-1}, x_k]$ intervallumon. Az, hogy az intervallum belsejében folytonos, onnét következik, hogy ha a (8.3.1) egyenlőséget elosztjuk az $(x - x_{k-1})(x - x_k)$ szorzattal, akkor a bal oldalon álló függvény kétszer folytonosan deriválható x szerint. Tehát g-nek is ilyennek kell lennie. Másrészt

$$\lim_{x\to x_{k-1}} g(x) = \lim_{x\to x_{k-1}} f''(\xi_x) = \lim_{x\to x_{k-1}} \frac{-2(L_1(x) - f(x))}{(x - x_{k-1})(x - x_k)} = f''(x_{k-1}) = g(x_{k-1}),$$

ahol az utolsó előtti egyenlőséget a L'Hospital-szabály kétszeri alkalmazásával nyertük, és hasonlóan

$$\lim_{x\to x_k} g(x) = g(x_k).$$

Tehát g folytonos a teljes $[x_{k-1}, x_k]$ intervallumon. Integráljuk a (8.3.1) egyenlőség mindkét oldalát az $[x_{k-1}, x_k]$ intervallumon, és alkalmazzuk a 8.3.1. integrál-középértéktételt ($-(x - x_{k-1})(x - x_k)/2$ nemnegatív az adott intervallumon)

$$\begin{aligned}
\frac{f_k + f_{k-1}}{2}h - \int_{x_{k-1}}^{x_k} f(x)\,\mathrm{d}x &= -\int_{x_{k-1}}^{x_k} \frac{f''(\xi_x)(x - x_{k-1})(x - x_k)}{2}\,\mathrm{d}x \\
&= -\int_{x_{k-1}}^{x_k} \frac{g(x)(x - x_{k-1})(x - x_k)}{2}\,\mathrm{d}x \\
&= g(\eta_k)\int_{x_{k-1}}^{x_k} \frac{(x - x_{k-1})(x_k - x)}{2}\,\mathrm{d}x \\
&= f''(\eta_k)\frac{h^3}{12},
\end{aligned}$$

ahol η_k egy az $[x_{k-1}, x_k]$ intervallumba eső megfelelő konstans.

Mivel n darab intervallum van, ezért a teljes hiba

$$I_{n,\mathrm{trap}}(f) - I(f) = \sum_{k=1}^{n} f''(\eta_k)\frac{h^3}{12} = nf''(\eta)\frac{h^3}{12} = \frac{(b-a)h^2}{12}f''(\eta),$$

ahol $\eta \in [a,b]$ egy megfelelően választott konstans. Azt, hogy η megválasztható a fenti módon, a következő módon láthatjuk be. A

$$\frac{\sum_{k=1}^{n} f''(\eta_k)}{n}$$

hányados az f'' folytonos függvény n darab $[a,b]$-beli függvényértékének számtani közepe. Ez az érték nyilvánvalóan a legkisebb és legnagyobb függvényérték közé esik. A Bolzano-tétel miatt pedig ezt az értéket fel is veszi az f'' függvény. Tehát η az a hely, ahol f'' felveszi a fenti átlagot. Ezt akartuk megmutatni. ■

8.3.3. megjegyzés. Vegyük észre, hogy a fenti tétel pontosan megmondja a kvadratúraképlet hibáját, amennyiben η értéke ismert. Mivel η értékét általában nem ismerjük, így a képlet nem használható jobban, mint a belőle következő

$$|I_{n,\mathrm{trap}}(f) - I(f)| \le \frac{(b-a)h^2}{12}M_2$$

becslés, ahol a szokott módon M_2 az f függvény második deriváltjának egy felső korlátja az (a,b) intervallumon. ◇

8.3.4. megjegyzés. A bizonyítás két fontos lépése volt annak igazolása, hogy $g(x)$ folytonosan függ x-től és az, hogy az η érték hogyan választható meg az η_k értékek ismeretében. Ezekhez hasonló állítások később is szerepelni fognak, de akkor már nem igazoljuk részletesen őket. ◇

8.3.2. Összetett érintőformula

Az összetett trapézformula esetén bevezetett jelöléseket fogjuk most is alkalmazni, kiegészítve még az $f_{i/2} = f((x_i + x_{i-1})/2)$ $(i = 1, \ldots, n)$ jelölésekkel. Az összetett érintőformula, hasonlóan az összetett trapézformulához, az érintőformulát alkalmazza az ekvidisztánsan felosztott $[a,b]$ intervallum részintervallumain. Az integrál pontos értékét tehát a formula az

$$I_{n,\text{érintő}}(f) = h(f_{1/2} + \ldots + f_{n-1/2})$$

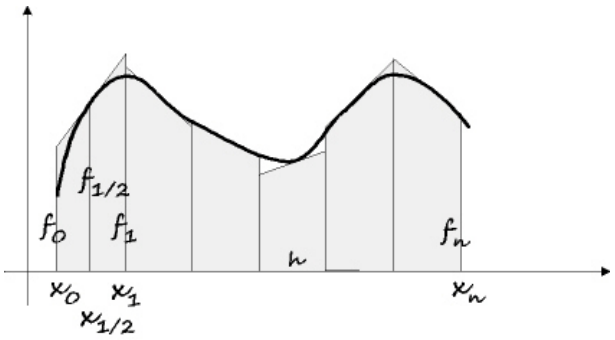
módon közelíti (8.3.2. ábra).

Az összetett érintőformula egy nyílt kvadratúraformula. Most is könnyen látható, hogy az általa adott közelítés integrálható függvények esetén tart a pontos integrálértékhez, ha az osztóintervallumok n száma végtelenhez tart. A formula pontossági rendje nyilvánvalóan 2. A konvergenciarendje is 2. Erről szól a következő tétel.

8.3.5. tétel.
Az összetett érintőformula hibája $f \in C^2[a,b]$ függvények esetén

$$I_{n,\text{érintő}}(f) - I(f) = -\frac{(b-a)h^2}{24}f''(\eta),$$

ahol $\eta \in [a,b]$ egy megfelelően választott konstans.



8.3.2. ábra: Az összetett érintőformula a besötétített tartomány területével közelíti az integrál pontos értékét.

Bizonyítás. A bizonyítást nem az interpolációs hibafüggvény integráljának becslésével hajtjuk végre, mint az összetett trapézformulára vonatkozó hasonló tétel esetén, mert azzal a módszerrel csak elsőrendű konvergencia igazolható. Ehelyett az integrálandó függvényt másodrendű Taylor-polinomjának és a hozzá tartozó Lagrange-féle maradéktagnak az összegeként állítjuk elő, majd kiszámítjuk az érintőformula által adott közelítés és a pontos integrál értékének különbségét az $[x_{k-1}, x_k]$ intervallumon.

$$f_{k-1/2}h - \int_{x_{k-1}}^{x_k} f(x)\,\mathrm{d}x$$

$$= f_{k-1/2}h - \int_{x_{k-1}}^{x_k} (f_{k-1/2} + f'_{k-1/2}(x - x_{k-1/2}) + f''(\xi_{k,x})(x - x_{k-1/2})^2/2)\,\mathrm{d}x$$

$$= -\frac{f''(\eta_k)h^3}{2 \cdot 12},$$

ahol $f'_{k-1/2} = f'((x_k + x_{k-1})/2)$ és $\xi_{k,x}$ megfelelő k-tól és x-től függő érték az (x_{k-1}, x_k) nyílt intervallumból $(k = 1, \ldots, n)$. Az utolsó lépésben azt a tényt alkalmaztuk, hogy az érintőformula képlete és a Taylor-polinom első két tagja integráljának különbsége nulla. Az η_k konstans értékét a 8.3.1. tétel szerint választjuk.

Mivel n intervallum van, ezért a teljes hiba valamilyen $\eta \in [a, b]$ értékkel

$$I_{n,\text{érintő}}(f) - I(f) = -\sum_{k=1}^{n} \frac{f''(\eta_k)h^3}{24} = -n\frac{f''(\eta)h^3}{24} = -\frac{(b-a)h^2}{24} f''(\eta).$$

Ezt akartuk megmutatni. ■

8.3.6. megjegyzés. Az előző tételben az η paraméter értékét általában nem ismerjük. Ekkor csak becslést tudunk adni az integrál hibájára az

$$|I_{n,\text{érintő}}(f) - I(f)| \leq \frac{(b-a)h^2 M_2}{24}$$

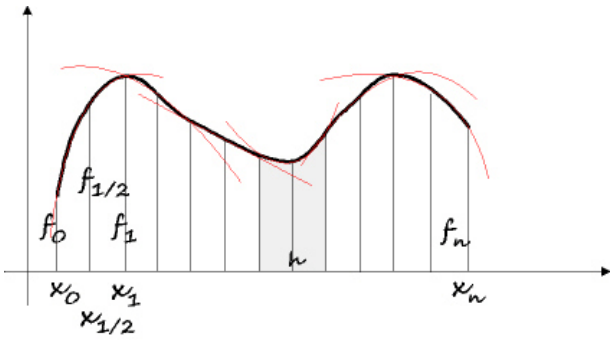
alakban. ◇

8.3.3. Összetett Simpson-formula

Az előző fejezetekben használt jelölések segítségével az összetett Simpson-formula az

$$I_{n,\mathrm{Simp}}(f) = \frac{h}{6}(f_0 + 4f_{1/2} + 2f_1 + 4f_{3/2} + 2f_2 + \ldots + 4f_{n-1/2} + f_n)$$

alakban adható meg. Ez a részintervallumok végpontjaiban és felezőpontjaiban lévő három függvényértékre illesztett legfeljebb másodfokú polinom integráljával közelíti a tényleges integrált. Azt, hogy ez a formula mely tartomány területével közelíti a tényleges integrálértéket, a 8.3.3. ábrán szemléltettük.



8.3.3. ábra: Az összetett Simpson-formula egy részintervallumon a besötétített tartomány területével közelíti az integrál pontos értékét. Az ábrán berajzoltuk az egyes részintervallumon interpoláló legfeljebb másodfokú polinomokat is.

A formula zárt kvadratúraformula, és értéke az integrál pontos értékéhez tart, ha az n felosztásszámmal végtelenhez tartunk. A formula pontossági rendje nyilván 4, hiszen a Simpson-formuláé is 4. A következő tétel mutatja, hogy a konvergenciarendje is 4.

8.3.7. tétel.
Az összetett Simpson-formula hibája $f \in C^4[a,b]$ függvények esetén

$$I_{n,\mathrm{Simp}}(f) - I(f) = \frac{(b-a)h^4}{2880} f^{(4)}(\eta),$$

ahol $\eta \in [a,b]$ megfelelően választott konstans.

Bizonyítás. Tekintsük az $[x_{k-1}, x_k]$ intervallumon az f függvényt, és fejtsük sorba a harmadfokú tagig

$$f(x) = f_{k-1/2} + f'_{k-1/2}(x - x_{k-1/2}) + \frac{f''_{k-1/2}(x - x_{k-1/2})^2}{2}$$
$$+ \frac{f'''_{k-1/2}(x - x_{k-1/2})^3}{6} + \frac{f^{(4)}(\xi_{k,x})(x - x_{k-1/2})^4}{24}.$$

Ekkor a k-adik intervallumon a számított integrál hibája

$$\frac{h}{6}(f_{k-1} + 4f_{k-1/2} + f_k) - \int_{x_{k-1}}^{x_k} f(x)\,\mathrm{d}x = \frac{h}{6}(f_{k-1} + 4f_{k-1/2} + f_k) - f_{k-1/2}h$$

$$-\left(\underbrace{\frac{f_{k-1}-2f_{k-1/2}+f_k}{(h/2)^2}-\frac{f^{(4)}(\eta_{k1})(h/2)^2}{12}}_{f''_{k-1/2}}\right)\frac{h^3}{2\cdot 12}-\frac{f^{(4)}(\eta_{k2})h^5}{24\cdot 80}$$

$$=\frac{f^{(4)}(\eta_{k1})h^5}{48\cdot 2\cdot 12}-\frac{f^{(4)}(\eta_{k2})h^5}{24\cdot 80}=\frac{h^5 f^{(4)}(\eta_k)}{2880},$$

ahol η_{k1}, η_{k2}, η_k $(k=1,\dots,n)$ megfelelő konstansok az $[x_{k-1},x_k]$ intervallumból $(k=1,\dots,n)$. A teljes hibára így

$$I_{n,Simp(f)}-I(f)=\sum_{k=1}^{n}\frac{h^5 f^{(4)}(\eta_k)}{2880}=n\frac{h^5}{2880}f^{(4)}(\eta)=\frac{(b-a)h^4}{2880}f^{(4)}(\eta)$$

adódik, ahol η megfelelő konstans az $[a,b]$ intervallumból. ■

8.3.8. megjegyzés. Az összetett Simpson-formula hibájában szereplő η értéke általában nem ismert. Ekkor a közelítés hibájára az

$$|I_{n,\mathrm{Simp}}(f)-I(f)|\le\frac{(b-a)h^4M_4}{2880}$$

becslést használhatjuk. ◇

8.3.9. megjegyzés. Az összetett trapéz- és az összetett érintőformulákat 1:2 arányban súlyozva az összetett Simpson-formulához jutunk. Ez következik a nem összetett formulákra megfogalmazott hasonló állításból. ◇

8.3.10. megjegyzés. Vegyük észre, hogy a korábban megismert hibabecslések alkalmasak arra, hogy az integrálandó függvény deriváltjára vonatkozó felső korlátot ismerve megmondjuk, hogy az adott intervallumot hány részre kell felosztanunk ahhoz, hogy a közelítés egy előre adott hibánál jobban megközelítse az integrál pontos értékét. ◇

8.3.11. megjegyzés. Bár az összetett formulákat csak ekvidisztáns felosztás esetén vizsgáltuk, ezek könnyen átfogalmazhatók arra az esetre is, ha a részintervallumok hossza nem egyforma. ◇

8.3.12. példa. Egyszerű példaként határozzuk meg az

$$\int_0^1 x^4\,\mathrm{d}x=\frac{1}{5}$$

integrál értékét a tanult összetett formulák segítségével 4 osztóintervallumot használva! Most tehát $f(x)=x^4$ és a négy osztóintervallum miatt $h=0.25$.

Az összetett trapézformula az

$$I_{4,\mathrm{trap}}(f)=h(f(0)+2(f(0.25)+f(0.5)+f(0.75))+f(1))=0.220703125$$

értéket adja. Az érintőformulával nyert közelítés

$$I_{4,\text{érintő}}(f) = h(f(0.125) + f(0.375) + f(0.625) + f(0.875)) = 0.189697265625.$$

A Simpson-formula pedig a fenti két képlet súlyozásával adódik

$$I_{4,\text{Simp}}(f) = \frac{2I_{4,\text{érintő}}(f) + I_{4,\text{trap}}(f)}{3} = 0.20003255208333.$$

Ez utóbbi jóval pontosabb integrálközelítést ad, mint a másik kettő formula. ⋄

8.4. Romberg-módszer

Az előző fejezet végén láttunk példát arra (8.3.9. megjegyzés), hogy két másodrendű kvadratúraképlet eredményét megfelelően súlyozva negyedrendű kvadratúraképlet állítható elő. Azt az eljárást, ahogy súlyozással alacsonyabbrendű kvadratúraképletekből magasabbrendű kvadratúraképleteket állítunk elő, Romberg[4]-módszernek nevezzük. A Romberg-módszer tulajdonképpen a Richardson-extrapoláció numerikus integrálási formulákra alkalmazva. Ebben a fejezetben ezen módszer lényegét ismertetjük röviden.

Legyen $I_n(f)$ egy adott kvadratúraképlet egy n egyenlő részre osztott $[a, b]$ intervallumon. Tegyük fel, hogy a kvadratúraformula konvergenciarendje r. Tegyük fel továbbá, hogy a képlet hibája jó közelítéssel az

$$I_n(f) - I(f) = Ch^r$$

alakban írható, alkalmas C konstassal. Ekkor ha megkétszerezzük az osztóintervallumok számát, akkor az új hibára körülbelül

$$I_{2n}(f) - I(f) = C\left(\frac{h}{2}\right)^r$$

adódna. A fenti két képletből $I(f)$ kifejezhető

$$I(f) = I_{2n}(f) + \frac{I(f) - I_n(f)}{2^r} = \frac{I_{2n}(f)2^r - I_n(f)}{2^r - 1}$$

alakban, ami természetesen csak egy újabb közelítése lesz a tényleges integrálértéknek. Igazolható azonban, hogy az így kapott értékek gyorsabban konvergálnak a pontos integrálértékhez, mint az eredeti kvadratúrával kapott értékek. Megmutatható, hogy az így kapott számsorozatok rendje $r+2$ lesz, azaz a rend kettővel növekszik a súlyozás által. Így tehát ahelyett, hogy tovább dupláznánk az osztóintervallumok számát (ami újabb függvényértékek kiszámolását igényli), egyszerű súlyozással pontosabb közelítés adható az integrál értékére.

8.4.1. példa. Alkalmazzuk a Romberg-módszert az

$$\int_0^1 e^{-x^2}\,\mathrm{d}x = 0.746824132812$$

[4] Werner Romberg (1909-2003), német matematikus.

integrál kiszámítására! A számított és a súlyozott értékeket az alábbi táblázatban tüntettük fel.

n	2	4	8	16	32
össz. trapéz	0.7313702518	0.7429840978	0.7458656148	0.7465845967	0.7467642546
össz. Simpson		0.7468553797	0.7468261205	0.7468242574	0.7468241406
			0.7468241699	0.7468241332	0.7468241328
				0.7468241326	0.7468241328
					0.7468241328

A felső sorban az osztóintervallumok száma szerepel, a második sor az összetett trapézformulával (másodrendű, tehát $r = 2$) számított értékeket tartalmazza az adott osztóintervallumszám esetén, az alatta lévő sorok a Romberg-módszer segítségével lettek számítva. A második sor éppen a Simpson-formulát adja. Látható, hogy az alsó sorokban szereplő számok sokkal pontosabbak, mint az összetett trapézmódszer által adott közelítések. ◇

8.5. Gauss-kvadratúra

Egy határozott integrál értékét közelíthetjük a korábban megismert összetett kvadratúraformulákkal. Ha módunk van több helyen kiszámítani a függvényértékeket, akkor az osztóintervallumok számának növelésével tetszőleges pontossággal megközelíthetjük az integrál értékét. Ha erre nincs lehetőség, mert pl. mérésekből csak jól meghatározott helyeken ismerjük a függvényértéket, akkor mondhatunk egy közelítést az integrálra, és feltéve az integrálandó függvény megfelelő simaságát, a hibabecslő formulákból mondhatunk egy felső korlátot a hiba nagyságára. Ha tetszőleges számú alappontban ki tudjuk számítani a függvényértékeket, akkor felmerül a kérdés, hogy mik legyenek ezek az alappontok. Az interpolációs feladatoknál láttuk, hogy az alappontok alkalmas megválasztásával az interpolációs hiba jelentősen csökkenthető. Most megvizsgáljuk, hogy az interpolációs kvadratúraképletek rendje növelhető-e az alappontok alkalmas megválasztásával.

Ebben a fejezetben a korábban vizsgált határozott integrálokat úgy módosítjuk, hogy az integrálandó f függvényt még beszorozzuk egy $[a, b]$-n folytonos és pozitív s súlyfüggvénnyel. Természetesen az $s(x) \equiv 1$ választással visszakapjuk a korábban vizsgált integrálok alakját. Célunk tehát az

$$I(f; s) := \int_a^b s(x) f(x)\, \mathrm{d}x$$

integrál minél pontosabb közelítése.

Ha a fenti integrál közelítéséhez interpolációs kvadratúraformulát használunk az

$$(x_0, f_0), \ldots, (x_n, f_n)$$

pontokon, akkor a kvadratúraformula

$$I_n(f; s) = \sum_{k=0}^{n} a_k f_k$$

alakú lesz, ahol a súlyok

$$a_k = \int_a^b s(x) l_k(x)\, \mathrm{d}x, \;\; k = 1, \ldots, n$$

alakban írhatók. Ez a kvadratúraformula pontos lesz minden legfeljebb n-edfokú polinomra (8.2.6. tétel). Hogyan válasszuk az alappontokat, hogy magasabbfokú polinomokra is pontos legyen a képlet? A következő tétel arról szól, hogy mennyivel növelhető meg a formula rendje.

8.5.1. tétel.
Az

$$I_n(f;s) = \sum_{k=0}^{n} a_k f_k$$

interpolációs kvadratúraformula pontosan akkor pontos minden P_{n+m}-beli polinomra, ha

$$\int_a^b w_{n+1}(x)s(x)p(x)\,\mathrm{d}x = 0 \tag{8.5.1}$$

minden $p \in P_{m-1}$ esetén. A w_{n+1} az alappontpolinom szokásos jelölése.

Bizonyítás. Igazoljuk először a feltétel elégségességét. Legyen $f \in P_{n+m}$, és azt kell megmutatnunk, hogy a kvadratúraformula pontos az f polinomra. Az f polinom a polinomokra ismert maradékos osztás segítségével felírható

$$f(x) = w_{n+1}(x)q(x) + r(x)$$

alakban, ahol $r \in P_n$ és $q \in P_{m-1}$. Ekkor

$$\sum_{k=0}^{n} a_k r(x_k) = \int_a^b r(x)s(x)\,\mathrm{d}x = \int_a^b f(x)s(x)\,\mathrm{d}x - \underbrace{\int_a^b w_{n+1}(x)s(x)q(x)\,\mathrm{d}x}_{=0}.$$

Tehát

$$\int_a^b s(x)f(x)\,\mathrm{d}x = \sum_{k=0}^{n} a_k r(x_k) = \sum_{k=0}^{n} a_k f(x_k),$$

azaz a kvadratúraformula pontos f-re.

A másik irány igazolásához legyen $p \in P_{m-1}$. Ekkor a formula pontos a $p \cdot w_{n+1}$ polinomra, hiszen legfeljebb $n+m$ fokú, azaz

$$\int_a^b s(x)p(x)w_{n+1}(x)\,\mathrm{d}x = \sum_{k=0}^{n} a_k p(x_k)w_{n+1}(x_k) = 0.$$

Ezt akartuk megmutatni. ■

Megvizsgáljuk, hogy mekkora lehet a tételben szereplő m maximális értéke, azaz mennyivel növelhető meg a pontossági rend. Látható, hogy a kvadratúraformula nem lehet pontos minden P_{2n+2}-beli polinomra, mert akkor $p = w_{n+1}$ esetén az

$$\int_a^b s(x)w_{n+1}^2(x)\,\mathrm{d}x = 0$$

egyenlőségből a $w_{n+1} \equiv 0$ azonosság következne, ami ellentmondás. Tehát m maximálisan $n+1$ lehet. Most azt mutatjuk meg, hogy el is érhető, hogy a formula minden P_{2n+1}-beli polinomra pontos legyen. Tehát a Gauss-kvadratúra pontossági rendje így $2n+2$ lesz. Azt kell megmutatnunk, hogy a (8.5.1) feltétel teljesíthető minden P_{2n+1}-beli polinomra, ha alkalmasan választjuk

az alappontokat. Tekintsük az $[a,b]$ intervallumon az s súlyfüggvényre nézve s-ortogonális polinomokat (súlyfüggvénytől függően Csebisev-, Legendre- stb. polinomok) és legyenek $x_0,\dots,x_n$ az $n+1$-edfokú s-ortogonális polinom zérushelyei. Ekkor ezen alappontokon a w_{n+1} polinom s-ortogonális lesz minden nála kisebb fokszámú polinomra, azaz minden n-edfokú polinomra. Így (8.5.1) teljesül minden legfeljebb n-edfokú p polinomra. Tehát a formula pontos minden P_{2n+1}-beli polinomra. Attól függően, hogy milyen ortogonális polinomokat használunk, az így nyert kvadratúraképletet Gauss–Csebisev- vagy Gauss–Legendre-kvadratúrának nevezzük.

A Gauss–Csebisev és Gauss–Legendre kvadratúraformulák által használt alappontokat és a súlyokat az alábbi táblázatban gyűjtöttük össze.

név	$s(x)$	p_1,p_2,p_3 zérushelyei	súlyok
Csebisev	$1/\sqrt{1-x^2}$	$0;\ \pm1/\sqrt{2}\ ;\ 0,\pm\sqrt{3}/2$	$\pi;\pi/2,\pi/2;\pi/3,\pi/3,\pi/3$
Legendre	1	$0;\ \pm1/\sqrt{3}\ ;\ 0,\pm\sqrt{3/5}$	$2;1,1;5/9,8/9,5/9$

8.5.1. táblázat: A Gauss–Csebisev- és Gauss–Legendre-kvadratúrák alappontjai és súlyai.

8.5.2. megjegyzés. Bár a vizsgált ortogonális polinomok a $[-1,1]$ intervallumon ortogonálisak, egy egyszerű koordinátatranszformációval tetszőleges $[a,b]$ intervallumra áttranszformálhatók. ⋄

8.5.3. példa. Készítsük el a három pontra illeszkedő Csebisev–Gauss formulát! A harmadfokú Csebisev-polinom zérushelyei 0 és $\pm\sqrt{3}/2$. Ezek lesznek az alappontok. A súlyok, mivel a súlyfüggvény ebben az esetben az $1/\sqrt{1-x^2}$ függvény,

$$a_0=\int_{-1}^{1}\frac{x(x-\sqrt{3}/2)}{-\sqrt{3}/2(-\sqrt{3}/2-\sqrt{3}/2)}\,\frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x=\pi/3,$$

hasonlóan $a_1=a_2=\pi/3$. Így a formula

$$\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}\,\mathrm{d}x\approx\frac{\pi}{3}(f(-\sqrt{3}/2)+f(0)+f(\sqrt{3}/2)),$$

amely pontos lesz minden legalább ötödfokú polinomra. ⋄

A Gauss-féle kvadratúraformula hibájáról szóló tételt bizonyítás nélkül közöljük.

8.5.4. tétel.
A Gauss-féle kvadratúraformula hibája $n+1$ alappont esetén

$$I_n(f;s)-I(f;s)=-\frac{f^{(2n+2)}(\eta)}{(2n+2)!}\int_a^b s(x)(w_{n+1}(x))^2\,\mathrm{d}x,$$

ahol η (a,b)-be eső megfelelő konstans.

8.6. Numerikus integrálási eljárások a MATLAB-ban

A MATLAB-ban mindegyik tanult numerikus integrálási eljárás nagyon egyszerűen programozható. Az alábbi függvény az összetett Simpson-formula esetét mutatjuk be, de a másik kettő függvény is hasonlóan egyszerű. A programban értelemszerűen `a` és `b` adja meg az integrálási intervallum két végpontját, `n` az osztóintervallumok száma és `fv` az integrálandó függvény.

```
function int = osszsimpson(a,b,n,fv)
h=(b-a)/n;
x=[a:h/2:b];
y=eval(fv);
int=(h/6)*(y(1)+2*sum(y(3:2:2*n-1))+4*sum(y(2:2:2*n))+y(2*n+1));
```

A program futtatható pl. az

$$\int_0^1 x\sin(x)\,\mathrm{d}x$$

integrál esetén az alábbi módon.

```
>> osszsimpson(0,1,20,'x.*sin(x)') % 20 osztóintervallumot alkalmazunk.

ans =

   0.30116867228813
```

A MATLAB önmagában is kínál néhány numerikus integrálási eljárást. Ezek közül mutatunk be kettőt. Az első a `trapz` parancs, ami egy egyszerű függvény az összetett trapézformulára.

```
>> x=0:1/100:1; y=x.*sin(x); trapz(x,y)
% Az x vektor tartalmazza az alappontokat,
% az y vektor pedig a függvényértékeket.

ans =

   0.30118019375974
```

A másik a `quad` parancs, amely az összetett Simpson-formulát alkalmazza rekurzív módon addig, míg a hiba 10^{-6}-nál kisebb nem lesz (lásd 8.7.6. feladat).

```
>> quad('x.*sin(x)',0,1)

ans =

   0.30116867962220
```

8.7. Feladatok

Numerikus integrálás

8.7.1. feladat. Alkalmazzuk az összetett trapézformulát az

$$\int_0^2 \frac{1}{x+2}\,\mathrm{d}x$$

integrál kiszámítására úgy, hogy a hiba 10^{-3}-nál kisebb legyen!

8.7.2. feladat. Határozzuk meg az előző feladatbeli integrál pontos értékét, és kísérletileg vizsgáljuk meg az egyes módszerek konvergenciarendjét számítógép segítségével!

8.7.3. feladat. Határozzuk meg az $N^{4,k}_{\text{zárt}}$ Newton–Cotes-együtthatókat! Használjuk a MATLAB `int` parancsát a polinomok integráláshoz!

8.7.4. feladat. Határozzuk meg az e^{-x^2} függvény közelítő integrálját a $[0,1]$ intervallumon az előző feladatban kiszámolt együtthatók segítségével!

8.7.5. feladat. Határozzuk meg a $\sin x$ függvény közelítő integrálját a $[0,\pi/2]$ intervallumon a Romberg-módszer alkalmazásával. Határozzuk meg hasonlóan az $\int_0^{0.8} e^{-x^2}\,\mathrm{d}x$ integrált. Mindkét esetben legyen az elérni kívánt pontosság 10^{-6}. Addig számoljunk, míg két egymás utáni közelítés már közelebb van egymáshoz, mint 10^{-6}!

8.7.6. feladat. Egy közelítő integrált határoztunk meg a Simpson-szabállyal ($I_n(f)$), majd kétszer annyi osztóintervallummmal újra kiszámoltunk egy közelítést ($I_{2n}(f)$). Adjunk becslést a fenti értékek segítségével a durvább közelítés hibájára!

8.7.7. feladat. Közelítsük az $\int_0^h f(x)\,\mathrm{d}x$ integrált a következő kvadratúraképlettel:

$$h(a_0 f_0 + a_1 f_1) + h^2(b_0 f_0' + b_1 f_1'),$$

ahol f_i, f_i' a 0 ill. h pontokbeli függvényértékek ill. deriváltak! Hogyan válasszuk meg az a_0, a_1, b_0 és b_1 együtthatókat, hogy a formula minden legfeljebb negyedfokú polinomra pontos legyen?

8.7.8. feladat. Tekintsük az $\int_0^2 \sqrt{x} f(x)\,\mathrm{d}x$ integrál közelítésére az $I_2(f) = a_0 f_0 + a_1 f_1 + a_2 f_2$ kvadratúraképletet. Határozzuk meg úgy az együtthatókat, hogy a képlet minden legfeljebb másodfokú polinomra pontos legyen!

8.7.9. feladat. Készítsük el a három pontra illeszkedő Gauss–Legendre-formulát! Határozzuk meg ezzel a módszerrel az $\int_{-1}^1 1/(1+x^2)\,\mathrm{d}x$ integrál közelítését! Adjunk becslést előre a hibára!

8.7.10. feladat. Készítsük el a két pontra illeszkedő Gauss–Csebisev-formulát! Határozzuk meg ezzel a módszerrel az $\int_{-1}^1 1/\sqrt{1-x^2}\,\mathrm{d}x$, $\int_{-1}^1 x^3/\sqrt{1-x^2}\,\mathrm{d}x$ és $\int_{-1}^1 x^4/\sqrt{1-x^2}\,\mathrm{d}x$ integrálok "közelítő" értékét! Adjunk becslést előre a hibára!

Ellenőrző kérdések

1. Miért van szükség kvadratúraformulákra?
2. Milyen elven alapulnak az interpolációs kvadratúraformulák?
3. Mennyiben speciális kvadratúraformulák a Newton–Cotes-formulák?
4. Mik azok a Newton–Cotes-együtthatók?
5. Melyik a három nevezetes Newton–Cotes-formula?
6. Adjuk meg a három nevezetes összetett kvadratúraformula képletét és konvergenciarendjét!
7. Ismertessük a Romberg-eljárást!
8. Mekkora pontossági rend érhető el a Gauss-kvadratúra segítségével?
9. Adjuk meg a Gauss-kvadratúra kiszámításának módját!

9. A közönséges differenciálegyenletek kezdetiérték-feladatainak numerikus módszerei

Ebben a fejezetben a közönséges differenciálegyenletek kezdetiérték-feladatainak elméleti összefoglalása után azok numerikus megoldási módszereivel foglalkozunk. Ismertetjük a legtipikusabb egylépéses módszereket, és ezek általánosításaként a Runge-Kutta típusú egylépéses illetve a lineáris többlépéses módszereket tárgyaljuk. Ezután megvizsgáljuk a speciális tulajdonsággal rendelkező ún. merev rendszerek tulajdonságait. A fejezet végén az ismertetett módszerek számítógépes realizálásával, és Matlab programjaikkal foglalkozunk.

9.1. Bevezetés

A differenciálegyenletek gyakori eszközei a természettudományos, műszaki, közgazdasági folyamatok leírásának, azaz a folytonos matematikai modelleket többnyire ezek segítségével lehetséges (és szokásos) leírni. Az ilyen modellek elemzésével a közönséges (illetve parciális) differenciálegyenletek elmélete foglalkozik. Ezek a vizsgálatok elsősorban a különböző jellegű feladatok megoldhatóságával foglalkoznak, tehát elsősorban azt vizsgáljuk, hogy a kitűzött feladat milyen feltételek mellett korrekt kitűzésű. A megoldás konkrét előállítása zárt alakban (azaz megadása olyan képletek segítségével, amelyek ismert és könnyen kiértékelhető függvényeket tartalmaznak) csak ritkán lehetséges. Ezért gyakorlati szempontból megkerülhetetlen annak a vizsgálata, amikor a megoldást valamilyen *numerikus módszer* segítségével *közelítő alakban* keressük. Mint látni fogjuk, ezek a módszerek lehetővé teszik a numerikus megoldás nagy pontosságú és megbízható előállítását. Ez utóbbi azt jelenti, hogy becslést tudunk adni az eredeti feladat elemi eszközökkel nem meghatározható pontos megoldása és az alkalmazott közelítő módszerrel nyert numerikus megoldás közötti eltérésre.

9.2. A közönséges differenciálegyenletek kezdetiérték-feladata

9.2.1. definíció.

Legyen $G \subset \mathbb{R} \times \mathbb{R}^d$ egy tartomány (azaz összefüggő, nyílt halmaz), $(t_0, \overline{\mathbf{u}}_0) \in G$ egy adott pont ($t_0 \in \mathbb{R}$, $\overline{\mathbf{u}}_0 \in \mathbb{R}^d$), $\mathbf{f} : G \to \mathbb{R}^d$ egy folytonos leképezés. A

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{f}(t, \mathbf{u}), \quad \mathbf{u}(t_0) = \overline{\mathbf{u}}_0 \tag{9.2.1}$$

feladatot *kezdetiérték-feladatnak*, avagy más szóval *Cauchy-feladatnak* nevezzük.

A könnyebb áttekinthetőség kedvéért írjuk ki a (9.2.1) feladatot koordinátánként! Jelölje $u_i(\cdot)$ az ismeretlen $\mathbf{u}(t)$ vektorértékű függvény i-edik koordináta-függvényét, $f_i : G \to \mathbb{R}$ az $\mathbf{f}$ és u_{0i}

$(i = 1, 2, \dots, d)$ pedig az $\overline{\mathbf{u}}_0$ vektor koordinátáit. Ekkor a Cauchy-feladat felírható a következő ún. koordinátánkénti alakban:

$$\begin{aligned} \frac{du_i(t)}{dt} &= f_i(t, u_1(t), u_2(t), \dots u_d(t)), \\ u_i(t_0) &= u_{0i} \end{aligned} \tag{9.2.2}$$

ahol $i = 1, 2, \dots, d$.

Egy Cauchy-feladat megoldása azt jelenti, hogy meghatározzuk az összes olyan $\mathbf{u} : \mathbb{R} \to \mathbb{R}^d$ függvényt, amely valamely $I \subset \mathbb{R}$ intervallum pontjaiban egyrészt behelyettesíthető a (9.2.1) feladatba, másrészt pedig azt ki is elégíti.

9.2.2. definíció.

Az olyan $\mathbf{u} : I \to \mathbb{R}^d$ (I egy nyílt intervallum) folytonosan differenciálható függvényt, amelyre

- $\{(t, \mathbf{u}(t)) : \quad t \in I\} \subset G$;
- $\dfrac{d\mathbf{u}(t)}{dt} = \mathbf{f}(t, \mathbf{u}(t)), \quad \text{minden} \quad t \in I,$
- $t_0 \in I$ és $\mathbf{u}(t_0) = \overline{\mathbf{u}}_0$

a (9.2.1) Cauchy-feladat megoldásának nevezzük.

Amikor a (9.2.1) Cauchy-feladat egy természettudományos, műszaki vagy közgazdasági folyamat matematikai modellje, akkor alapvető követelmény, hogy létezzen egyértelmű megoldása. Ennek biztosítására vezessük be valamely α és β pozitív számok mellett a $H_{\alpha,\beta}(t_0, \overline{\mathbf{u}}_0) = \{(t, \mathbf{u}) : |t - t_0| \le \alpha, \|\mathbf{u} - \overline{\mathbf{u}}_0\|_\infty \le \beta\} \subset G$ jelölést. (Tehát $H_{\alpha,\beta}(t_0, \overline{\mathbf{u}}_0)$ egy $(t_0, \overline{\mathbf{u}}_0)$ közepű, zárt, $d+1$-dimenziós téglalap.) Mivel $\mathbf{f}$ folytonos a zárt $H_{\alpha,\beta}(t_0, \overline{\mathbf{u}}_0)$ halmazon, ezért értelmes az $M = \max_{H_{\alpha,\beta}(t_0,\overline{\mathbf{u}}_0)} \|\mathbf{f}(t, \mathbf{u})\|_\infty$ valós szám bevezetése. Ekkor minden t, $|t - t_0| \le \min\{\alpha, \beta/M\}$ esetén létezik a (9.2.1) Cauchy-feladatnak $\mathbf{u}(t)$ megoldása. Ha emellett a $H_{\alpha,\beta}(t_0, \overline{\mathbf{u}}_0)$ halmazon az $\mathbf{f}$ függvény *a második változójában lipschitzes*, azaz valamely $L > 0$ állandó mellett minden $(t, \mathbf{u}_1), (t, \mathbf{u}_2) \in H_{\alpha,\beta}(t_0, \overline{\mathbf{u}}_0)$ pontban teljesül a

$$\|\mathbf{f}(t, \overline{\mathbf{u}}_1) - \mathbf{f}(t, \overline{\mathbf{u}}_2)\|_\infty \le L\|\overline{\mathbf{u}}_1 - \overline{\mathbf{u}}_2\|_\infty \tag{9.2.3}$$

ún. *Lipschitz-féle feltétel*, akkor ez a megoldás egyértelmű is.

A továbbiakban a (9.2.1) feladatra mindig feltesszük, hogy létezik olyan alkalmasan megválasztott $H_{\alpha,\beta}(t_0, \overline{\mathbf{u}}_0) \subset G$ részhalmaz, amelyen $\mathbf{f}$ folytonos és a második változójában lipschitzes, azaz létezik egyértelmű megoldása az $I_0 := \{t \in I : |t - t_0| \le T\}$ intervallumon, ahol $T = \min\{\alpha, \beta/M\}$, és a közelítő megoldást az I_0 intervallumon állítjuk elő.

Mivel a t változó az időt jelöli, ezért egy Cauchy-feladat megoldása azt írja le, hogy a rendszer időben hogyan változik. Mi a gyakorlati problémák vizsgálata során általában erre az időbeli fejlődésre (változásra) vagyunk kíváncsiak. Ez azt jelenti, hogy ismerve a rendszer állapotát egy rögzített kezdeti időpontban, annak ezen időpontot követő állapotát szeretnénk meghatározni, azaz $\mathbf{u}(t_0)$ ismeretében az $\mathbf{u}(t)$ függvényt a $t > t_0$ értékeire vagyunk kíváncsiak. A $t = t_0$ időpontot *kezdőpontnak*, a megoldásfüggvény ezen pontbeli értékét *kezdeti értéknek*, a $(t_0, \mathbf{u}(t_0))$ párt pedig *kezdeti feltételnek* nevezzük. Nyilvánvalóan nem jelent megszorítást, ha a kezdőpontot $t_0 = 0$ értéknek vesszük. Így a (9.2.1) feladat megoldásának értelmezési tartománya a $[0, T] \subset I$

intervallum, és ekkor feladatunk a következő alakot ölti:

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{f}(t, \mathbf{u}(t)), \quad t \in [0, T], \tag{9.2.4}$$

$$\mathbf{u}(0) = \overline{\mathbf{u}}_0. \tag{9.2.5}$$

Célunk a továbbiakban ezen $\mathbf{u}(t)$ függvény meghatározása.

9.2.3. megjegyzés. Az $\mathbf{f}$ függvény folytonossága esetén (azaz $\mathbf{f} \in C(H)$ mellett) a Cauchy-feladat $\mathbf{u}(t)$ megoldása egyszeresen folytonosan differenciálható is, tehát $\mathbf{u} \in C^1[0, T]$. Ugyanakkor, ha $\mathbf{f}$ magasabb rendben sima, akkor a megoldás is simábbá válik: ha $\mathbf{f} \in C^p(H)$, akkor $\mathbf{u} \in C^{p+1}[0, T]$, ahol $p \in \mathbb{N}$. Így az adott $\mathbf{f}$ függvény megfelelő simaságával a megoldás szükséges simasága mindig biztosítható. Ezért tehát nem jelent lényeges megszorítást, ha a továbbiakban – ahol ez szükséges – feltesszük, hogy a megoldás *megfelelően sima*. ◇

A numerikus módszereket – a könnyebb áttekinthetőség kedvéért – a skaláris egyenletekre fogalmazzuk meg, azaz a $d = 1$ esetet tekintjük. Legyen $Q_T := [0, T] \times \mathbb{R} \subset \mathbb{R}^2$, $f : Q_T \to \mathbb{R}$. A továbbiakban a

$$\frac{du}{dt} = f(t, u), \quad u(0) = u_0 \tag{9.2.6}$$

feladatot nevezzük Cauchy-feladatnak, ahol mindvégig feltesszük, hogy $f \in C(Q_T)$ és a második változójában lipschitzes függvény, azaz

$$|f(t, u_1) - f(t, u_2)| \leq L\,|u_1 - u_2|\,, \quad \forall (t, u_1), (t, u_2) \in Q_T, \tag{9.2.7}$$

továbbá $u_0 \in \mathbb{R}$ adott szám. Tehát feladatunk a következő: keressük azon megfelelően sima $u : [0, T] \to \mathbb{R}$ függvényt, amelyre

$$\frac{du(t)}{dt} = f(t, u(t)), \quad \forall t \in [0, T], \quad u(0) = u_0. \tag{9.2.8}$$

9.2.4. megjegyzés. Felmerülhet a kérdés: van-e kapcsolat valamely $g : \mathbb{R}^2 \to \mathbb{R}$ függvény folytonossága és a második változójában való lipschitzessége között? A válasz nemleges, ugyanis, mint azt a következő két példa is mutatja, ezek egymástól független feltételek. Legyen először $g(x, y) = y^2$. Ez a függvény nyilván folytonos a $G = \mathbb{R}^2$ síkon, de nem lipschitzes, ugyanis

$$|g(x, y_1) - g(x, y_2)| = \left|y_1^2 - y_2^2\right| = |y_1 + y_2|\,|y_1 - y_2|\,,$$

és így a (9.2.7) feltétel nem teljesülhet, hiszen y_1 és y_2 tetszőlegessége miatt $|y_1 + y_2|$ nem lehet felülről korlátos valamely L állandóval.

Legyen most $g(x, y) = D(x)y$, ahol $D(x)$ a jól ismert Dirichlet-függvény[1]. Ekkor g sehol sem folytonos, viszont

$$|g(x, y_1) - g(x, y_2)| = |D(x)|\,|y_1 - y_2| \leq |y_1 - y_2|\,,$$

azaz $L = 1$ értékkel a (9.2.7) összefüggés érvényes a $G = \mathbb{R}^2$ síkon. ◇

9.2.5. megjegyzés. Hogyan biztosítható a lipschitzesség? Tegyük fel, hogy valamely $g : \mathbb{R}^2 \to \mathbb{R}$ függvény az értelmezési tartományának valamely nyílt H_g részhalmazán a második változójában korlátos deriválttal rendelkezik. Ekkor a Lagrange-középértéktétel értelmében valamely

[1] A Dirichlet-függvény definiciója: $D(x) = 1$, ha x racionális, és $D(x) = 0$, ha x irracionális. Ez a függvény minden pontban szakad.

$\tilde{y} \in (y_1, y_2)$ érték mellett $g(x, y_1) - g(x, y_2) = \partial_2 g(x, \tilde{y})(y_1 - y_2)$, azaz a (9.2.7) feltétel teljesül az $L = \sup_{H_g} (|\partial_2 g(x, y)|) < \infty$ állandóval[2]. $\diamond$

A fenti megjegyzés következménye: ha a (9.2.8) Cauchy-feladat f függvénye a Q_T halmazon folytonos, és a második változójában korlátos parciális deriválttal rendelkezik, akkor létezik egyértelmű megoldása a $[0, T]$ intervallumon.

9.3. Egylépéses módszerek

Vegyük észre, hogy az előző szakaszban felsorolt tételek a megoldás létezésére és annak egyértelműségére vonatkoztak, de nem adnak választ annak előállítására. Általában a közönséges differenciálegyenletek kezdetiérték-feladatainak megoldásai csak nagyon speciális f függvények esetén adhatók meg képletek segítségével. Ehelyett *numerikus megoldást* állítunk elő, ami azt jelenti, hogy az értelmezési tartományának egyes pontjaiban az ismeretlen megoldásfüggvény értékeit *véges számú lépéssel közelítőleg* határozzuk meg. A fejezet további részeiben ilyen eljárásokat ismertetünk. Ebben a részben az olyan típusú eljárásokkal foglalkozunk, ahol valamely rögzített időpontbeli közelítést *egy azt megelőző időpontbeli közelítés felhasználásával* határozunk meg. Az ilyen módszereket *egylépéses módszereknek* nevezzük .
Tehát a továbbiakban a célunk a

$$\frac{du}{dt} = f(t, u), \quad t \in [0, T], \tag{9.3.1}$$

$$u(0) = u_0 \tag{9.3.2}$$

feladat egylépéses módszerekkel történő közelítő megoldása, ahol $T > 0$ olyan szám, amely mellett a (9.3.1)–(9.3.2) feladatnak létezik egyértelmű, megfelelően sima megoldása a $[0, T]$ intervallumon.

9.3.1. Taylor-sorba fejtéses módszer

Ez az egyik legrégebben ismert módszer. A definíció alapján a (9.3.1) egyenlet $u(t)$ megoldására érvényes az

$$u'(t) = f(t, u(t)), \quad t \in [0, T] \tag{9.3.3}$$

azonosság. Tegyük fel, hogy az f függvény analitikus, és így tetszőleges rendű parciális deriváltjai léteznek a Q_T halmazon. Ekkor az $u(t)$ megoldásfüggvény is analitikus, és ezért akárhányszor differenciálható [7, 31]. A láncszabály alkalmazásával rendre deriválva a (9.3.3) azonosságot a $t^\star \in [0, T]$ pontban, a következő egyenlőségeket nyerjük:

$$\begin{aligned} u'(t^\star) &= f(t^\star, u(t^\star)), \\ u''(t^\star) &= \partial_1 f(t^\star, u(t^\star)) + \partial_2 f(t^\star, u(t^\star))\ u'(t^\star), \\ u'''(t^\star) &= \partial_{11} f(t^\star, u(t^\star)) + 2\partial_{12} f(t^\star, u(t^\star))\ u'(t^\star) + \partial_{22} f(t^\star, u(t^\star))\ (u'(t^\star))^2 + \\ &\quad + \partial_2 f(t^\star, u(t^\star))\ u''(t^\star). \end{aligned} \tag{9.3.4}$$

Vegyük észre, hogy $u(t^\star)$ ismeretében mindegyik derivált pontosan kiszámítható. (Megjegyezzük, hogy tetszőleges, magasabb rendű derivált hasonló módon kiszámítható, csak a képletek egyre bonyolultabbá válnak.) Tegyük fel, hogy $t > t^\star$ olyan, amelyre $[t^\star, t] \subset [0, T]$. Mivel az $u(t)$

[2] A $\partial_2 g(x, y)$ jelölés a g függvény második változója szerinti parciális deriváltat jelenti az (x, y) pontban.

megoldásfüggvény analitikus, ezért Taylor-sora előállítja a $t^\star$ pont valamely környezetében. Tehát a

$$T_{n,u}(t) = \sum_{k=0}^{n} \frac{u^{(k)}(t^\star)}{k!}(t-t^\star)^k \tag{9.3.5}$$

Taylor-polinom $n \to \infty$ esetén konvergál az $u(t)$ megoldáshoz, ha t megfelelően közel van a $t^\star$ ponthoz. Ezért a konvergencia-tartományon belül a megoldás előállítható az

$$u(t) = \sum_{k=0}^{\infty} \frac{u^{(k)}(t^\star)}{k!}(t-t^\star)^k \tag{9.3.6}$$

egyenlőséggel. A megoldás a (9.3.6) képlet szerinti Taylor-soros előállítása a gyakorlati számítások során kivitelezhetetlen: feltételezi, hogy a $t^\star$ pontban az f függvény *végtelen sok* parciális deriváltját ismerjük, továbbá, hogy egy rögzített t pontban a jobb oldali *végtelen numerikus sort* pontosan tudjuk összegezni.

Az $u(t)$ pontos érték kiszámítása tehát a (9.3.6) képlet szerint nem valósítható meg. Ezért a továbbiakban ennek *közelítését* igyekszünk meghatározni. Kézenfekvő ötlet, hogy a Taylor-sor *véges szeletét* tekintjük közelítésnek, azaz

$$u(t) \simeq \sum_{k=0}^{p} \frac{u^{(k)}(t^\star)}{k!}(t-t^\star)^k =: T_{p,u}(t), \tag{9.3.7}$$

és ekkor az elhagyott rész (azaz a hiba) $\mathcal{O}((t-t^\star)^{p+1})$ nagyságrendű. Definíció alapján, $T_{p,u}(t)$ az $u(t)$ függvény $t^\star$ pontbeli p-ed fokú Taylor -polinomja.

A (9.3.7) és a (9.3.4) összefüggések segítségével az alábbi közelítő eljárások definiálhatók.

a) Taylor-módszer

Válasszuk a $t^\star = 0$ pontot, ahol a kezdeti feltételünk adott[3]. Ekkor $u(t^\star) = u(0)$ ismert a kezdeti feltételből, és (9.3.4) alapján a deriváltak *pontosan* kiszámíthatók. Tehát a (9.3.7) közelítés alapján

$$u(t) \simeq \sum_{k=0}^{p} \frac{u^{(k)}(0)}{k!}t^k. \tag{9.3.8}$$

b) Lokális Taylor-módszer

Tekintsük az alábbi közelítő algoritmust.

1. A $[0,T]$ intervallumon a $t_0, t_1, \ldots, t_N$ intervallumbeli pontok megadásával kijelölünk egy $\omega_h := \{0 = t_0 < t_1 < \ldots < t_{N-1} < t_N = T\}$ ún. *rácshálót*, amelynek *lépésközeit* $h_i = t_{i+1} - t_i$, (ahol $i = 0, 1, \ldots N-1$,) míg *finomságát* $h = \max_i h_i$-vel jelöljük. (A továbbiakban ezekben a pontokban határozzuk meg a közelítéseket, és $u(t_i)$ közelítését y_i-vel, míg $u^{(k)}(t_i)$ közelítését $y_i^{(k)}$-val jelöljük, ahol $k = 0, 1, \ldots, p$.[4]
2. $y_0^{(k)}$ értékei a szükséges $k = 0, 1, \ldots, p$ értékekre a (9.3.4) összefüggések segítségével, $t^\star = 0$ behelyettesítéssel *pontosan* kiszámíthatók.

[3] Az előző szakaszban leírtakból következik, hogy értelmes a megoldásfüggvény $t = 0$ pontbeli deriváltjairól beszélni.

[4] Szokásosan a nulladik deriválási rend ($k = 0$) a függvényértéket jelenti.

3. Az
$$y_1 = \sum_{k=0}^{p} \frac{y_0^{(k)}}{k!} h_0^k, \tag{9.3.9}$$
képlettel meghatározzuk az $u(t_1)$ közelítését.
4. Az $i = 1, 2, \ldots, N-1$ értékekre y_i ismeretében a (9.3.4) összefüggések segítségével a $t^\star = t_i$ és $u(t^\star) = u(t_i) \approx y_i$ behelyettesítéssel *közelítőleg* meghatározzuk $y_i^{(k)}$ értékeit a $k = 0, 1, \ldots, p$ értékekre.
5. Az
$$y_{i+1} = \sum_{k=0}^{p} \frac{y_i^{(k)}}{k!} h_i^k, \tag{9.3.10}$$
képlettel meghatározzuk az $u(t_{i+1})$ közelítését.

Írjuk ki a lokális Taylor-módszer algoritmusát (9.3.10) alapján a $p = 0, 1, 2$ esetekre!

- Ha $p = 0$, akkor $y_i = y_0$ minden i értékre, így ez az eset a gyakorlat szempontjából érdektelen.
- Legyen $p = 1$. Ekkor
$$y_{i+1} = y_i + y_i' h_i = y_i + h_i f(t_i, y_i), \quad i = 0, 1, \ldots N-1, \tag{9.3.11}$$
ahol $y_0 = u_0$ adott.
- Legyen $p = 2$. Ekkor a számítási algoritmus
$$y_{i+1} = y_i + h_i y_i' + \frac{h_i^2}{2} y_i'' = y_i + h_i f(t_i, y_i) + \frac{h_i^2}{2} \left(\partial_1 f(t_i, y_i) + \partial_2 f(t_i, y_i) f(t_i, y_i) \right), \tag{9.3.12}$$
ahol $i = 0, 1, \ldots N-1$ és $y_0 = u_0$ adott.

Hasonlítsuk össze a Taylor-módszert és a lokális Taylor-módszert!

1. Mindkét módszer esetén a p-ed fokú Taylor-polinomot használjuk, ezért a módszer szükségessé teszi a $p-1$-ed rendig bezárólag valamennyi parciális derivált meghatározását. Ezek száma $p(p+1)/2 - 1$, és mindegyikben szükséges a függvények kiértékelése is. Ez már viszonylag kis p értékek mellett is rendkívül munkaigényes, és ez a módszer gyakorlati alkalmazhatóságának komoly korlátja[5]. Ezért a Taylor-módszer gyakorlatban elérhető pontossága behatárolt.
2. A Taylor-módszer ugyan p növelésével egyre pontosabban közelíti a megoldást, valamint tetszőleges pontban közvetlenül kiszámítható a közelítés, de csak olyan t értékekre, amelyek a Taylor-sor konvergencia-sugaránál kisebbek. Ez a módszer egyik legnagyobb hátránya: a konvergencia-sugár általában nem ismert, vagy a konvergencia-sugár kisebb T-nél, így a teljes $[0, T]$ intervallumon nem lehetséges a közelítő megoldás előállítása.
3. A Taylor-módszer előnye, hogy ha csak egy rögzített $t = \hat{t}$ pontban vagyunk kíváncsiak a közelítésre, és ez a pont a konvergenciatartományon belül helyezkedik el, akkor közvetlenül, egy lépésben meghatározható a közelítés. A lokális Taylor-módszer kiküszöböli a fenti hiányosságokat, hiszen h a szükséges módon, akármilyen kicsinek választható. Ugyanakkor ennél a módszernél n darab feladat megoldása szükséges, ahol $h_0 + h_1 + \ldots + h_{n-1} = \hat{t}$, mivel csak a teljes $[0, \hat{t}]$ időintervallumon tudjuk előállítani a közelítést.

[5] Az utóbbi években elterjedt szimbolikus számításos programok ugyan lehetőséget adnak az automatikus deriválásra, de a probléma még továbbra is fennáll.

4. A Taylor-módszer alkalmazása esetén a pontos és a közelítő megoldás eltérésére a Taylor-polinom hibatagjával becslést tudunk adni. A lokális Taylor-módszer esetén viszont a módszer hibája közvetlenül nem látszik, ugyanis az eltérés két részből adódik:
   a) minden lépésnél a Taylor-módszerhez hasonlóan a függvény n-ed fokú Taylor-polinommal történő approximációjából,
   b) a Taylor-polinom együtthatóit (azaz a megoldásfüggvény deriváltjait) csak közelítőleg tudjuk meghatározni. (Ráadásul, az itt elkövetett hiba a lépések során felhalmozódhat.)
5. Vegyük észre, hogy a fenti Taylor-módszerek felépítéséhez nem szükséges a megoldás analitikussága. Elegendő csak a megoldás $p+1$-szeres folytonos differenciálhatósága, azaz elegendő az $f \in C^p(Q_T)$ simasági feltétel.

9.3.1. példa. Tekintsük az

$$\begin{aligned} u' &= -u + t + 1, \quad t \in [0,1], \\ u(0) &= 1 \end{aligned} \tag{9.3.13}$$

feladatot, amelynek pontos megoldása $u(t) = \exp(-t) + t$. Ebben a feladatban $f(t,u) = -u + t + 1$, ezért

$$\begin{aligned} u'(t) &= -u(t) + t + 1, \\ u''(t) &= -u'(t) + 1 = u(t) - t, \\ u'''(t) &= -u(t) + t, \end{aligned} \tag{9.3.14}$$

tehát $u(0) = 1$, $u'(0) = 0$, $u''(0) = 1$, $u'''(0) = -1$. A Taylor-módszer esetén a közelítő polinomok:

$$\begin{aligned} T_{1,u}(t) &= 1, \\ T_{2,u}(t) &= 1 + t^2/2, \\ T_{3,u}(t) &= 1 + t^2/2 - t^3/6. \end{aligned} \tag{9.3.15}$$

Ezért a $t = 1$ pontban $T_{1,u}(1) = 1$, $T_{2,u}(1) = 1.5$, $T_{3,u}(1) = 1.333$. (Könnyen kiszámítható, hogy $T_{4,u}(1) = 1.375$ és $T_{5,u}(1) = 1.3666$.) Mint látható, ezek az értékek csak viszonylag magas n esetén közelítik megfelelően az $u(1) = 1.367879$ értéket.

Alkalmazzuk a (9.3.10) lokális Taylor-módszert, figyelembe véve a (9.3.14) deriváltakat. Az elsőrendű módszer algoritmusa

$$y_{i+1} = y_i + h_i(-y_i + t_i + 1), \quad i = 0, 1, \ldots, N-1, \tag{9.3.16}$$

míg a másodrendű módszer algoritmusa

$$y_{i+1} = y_i + h_i(-y_i + t_i + 1) + \frac{h_i^2}{2}(y_i - t_i), \quad i = 0, 1, \ldots, N-1,$$

ahol $h_1 + h_2 + \ldots + h_N = T$. Számításainkat a $h_i = h = 0.1$ egyenközű (ún. ekvidisztáns) rácshálón végezzük el. A 9.3.1. táblázatban a $[0,1]$ intervallum rácspontjaiban hasonlítjuk össze a lokális és globális Taylor-módszereket. (LT1 és LT2 a lokális első- illetve másodrendű Taylor-módszert, míg T1, T2 és T3 a első-, másod- és harmadrendű Taylor-módszereket jelöli.)

◇

t_i	a pontos megoldás	LT1	LT2	T1	T2	T3
0.1	1.0048	1.0000	1.0050	1.0000	1.0050	1.0048
0.2	1.0187	1.0100	1.0190	1.0000	1.0200	1.0187
0.3	1.0408	1.0290	1.0412	1.0000	1.0450	1.0405
0.4	1.0703	1.0561	1.0708	1.0000	1.0800	1.0693
0.5	1.1065	1.0905	1.1071	1.0000	1.1250	1.1042
0.6	1.1488	1.1314	1.1494	1.0000	1.1800	1.1440
0.7	1.1966	1.1783	1.1972	1.0000	1.2450	1.1878
0.8	1.2493	1.2305	1.2500	1.0000	1.3200	1.2347
0.9	1.3066	1.2874	1.3072	1.0000	1.4050	1.2835
1.0	1.3679	1.3487	1.3685	1.0000	1.5000	1.3333

9.3.1. táblázat: A lokális Taylor-módszer és a Taylor-módszer összehasonlítása a $h = 0.1$ lépésközű rácshálón

a h lépésköz	LT1	LT2	T1	T2	T3
0.1	$1.92e-02$	$6.62e-04$	0.3679	0.1321	0.0345
0.01	$1.80e-03$	$6.12e-06$	0.3679	0.1321	0.0345
0.001	$1.85e-04$	$6.14e-08$	0.3679	0.1321	0.0345
0.0001	$1.84e-05$	$6.13e-10$	0.3679	0.1321	0.0345

9.3.2. táblázat: A lokális Taylor-módszer és a Taylor-módszer hibája h lépésközű rácshálón a maximumnormában

A vizsgált módszerekkel nyert numerikus megoldás és a pontos megoldás eltérése a rácsháló pontjaiban meghatározza az $e_i = y_i - u(t_i$ koordinátájú hibavektort, és ennek maximumnormáját hasonlítjuk össze a 9.3.2. táblázatban a különböző, egyre finomodó rácshálókon. Jól látható, hogy h csökkenése esetén a lokális Taylor-módszer hibája csökken, viszont a Taylor-módszer által nyert eredmény változatlan marad. (Ez utóbbi természetes következménye annak, hogy a módszer független a lépésköz megválasztásától.)

A lokális Taylor-módszer egylépéses módszer, hiszen a t_i pontbeli értékek határozzák meg a t_{i+1}-beli közelítést. A hibaanalízise meglehetősen bonyolult. A korábbiakban leírtaknak megfelelően (és amit a fenti példa is jól mutat), y_{i+1} közelítésnek az $u(t_{i+1})$ pontos értéktől való eltérését több tényező okozza.

- Az ún. *lokális approximációs hiba*, ami a Taylor-sor Taylor-polinommal való helyettesítéséből ered, feltételezve, hogy a t_i pontbeli értéket pontosan ismerjük. Ennek a $[t_i, t_i + h_i]$ intervallum hossza szerinti rendjét, vagyis az $u(t_{i+1}) - T_{n,u}(t_{i+1})$ eltérés h_i szerinti rendjét *lokális hibarendnek* nevezzük. (Megfelelően sima függvények esetén ez a rend $\mathcal{O}(h_i^{p+1})$.)

- Minden lépésben (kivéve az elsőt) a sorbafejtésben nem a pontos t_i-beli értékek, hanem azoknak közelítései szerepelnek, és ezek az eltérések a lépések során felhalmozódhatnak.

- Minden számításnál *kerekítési hibák* is fellépnek, amelyek jelentősen torzíthatják a közelítést. Ez természetes velejárója a számítógépek korlátozott pontosságának, és nagysága függ a gépi pontosságtól. (A módszerek vizsgálata során mi ezen hibával nem foglalkozunk.)

- Amikor a $[0, T]$ intervallumon megoldjuk a feladatunkat, akkor egy $t^\star \in [0, T]$ pontbeli közelítés első két hibaforrásból eredő hibáját *globális hibának* nevezzük. Intuitív módon azt

mondjuk, hogy a módszer konvergens a $t = t^\star$ pontban, amikor a h maximális lépésköz nullához tartása esetén ez a globális hiba is nullához tart. A globális hiba nullához tartásának rendjét a módszer *konvergenciarendjének* nevezzük. Ez a rend független a kerekítési hibáktól. Mivel a $t = t^\star$ pontbeli közelítés meghatározásához kb. n lépést kell tennünk, ahol $nh = t^\star$, ezért $\mathcal{O}(h^{p+1})$ lokális csonkolási hiba mellett a globális konvergencia várható rendje $\mathcal{O}(h^p)$. (A 9.3.2. táblázat LT1 és LT2 módszereihez tartozó eredmények ezt alátámasztják: az LT1 elsőrendben, míg LT2 másodrendben konvergens a $t^\star = 1$ pontban.)

9.3.2. megjegyzés. A Taylor-módszer alkalmazása és viselkedése az $u' = 1 - t\sqrt[3]{u}$ differenciálegyenleten jól látható a
http://math.fullerton.edu/mathews/a2001/Animations/OrdinaryDE/Taylor/Taylor.html
linken lévő animáción. ◇

9.3.2. Néhány nevezetes egylépéses módszer

Az előző részben láttuk, hogy numerikus szempontból a lokális Taylor-módszer különösen $p = 1$ esetén előnyös: a (9.3.11) képlethez nem kell meghatározni az f függvény parciális deriváltjait, és a lépésközök csökkentésével a rácspontokban az ismeretlen függvény jól közelíthető. Ebben a részben az a célunk, hogy újabb, hasonló tulajdonságokkal rendelkező egylépéses módszereket definiáljunk.

Az LT1 módszert az ismeretlen $u(t)$ megoldásfüggvénynek a $[t_i, t_{i+1}]$ intervallumon $T_{1,u}(t)$ elsőrendű Taylor-polinommal való approximációjából nyertük.[6] Ekkor az elkövetett hiba (a lokális csonkolási hiba)

$$|u(t_{i+1}) - T_{1,u}(t_{i+1})| = \mathcal{O}(h_i^2), \quad i = 0, 1, \dots, N-1, \tag{9.3.17}$$

azaz másodrendben pontos az approximáció. Adjunk meg $T_{1,u}(t)$ helyett olyan más, $P_1(t)$ elsőfokú polinomot, amely mellett a (9.3.17) becslés továbbra is érvényben marad, azaz

$$|u(t_{i+1}) - P_1(t_{i+1})| = \mathcal{O}(h_i^2). \tag{9.3.18}$$

Mivel $T_{1,u}(t)$ a megoldásgörbéhez a $(t_i, u(t_i))$ pontbeli érintő, ezért olyan $P_1(t)$ polinomot keresünk, amely szintén átmegy ezen a ponton, de irányát – mivel a megoldásgörbe rácspontbeli értékeiből akarjuk meghatározni – az $u(t)$ függvény t_i és t_{i+1} pontbeli érintőinek iránya határozza meg. Ezért legyen $P_1(t) := u(t_i) + \alpha(t - t_i)$ $(t \in [t_i, t_{i+1}])$ alakú, ahol $\alpha = \alpha(u'(t_i), u'(t_{i+1}))$ egy adott függvény. (Például, az $\alpha = u'(t_i)$ megválasztással $P_1(t) = T_{1,u}(t)$, és ekkor természetesen (9.3.18) érvényes.)

Lehet-e más alkalmas megválasztás is? Mivel

$$u(t_{i+1}) = u(t_i) + u'(t_i)h_i + \mathcal{O}(h_i^2), \tag{9.3.19}$$

ezért

$$u(t_{i+1}) - P_1(t_{i+1}) = h_i(u'(t_i) - \alpha) + \mathcal{O}(h_i^2),$$

azaz (9.3.18) pontosan akkor teljesül, amikor az

$$\alpha - u'(t_i) = \mathcal{O}(h_i) \tag{9.3.20}$$

becslés érvényes.

[6] Mindegyik $[t_i, t_{i+1}]$ intervallumon más polinomot határozunk meg, de a polinomok ezen i-től való függését a jelöléseinkben nem hangsúlyozzuk.

9.3.3. tétel.
Tetszőleges $\theta \in \mathbb{R}$ esetén az

$$\alpha = (1-\theta)u'(t_i) + \theta u'(t_{i+1}) \tag{9.3.21}$$

megválasztású α függvény esetén a (9.3.20) becslés érvényes.

Bizonyítás. Alkalmazzuk a (9.3.19) felbontást az $u'(t)$ függvényre!

$$u'(t_{i+1}) = u'(t_i) + u''(t_i)h_i + \mathcal{O}(h_i^2), \tag{9.3.22}$$

és behelyettesítve a (9.3.22) összefüggést a (9.3.21) képletbe,

$$\alpha - u'(t_i) = \theta u''(t_i)h_i + \mathcal{O}(h_i^2) \tag{9.3.23}$$

összefüggést nyerjük, ami az állításunkat bizonyítja. ■

9.3.4. következmény. A fenti $P_1(t)$ polinom az

$$y_{i+1} = y_i + \alpha h_i \tag{9.3.24}$$

egylépéses numerikus módszert határozza meg, ahol a (9.3.21) és a (9.3.1) összefüggések alapján

$$\alpha = (1-\theta)f(t_i, y_i) + \theta f(t_{i+1}, y_{i+1}). \tag{9.3.25}$$

◇

9.3.5. definíció.
A (9.3.24)-(9.3.25) numerikus módszert θ-módszernek nevezzük.

9.3.6. megjegyzés. A θ-módszer esetén is jellemző, hogy y_i valamilyen közelítése az $u(t_i)$ pontos értéknek, és az eltérés – a Taylor-módszernél is említettekkel megegyezően – alapvetően a következők miatt van:

a) minden lépésnél az $u(t)$ megoldásfüggvényt az elsőfokú $P_1(t)$ polinommal approximáljuk,

b) a $P_1(t)$ polinomban az α együtthatót (azaz a megoldásfüggvény deriváltjait) csak közelítőleg tudjuk meghatározni.

◇

Mivel az α irányt a megoldásfüggvény t_i és t_{i+1} pontbeli érintőinek iránya határozza meg, ezért általában úgy választjuk meg, hogy ezen két érték közé essék. Ezért a θ paramétert csak a $[0,1]$ intervallumból szokásos megválasztani. A továbbiakban három, speciálisan megválasztott $\theta \in [0,1]$ értékhez tartozó numerikus módszert vizsgálunk meg.

Az explicit Euler-módszer

Tekintsük a θ-módszert a $\theta = 0$ megválasztással! Ekkor (9.3.24) és (9.3.25) a következő numerikus módszert generálják:

$$y_{i+1} = y_i + h_i f(t_i, y_i), \quad i = 0, 1, \ldots, N-1. \tag{9.3.26}$$

Mivel y_i az ismeretlen $u(t)$ függvény t_i pontbeli közelítése, ezért értelemszerűen

$$y_0 = u(0) = u_0, \tag{9.3.27}$$

vagyis a (9.3.26) iterációban az $i = 0$ értékhez tartozó y_0 adott érték.

9.3.7. definíció.
A (9.3.26)–(9.3.27) képletekkel definiált egylépéses módszert *explicit Euler-módszernek* nevezzük.

Mivel a $\theta = 0$ esetén $\alpha = u'(t_i)$, ezért ebben az esetben a módszert definiáló P_1 polinom megegyezik az elsőrendű Taylor-polinommal. Tehát az explicit Euler-módszer azonos a (9.3.11) képlettel definiált elsőfokú közelítéses lokális Taylor-módszerrel.

9.3.8. megjegyzés. A (9.3.26)–(9.3.27) módszert azért nevezzük explicitnek, mert a t_i pontbeli érték ismeretében közvetlenül, egy egyszerű függvénybehelyettesítéssel kiszámítható a t_{i+1} pontbeli közelítés. ◇

Vizsgáljuk meg először, hogy egy rögzített

$$\omega_h := \{t_i = ih;\ i = 0, 1, \dots, N;\ h = T/N\}$$

ekvidisztáns rácshálón milyen becslés adható az explicit Euler-módszer által nyert numerikus közelítés és a pontos megoldás eltérésére valamely $t_n \in \omega_h$ pontban! (Az előzőekben leírtaknak megfelelően továbbra is feltesszük, hogy az f függvény a második változójában lipschitzes, és a megoldás megfelelően sima.)

Jelölje

$$e_i = y_i - u(t_i),\ i = 0, 1, \dots, N \tag{9.3.28}$$

egy tetszőleges $t_i \in \omega_h$ rácspontbeli pontbeli hibát.

A (9.3.26) explicit Euler-módszer képletébe behelyettesítve a (9.3.28) definícióból következő $y_i = e_i + u(t_i)$ kifejezést, érvényes a következő egyenlőség:

$$\begin{aligned} e_{i+1} - e_i &= -(u(t_{i+1}) - u(t_i)) + hf(t_i, e_i + u(t_i)) \\ &= [hf(t_i, u(t_i)) - (u(t_{i+1}) - u(t_i))] + h\left[f(t_i, e_i + u(t_i)) - f(t_i, u(t_i))\right]. \end{aligned} \tag{9.3.29}$$

Így, bevezetve a

$$g_i = hf(t_i, u(t_i)) - (u(t_{i+1}) - u(t_i)), \quad \psi_i = f(t_i, e_i + u(t_i)) - f(t_i, u(t_i)) \tag{9.3.30}$$

jelöléseket az

$$e_{i+1} - e_i = g_i + h\psi_i \tag{9.3.31}$$

ún. *hibaegyenletet* kapjuk.

9.3.9. megjegyzés. Vizsgáljuk meg a (9.3.30) definícióban szereplő két tagot! A g_i tag azt mutatja, hogy a pontos megoldás az explicit Euler-módszer (9.3.26) $hf(t_i, y_i) - (y_{i+1} - y_i) = 0$ alakban felírt képletét milyen pontosan elégíti ki. Ez a kifejezés azt a hibát tartalmazza, ami az $u(t)$ megoldásfüggvénynek a $[t_i, t_{i+1}]$ intervallumon történő, elsőfokú Taylor-polinommal való approximációjából ered. (Ezt neveztük lokális approximációs hibának.) A ψ_i tag azt jellemzi, hogy a módszer egy lépése során mekkora hiba keletkezik abból, hogy az y_{i+1} érték kiszámolására szolgáló képletben a pontos $u(t_i)$ érték helyett annak y_i közelítésével számolunk. ◇

Az f függvény lipschitzessége következtében

$$|\psi_i| = |f(t_i, e_i + u(t_i)) - f(t_i, u(t_i))| \leq L|(e_i + u(t_i)) - u(t_i)| = L|e_i|. \tag{9.3.32}$$

Így a (9.3.31) és a (9.3.32) összefüggések alapján

$$|e_{i+1}| \leq |e_i| + |g_i| + h|\psi_i| \leq (1 + hL)|e_i| + |g_i| \tag{9.3.33}$$

minden $i = 0, 1, \ldots, n-1$ értékre. Ezért

$$\begin{aligned} |e_n| &\leq (1 + hL)|e_{n-1}| + |g_{n-1}| \leq (1 + hL)\left[(1 + hL)|e_{n-2}| + |g_{n-2}|\right] + |g_{n-1}| \\ &= (1 + hL)^2|e_{n-2}| + \left[(1 + hL)|g_{n-2}| + |g_{n-1}|\right] \\ &\leq (1 + hL)^n|e_0| + \sum_{i=0}^{n-1}(1+hL)^i|g_{n-1-i}| < (1 + hL)^n \left[|e_0| + \sum_{i=0}^{n-1}|g_{n-1-i}|\right]. \end{aligned} \tag{9.3.34}$$

(Az utolsó lépésben az $(1 + hL)^i < (1 + hL)^n$, $i = 0, 1, \ldots, n-1$ egyenlőtlenséget alkalmaztuk.) Mivel tetszőleges pozitív x esetén $1 + x < \exp(x)$, ezért az $nh = t_n$ reláció következtében $(1 + hL)^n < \exp(nhL) = \exp(Lt_n)$. Így (9.3.34) alapján

$$|e_n| \leq \exp(Lt_n)\left[|e_0| + \sum_{i=0}^{n-1}|g_{n-1-i}|\right]. \tag{9.3.35}$$

Adjunk becslést a $|g_i|$ kifejezésre! Könnyen láthatóan

$$u(t_{i+1}) - u(t_i) = u(t_i + h) - u(t_i) = hu'(t_i) + \frac{1}{2}u''(\xi_i)h^2, \tag{9.3.36}$$

ahol $\xi_i \in (t_i, t_{i+1})$ egy adott pont. Mivel $f(t_i, u(t_i)) = u'(t_i)$, ezért a

$$\max_{[0,t_n]}|u''(t)| \leq \max_{[0,T]}|u''(t)| =: M_2$$

reláció miatt a g_i tagra - a (9.3.30) szerinti definíciója következtében- érvényes a

$$|g_i| \leq \frac{M_2}{2}h^2 \tag{9.3.37}$$

egyenlőtlenség. Ekkor a (9.3.35) és a (9.3.37) becslések alapján

$$|e_n| \leq \exp(Lt_n)\left[|e_0| + hn\frac{M_2}{2}h\right] = \exp(Lt_n)\left[|e_0| + \frac{t_nM_2}{2}h\right] \tag{9.3.38}$$

becslés. Mivel $e_0 = 0$, ezért érvényes a

$$|e_n| \leq \exp(Lt_n)\frac{t_nM_2}{2}h, \quad n = 0, 1, \ldots, N \tag{9.3.39}$$

becslés.

A 9.3.6. megjegyzésben felsoroltuk a módszer hibájának, azaz a pontos és a közelítő megoldás eltérésének forrásait. A módszer alkalmazásának fő célja annak biztosítása, hogy a finomodó rácshálók sorozatán előállított numerikus közelítések tartsanak (*konvergáljanak*) a pontos megoldáshoz. Ezt akérédst több szempontból is megvizsgálhatjuk.

a. Egy rögzített $t^\star \in [0,T]$ pontban a rácsháló finomításával hogyan viselkedik a közelítő és pontos érték különbsége?

b. Tetszőleges rögzített $t^\star \in [0,T]$ pont esetén a rácsháló finomításával hogyan viselkedik a közelítő és pontos érték különbsége a $[0,t^\star]$ intervallumon?

Először az a. kérdéssel foglalkozunk.

Ekvidisztáns rácshálók sorozatán viszonylag egyszerűen belátható a következő állítás.

9.3.10. tétel.
Legyen $t^\star \in [0,T]$ tetszőleges rögzített pont. Tekintsük $h \to 0$ mellett a $[0,t^\star]$ intervallumon az

$$\omega_h := \{t_i = ih;\ i = 0,1,\ldots,n;\ h = t^\star/n\} \tag{9.3.40}$$

ekvidisztáns rácshálók sorozatát. Ekkor a $t^\star$ pontban az explicit Euler módszer elsőrendben konvergens.

Bizonyítás. Nyilvánvalóan n index h-tól függ, és $h \to 0$ esetén $n \to \infty$. Emellett tetszőleges h esetén $t_n = nh = t^\star$. Azt kell megmutatnunk, hogy az $e_n = y_n - u(t^\star)$ hibára $e_n = Ch$, ahol C egy h-tól független állandó. Tekintsük a (9.3.39) egyenlőtlenséget! Mivel $t_n = t^\star$, ezért

$$|e_n| \le \exp(Lt^\star)\frac{t^\star M_2}{2}h. \tag{9.3.41}$$

Így az állításunk közvetlelül következik a $C = 0.5\exp(Lt^\star)t^\star M_2$ állandóval. ■

Az explicit Euler-módszer rögzített pontbeli konvergenciájának fenti bizonyítása viszonylag egyszerűen kiterjeszthető az alkalmasan megválasztott nemekvidisztáns rácshálókra is. Legyen most

$$\omega_{h_v} := \{0 = t_0 < t_1 < \ldots < t_{n-1} < t_n = t^\star\} \tag{9.3.42}$$

egy változó lépéshosszúságú rácsháló a $[0,t^\star]$ intervallumon. Vezessük be a

$$h_i = t_{i+1} - t_i, \quad h = \max_{i=0,\ldots,n-1} h_i, \quad h_{min} = \min_{i=0,\ldots,n-1} h_i$$

jelöléseket.

A

$$g_i = h_i f(t_i, u(t_i)) - (u(t_{i+1}) - u(t_i)), \quad \psi_i = f(t_i, e_i + u(t_i)) - f(t_i, u(t_i)) \tag{9.3.43}$$

jelölésekkel a (9.3.33) becslés így írható át:

$$\begin{aligned} |e_{i+1}| &\le |e_i| + |g_i| + h_i|\psi_i| \le |e_i| + |g_i| + h_i L|e_i| \le \\ (1 + h_i L)|e_i| + |g_i| &\le \exp(h_i L)|e_i| + |g_i| \le \exp(h_i L)\left[|e_i| + |g_i|\right]. \end{aligned} \tag{9.3.44}$$

Ekkor a $t^\star = t_n \in \omega_{h_v}$ rácspontbeli a (9.3.34) hiba becslése, a (9.3.44) reláció figyelembevételével így alakul:

$$\begin{aligned} |e_n| &\le \exp(h_{n-1}L)\left[|e_{n-1}| + |g_{n-1}|\right] \\ &\le \exp(h_{n-1}L)\left[\exp(h_{n-2}L)\left(|e_{n-2}| + |g_{n-2}|\right) + |g_{n-1}|\right] \\ &= \exp((h_{n-1} + h_{n-2})L)\left(|e_{n-2}| + |g_{n-2}| + g_{n-1}|\right) \\ &\le \exp((h_{n-1} + h_{n-2} + \ldots + h_0)L)\left[|e_0| + \sum_{i=1}^{n} |g_{n-i}|\right]. \end{aligned} \tag{9.3.45}$$

Jelölje

$$K_v \equiv K(\omega_{h_v}) = h/h_{min} \tag{9.3.46}$$

az ω_{h_v} rácsháló felosztását jellemző számot. Ekkor a

$$|g_i| \leq \frac{M_2}{2} h_i^2 \tag{9.3.47}$$

becslés alapján

$$|g_i| \leq \frac{M_2}{2} h^2 \leq \frac{M_2}{2} h K_v h_{min}. \tag{9.3.48}$$

Nyilvánvalóan $h_{n-1} + h_{n-2} + \ldots + h_0 = t_n = t^\star$ és $nh_{min} \leq t_n = t^\star$, ezért (9.3.45) és (9.3.48) alapján

$$|e_n| \leq \exp(t^\star L) \left[|e_0| + t^\star \frac{M_2 K_v}{2} h \right]. \tag{9.3.49}$$

Mivel $e_0 = 0$, ezért

$$|e_n| \leq \exp(L t^\star) \left(\frac{M_2 t^\star K_v}{2} \right) h. \tag{9.3.50}$$

A változó rácshálón való konvergenciához tekintsük az (ω_{h_v}) rácshálók olyan sorozatát, amelyek az alábbi tulajdonságokkal rendelkeznek:

- Létezik olyan $K > 0$ állandó, hogy mindegyik rácshálóra

$$K_v \leq K. \tag{9.3.51}$$

- Az egyes rácshálók h maximális lépésköze nullához tart.

A továbbiakban a fenti tulajdonságú rácshálósorozatot *regulárisnak* nevezzük.

9.3.11. megjegyzés. Az osztásrészek számának növelésével előállított ekvidisztáns rácshálósorozatok regulárisak, hiszen (9.3.51) a $K = 1$ állandóval teljesül. Továbbá az is könnyen meggondolható, hogy a reguláris rácshálósorozat második feltételéből nem következik a (9.3.51) tulajdonság. ⋄

A (9.3.50) becslés alapján közvetlenül belátható az alábbi állítás.

9.3.12. tétel.
Legyen $t^\star \in [0, T]$ egy rögzített pont az (ω_{h_v}) a $[0, t^\star]$ intervallumon értelmezett reguláris rácshálósorozat. Ekkor a $t^\star$ pontban az explicit Euler módszer elsőrendben konvergens.

Bizonyítás. Mivel a rácshálósorozat reguláris, ezért a (9.3.50) becslés felírható az

$$|e_n| \leq \exp(L t^\star) \left(\frac{M_2 t^\star K}{2} \right) h. \tag{9.3.52}$$

alakban. Innen a 9.3.10. tétel bizonyítása megismételhető. ■

Térjünk át a b. kérdésre! Az eddigi konvergenciavizsgálatok során azt vizsgáltuk, hogy a $[0, T]$ intervallum egy rögzített $t^\star$ pontjában a $[0, t^\star]$ intervallumon értelmezett rácshálókon a közelítő megoldások sorozata tart-e a pontos megoldáshoz a $t = t^\star$ pontban. Most azt vizsgáljuk meg, hogy a megfelelő rácshálósorozatokon előállított numerikus megoldások sorozata vajon az intervallum mindegyik pontjában konvergens-e?

9.3.13. tétel.
Legyen (ω_{h_v}) egy reguláris rácshálósorozat a $[0, t^\star]$ intervallumon. Ekkor az intervallum mindegyik pontjában az explicit Euler módszer elsőrendben konvergens.

Bizonyítás. A rácshálósorozat $\lim h = 0$ tulajdonsága következtében minden $t \in [0, t^\star]$ ponthoz létezik olyan (t_{n_v}) sorozat, hogy mindegyik rácshálóra $t_{n_v} \in \omega_{h_v}$ és $\lim_{n_v \to \infty} t_{n_v} = t$. Ezért ezekben a t_{n_v} pontokban felírva a (9.3.52) becslést, az

$$|e_{n_v}| = |y_{n_v} - u(t_{n_v})| \leq Ch \tag{9.3.53}$$

alakban, ahol

$$C = \exp(Lt^\star) \left(\frac{M_2 t^\star K}{2} \right) \tag{9.3.54}$$

adott állandó. Mivel $h \to 0$ esetén $n_v \to \infty$, ezért a (9.3.53) egyenlőtlenségben áttérve a $h \to 0$ határérékre, valamint felhasználva az u függvény folytonosságát, a tételünk állítását jelentő $\lim_{h \to 0} y_{n_v} = u(t)$ összefüggést kapjuk. ■

Nyilvánvalóan elegendő volt az állítást csak a reguláris rácshálósorozaton megmutatni, hiszen az ekvidisztáns felosztáson való konvergencia ebből már következik.

9.3.14. következmény. Vegyük észre, hogy a (9.3.54) módon definiált C kifejezés $t^\star$ monoton növő függvénye. Ezért tehát a 9.3.13. tétel alapján érvényes az alábbi állítás.

9.3.15. tétel.
Legyen (ω_{h_v}) egy reguláris rácshálósorozat a $[0, T]$ intervallumon. Ekkor az intervallumon mindegyik pontjában az explicit Euler módszer elsőrendben konvergens a $C = 0.5 M_2 T K \exp(LT)$ állandóval.

◇

9.3.16. megjegyzés. Látható, hogy az explicit Euler-módszer esetén a $\lim_{h \to 0} e_n = 0$ egyenlőséghez nem szükséges az $y_0 = u_0$ megválasztás, elegendő, ha $y_0 = u_0 + \mathcal{O}(h)$, mert ebben az esetben $e_0 = \mathcal{O}(h)$. (Emellett, továbbra is $e_n = \mathcal{O}(h)$.) ◇

Az implicit Euler-módszer

Tekintsük a θ-módszert a $\theta = 1$ megválasztással! Ekkor (9.3.24) és (9.3.25) a következő numerikus módszert generálja:

$$y_{i+1} = y_i + h_i f(t_{i+1}, y_{i+1}), \quad i = 0, 1, \ldots, N-1, \tag{9.3.55}$$

$$y_0 = u_0. \tag{9.3.56}$$

9.3.17. definíció.
A (9.3.55)–(9.3.56) képletekkel definiált egylépéses módszert *implicit Euler-módszernek* nevezzük.

9.3.18. megjegyzés. A (9.3.55) implicit Euler-módszert azért nevezzük implicitnek, mert az időben való előrehaladáshoz y_i ismeretében y_{i+1} értékét minden egyes időlépésben egy (tipikusan nemlineáris) egyenlet megoldásával tudjuk csak meghatározni. ◇

Az implicit Euler-módszer e_i hibafüggvényére a hibaegyenlet a következő módon írható fel:

$$\begin{aligned} e_{i+1} - e_i &= -(u(t_{i+1}) - u(t_i)) + h_i f(t_{i+1}, u(t_{i+1}) + e_{i+1}) \\ &= [h_i f(t_{i+1}, u(t_{i+1})) - (u(t_{i+1}) - u(t_i))] + h_i \left[f(t_{i+1}, u(t_{i+1}) + e_{i+1}) - f(t_{i+1}, u(t_{i+1}))\right]. \end{aligned} \tag{9.3.57}$$

Így a

$$g_i = h_i f(t_{i+1}, u(t_{i+1})) - (u(t_{i+1}) - u(t_i)), \quad \psi_i = f(t_{i+1}, u(t_{i+1}) + e_{i+1}) - f(t_{i+1}, u(t_{i+1})) \tag{9.3.58}$$

jelölésekkel ismételten a (9.3.31) alakú hibaegyenletet nyerjük.
Nyilvánvalóan

$$u(t_{i+1}) - u(t_i) = u(t_{i+1}) - u(t_{i+1} - h_i) = h_i u'(t_{i+1}) - \frac{1}{2} u''(\xi_i) h_i^2, \tag{9.3.59}$$

ahol $\xi_i \in (t_i, t_{i+1})$ egy adott pont. Másrészt, $f(t_{i+1}, u(t_{i+1})) = u'(t_{i+1})$. Ezért g_i (9.3.58) szerinti definíciója következtében érvényes a

$$|g_i| \leq \frac{M_2}{2} h_i^2 \tag{9.3.60}$$

egyenlőtlenség. Ugyanakkor az implicit Euler-módszer esetén ψ_i a t_{i+1} pontbeli közelítéstől és a pontos megoldástól egyaránt függ, ezért az explicit Euler-módszer vizsgálata ebben az esetben változatlan formában nem ismételhető meg. (A módszer e_n hibafüggvényének nullához tartásával később foglalkozunk.)

A Crank–Nicolson-módszer

Tekintsük a θ-módszert a $\theta = 0.5$ megválasztással! Ekkor (9.3.24) és (9.3.25) a következő numerikus módszert generálja:

$$y_{i+1} - y_i = \frac{h_i}{2} \left[f(t_i, y_i) + f(t_{i+1}, y_{i+1})\right], \quad i = 0, 1, \ldots, N-1, \tag{9.3.61}$$

ahol $y_0 = u_0$.

9.3.19. definíció.
A (9.3.61) egylépéses módszert *Crank–Nicolson-módszernek* nevezzük.

Vegyük észre, hogy a Crank–Nicolson-módszer is implicit.

A Crank–Nicolson-módszer hibafüggvényére az explicit és implicit Euler-módszerek hibaegyenleteinek kombinálásával könnyen nyerhető a (9.3.31) alakú hibaegyenletet, ahol most

$$\begin{aligned} g_i &= \frac{1}{2} h_i \left[f(t_i, u(t_i)) + f(t_{i+1}, u(t_{i+1}))\right] - (u(t_{i+1}) - u(t_i)), \\ \psi_i &= \frac{1}{2} \left[f(t_i, u(t_i) + e_i) - f(t_i, u(t_i))\right] + \frac{1}{2} \left[f(t_{i+1}, u(t_{i+1}) + e_{i+1}) - f(t_{i+1}, u(t_{i+1}))\right]. \end{aligned} \tag{9.3.62}$$

t_i	a pontos megoldás	EE	IE	CN
0.1	1.0048	1.0000	1.0091	1.0048
0.2	1.0187	1.0100	1.0264	1.0186
0.3	1.0408	1.0290	1.0513	1.0406
0.4	1.0703	1.0561	1.0830	1.0701
0.5	1.1065	1.0905	1.1209	1.1063
0.6	1.1488	1.1314	1.1645	1.1485
0.7	1.1966	1.1783	1.2132	1.1963
0.8	1.2493	1.2305	1.2665	1.2490
0.9	1.3066	1.2874	1.3241	1.3063
1.0	1.3679	1.3487	1.3855	1.3676

9.3.3. táblázat: Az explicit Euler-módszer (EE), az implicit Euler-módszer (IE) és a Crank–Nicolson módszer (CN) összehasonlítása a $h = 0.1$ lépésközű rácshálón

Adjunk becslést a (9.3.62) képletben szereplő g_i kifejezésre! A $t_{i+\frac{1}{2}} = t_i + 0.5h_i$ jelöléssel fejtsük sorba az $u(t_i) = u(t_{i+\frac{1}{2}} - h_i/2)$ és az $u(t_{i+1}) = u(t_{i+\frac{1}{2}} + h_i/2)$ kifejezéseket a $t = t_{i+\frac{1}{2}}$ pont körül. Ekkor

$$u(t_{i+1}) - u(t_i) = h_i u'(t_{i+1/2}) + \frac{h_i^3}{48}(u'''(\xi_i^1) + u'''(\xi_i^2)), \tag{9.3.63}$$

ahol $\xi_i^1, \xi_i^2 \in (t_i, t_{i+1})$ adott pontok. Másrészt

$$f(t_i, u(t_i)) + f(t_{i+1}, u(t_{i+1})) = u'(t_i) + u'(t_{i+1}). \tag{9.3.64}$$

Sorba fejtve a (9.3.64) jobb oldali függvényeit a $t = t_{i+\frac{1}{2}}$ pont körül, az

$$\frac{1}{2}\left[f(t_i, u(t_i)) + f(t_{i+1}, u(t_{i+1}))\right] = u'(t_{i+\frac{1}{2}}) + \frac{h_i^2}{16}(u'''(\xi_i^3) + u'''(\xi_i^4)). \tag{9.3.65}$$

egyenlőséget kapjuk. Ezért (9.3.63) és (9.3.65) alapján a g_i kifejezésre érvényes a

$$|g_i| \le \frac{M_3}{6} h_i^3, \quad M_3 = \max_{[0,t^\star]} |u'''(t)| \tag{9.3.66}$$

egyenlőtlenség.

A 9.3.3. táblázatban a (9.3.13) tesztfeladat fenti három numerikus módszerrel való megoldását ismertetjük. A 9.3.4. táblázatban a hibákat a maximumnormában hasonlítjuk össze a különböző, egyre finomodó rácshálókon. Az eredményekből megállapítható, hogy rögzített rácshálón a numerikus megoldás explicit Euler-módszer és implicit Euler-módszer esetén nagyjából hasonló pontosságot ad, míg a Crank–Nicolson-módszer pontosabb az előző két módszernél. A finomodó rácshálókon azt figyelhetjük meg, hogy a Crank–Nicolson-módszer hibafüggvénye $\mathcal{O}(h^2)$, az explicit Euler-módszer és az implicit Euler-módszer hibafüggvénye viszont csak $\mathcal{O}(h)$ rendben tart nullához. (Az általános alakú θ-módszer esetén a hibafüggvény nullához tartását a későbbiekben bizonyítjuk be.)

Ezen rész befejezéseként megemlítjük, hogy a fentiekben tárgyalt egylépéses módszerek más módon is bevezethetők. Az egyik lehetséges út a következő. Legyen $u(t)$ a (9.2.6) egyenlet megoldása, azaz érvényes rá a (9.3.3) azonosság a $[0, T]$ intervallum mindegyik pontjában. Az azonosság mindkét oldalát integrálva az ω_h rácsháló két tetszőleges szomszédos pontja között az

$$u(t_{i+1}) - u(t_i) = \int_{t_i}^{t_{i+1}} f(t, u(t))\,\mathrm{d}t, \quad t \in [0, T] \tag{9.3.67}$$

a h lépésköz	EE	IE	CN
0.1	$1.92e-02$	$1.92e-02$	$3.06e-04$
0.01	$1.84e-03$	$1.84e-03$	$3.06e-06$
0.001	$1.84e-04$	$1.84e-04$	$3.06e-08$
0.0001	$1.84e-05$	$1.84e-05$	$3.06e-10$
0.00001	$1.84e-06$	$1.84e-06$	$5.54e-12$

9.3.4. táblázat: Az explicit Euler-módszer (EE), az implicit Euler-módszer (IE) és a Crank–Nicolson módszer (CN) hibája h lépésközű rácshálón a maximumnormában.

($i = 0, 1, \ldots, N-1$ tetszőleges) egyenlőséget nyerjük. A közelítő módszerek megkonstruálásához a jobb oldalon lévő integrált valamely közelítő formulával számoljuk ki a $[t_i, t_{i+1}]$ intervallumon. A legegyszerűbb numerikus integrálás, amikor az integrálandó függvény intervallum valamely végpontjában, avagy mindkettőben felvett értéke szerepel csak a numerikus integráló formulákban. A különböző numerikus integráló formulák elvezetnek a fenti módszerekhez. Nevezetesen,

- A legegyszerűbb módszer, amikor a téglalapszabályt alkalmazzuk az integrálandó függvény bal oldali végpontjában felvett értékének felhasználásával, azaz

$$\int_{t_i}^{t_{i+1}} f(t, u(t))\,\mathrm{d}t \approx h_i f(t_i, u(t_i)). \tag{9.3.68}$$

  Ekkor a (9.3.67) és a (9.3.68) összefüggések a (9.3.26) képlettel megadott explicit Euler-módszert eredményezik.

- Egy további lehetséges módszer a (9.3.67) azonosságban szereplő integrál közelítő meghatározására, hogy az intervallum jobb oldali végpontbeli függvényértéket felhasználva a téglalapszabályt alkalmazzuk, azaz

$$\int_{t_i}^{t_{i+1}} f(t, u(t))\,\mathrm{d}t \approx h_i f(t_{i+1}, u(t_{i+1})). \tag{9.3.69}$$

  Ekkor (9.3.67) és (9.3.69) felhasználásával a (9.3.55) alakú implicit Euler-módszert kapjuk.

- Ha (9.3.67) közelítő integrálására a trapézszabályt alkalmazzuk, azaz

$$\int_{t_i}^{t_{i+1}} f(t, u(t))\,\mathrm{d}t \approx \frac{h_i}{2}\left[f(t_i, u(t_i)) + f(t_{i+1}, u(t_{i+1}))\right], \tag{9.3.70}$$

  akkor a (9.3.67) és a (9.3.70) képletek felhasználásával a (9.3.61) alakú Crank–Nicolson-módszert kapjuk. (Emiatt szokásos a Crank–Nicolson-módszert *trapézszabálynak* is nevezni.)

Egy másik lehetséges mód a fenti módszerek származtatására, amikor a (9.3.3) azonosságot valamely rögzített rácspontban felírva, a bal oldalon lévő deriváltra egy numerikus deriválási formulát alkalmazunk.

- Ha a $t = t_i$ pontban írjuk fel a (9.3.3) azonosságot, akkor az $u'(t_i) = f(t_i, u(t_i))$ egyenlőséget kapjuk. A bal oldalon szereplő deriváltat a haladó véges differenciával közelítve az $u'(t_i) \simeq (u(t_{i+1}) - u(t_i))/h_i$ közelítést kapjuk. Ez a két formula pedig az explicit Euler-módszert generálja.

- Ha a $t = t_{i+1}$ pontban írjuk fel a (9.3.3) azonosságot, és a retrográd numerikus deriválást alkalmazzuk, akkor az implicit Euler-módszert kapjuk.
- A fenti két képlet számtani közepe szolgáltatja a Crank–Nicolson-módszer képletét.

9.3.20. megjegyzés. Az Euler-módszer viselkedése az $u' = 1 - t\sqrt[3]{u}$ differenciálegyenletre jól látható a
http://math.fullerton.edu/mathews/a2001/Animations/Animations9.html
linken megtalálható animáción. ◇

9.3.3. Az általános alakú egylépéses módszerek alapfogalmai és pontbeli konvergenciája

Ebben a szakaszban az

$$\omega_h := \{t_i = ih;\ i = 0, 1, \dots, N;\ h = T/N\}$$

ekvidisztáns rácshálón (illetve azok sorozatán) megadjuk az egylépéses módszerek általános alakját, és definiáljuk a numerikus módszerek alapfogalmait. Tekintsük az

$$y_{i+1} = y_i + h\Phi(h, t_i, y_i, y_{i+1}) \tag{9.3.71}$$

egylépéses módszert, ahol Φ a numerikus módszert meghatározó adott függvény. A továbbiakban azt a numerikus módszert, amelyet a (9.3.71) képlet realizál, Φ *numerikus módszernek* (röviden: Φ-*módszernek*) nevezzük.

9.3.21. megjegyzés. Speciálisan megválasztott Φ függvények esetén a korábbi módszereink előállíthatók a (9.3.71) alakban. Például,

- $\Phi(h, t_i, y_i, y_{i+1}) = f(t_i, y_i)$ esetén az explicit Euler-módszert;
- $\Phi(h, t_i, y_i, y_{i+1}) = f(t_i + h, y_{i+1})$ esetén az implicit Euler-módszert;
- $\Phi(h, t_i, y_i, y_{i+1}) = 0.5\,[f(t_i, y_i) + f(t_i + h, y_{i+1})]$ esetén a Crank–Nicolson-módszert,
- $\Phi(h, t_i, y_i, y_{i+1}) = (1 - \theta)f(t_i, y_i) + \theta f(t_i + h, y_{i+1})$ esetén a θ-módszert

kapjuk. ◇

9.3.22. definíció.
Azokat a módszereket, amelyekre $\Phi = \Phi(h, t_i, y_i)$ (tehát a Φ függvény nem függ y_{i+1}-től, azaz a Φ függvényben nem szerepel y_{i+1}) *explicit módszereknek* nevezzük. Amennyiben $\Phi = \Phi(h, t_i, y_i, y_{i+1})$ (azaz a Φ függvényben szerepel y_{i+1} is), a módszert *implicitnek* nevezzük.

Jelölje továbbra is $u(t)$ a (9.3.1)–(9.3.2) feladat pontos megoldását. A Φ-módszer lokális viselkedését jól jellemzi, hogy az a közelítés, amelyet a módszerrel – a pontos megoldásból indulva – egy lépés elvégzése után nyerünk, milyen közel van a pontos megoldás értékéhez ebben a pontban, azaz az

$$\hat{y}_{i+1} = u(t_i) + h\Phi(h, t_i, u(t_i), \hat{y}_{i+1}) \tag{9.3.72}$$

egyenlettel definiált $\hat{y}_{i+1}$ közelítő érték milyen közel van az $u(t_{i+1})$ értékhez.

9.3.23. definíció.
Az $l_i(h) = \hat{y}_{i+1} - u(t_{i+1})$ függvényt a (9.3.71) alakú Φ numerikus módszer *lokális diszkretizációs hibafüggvényének* nevezzük.

Vezessük be a

$$g_i(h) = u(t_i) + h\Phi(h, t_i, u(t_i), u(t_{i+1})) - u(t_{i+1}) \tag{9.3.73}$$

függvényt, amelynek rendje (a kiinduló Cauchy-feladat egyenletének felhasználásával) sorfejtéssel a pontos megoldás ismerete nélkül is meghatározható.

9.3.24. definíció.
A $g_i(h)$ függvényt a (9.3.71) alakú Φ numerikus módszer $t_i \in \omega_h$ pontbeli *képlethibájának* (más szóval, *lokális approximációs hibájának*) nevezzük. Azt mondjuk, hogy a Φ numerikus módszer *p-ed rendben konzisztens* a $t_i \in \omega_h$ rácspontban, ha

$$g_i(h) = \mathcal{O}(h^{p+1}) \tag{9.3.74}$$

valamely $p > 0$ állandóval.

A lokális approximációs hiba rendje tehát azt mutatja meg, hogy a pontos megoldás milyen pontossággal elégíti ki a Φ-módszer egyenletét.[7]

9.3.25. megjegyzés. A (9.3.37), (9.3.60) és a (9.3.66) becslések alapján látható, hogy az explicit és implicit Euler-módszerek elsőrendűek, míg a Crank–Nicolson-módszer másodrendű. Egyszerű számolással ellenőrizhető, hogy a θ-módszer csak $\theta = 0.5$ esetén (azaz amikor a Crank–Nicolson-módszert jelenti) másodrendű, egyébként elsőrendű. $\diamond$

A továbbiakban feltesszük, hogy a (9.3.71) alakú Φ numerikus módszerben a Φ függvény a harmadik és negyedik változójában egyaránt lipschitzes, azaz léteznek olyan $L_3 \geq 0$ és $L_4 \geq 0$ állandók, amelyek mellett tetszőleges s_1, s_2, p_1 és p_2 számok esetén

$$|\Phi(h, t_i, s_1, p_1) - \Phi(h, t_i, s_2, p_2)| \leq L_3|s_1 - s_2| + L_4|p_1 - p_2| \tag{9.3.75}$$

tetszőleges $t_i \in \omega_h$ és $h > 0$ esetén. Ha a Φ függvény nem függ y_i-től (avagy y_{i+1}-től), akkor $L_3 = 0$ (avagy $L_4 = 0$).

9.3.26. megjegyzés. A 9.3.21. megjegyzés alapján könnyen megmutatható, hogy az f függvény második változója szerinti lipschitzessége esetén tetszőleges θ esetén a θ-módszerre (és így az explicit és implicit Euler-módszerekre valamint a Crank–Nicolson-módszerre egyaránt) alkalmasan megválasztott L_3 és L_4 állandókkal érvényes a (9.3.75) egyenlőtlenség. $\diamond$

Az egylépéses módszerek pontbeli konvergenciája

A továbbiakban megvizsgáljuk a (9.3.75) tulajdonságú, r-ed rendben konzisztens Φ-módszerek konvergenciáját egy $t^\star \in (0, T]$ rögzített pontban. Legyen (ω_{h_v}) egy reguláris rácshálósorozat a $[0, t^\star]$ intervallumon, és mindegyik rácshálón n jelöli azt az indexet, amelyre $t_n = t^\star$.

[7] Szokásos a $g_i(h)/h$ függvényt képlethibának nevezni. Ekkor a képlethiba rendje és a módszer konvergenciájának rendje megegyezik.

9.3.27. definíció.
Az $e_n(h) = y_n - u(t^\star), (nh = t^\star)$ kifejezést a Φ numerikus módszer $t^\star$ pontbeli *globális approximációs hibájának* nevezzük.

9.3.28. definíció.
Azt mondjuk, hogy a Φ numerikus módszer *konvergens a $t^\star$ pontban*, ha

$$\lim_{h\to 0} e_n(h) = 0. \tag{9.3.76}$$

A (9.3.76) konvergenciájának rendjét a Φ-módszer *konvergenciarendjének* nevezzük.

Az egyszerűség kedvéért vezessük be az alábbi jelöléseket:

$$e_i(h) = e_i, \quad g_i(h) = g_i, \quad l_i(h) = l_i. \tag{9.3.77}$$

A fenti definíciók alapján fennáll az

$$e_{i+1} = y_{i+1} - u(t_{i+1}) = (y_{i+1} - \hat{y}_{i+1}) + (\hat{y}_{i+1} - u(t_{i+1})) = (y_{i+1} - \hat{y}_{i+1}) + l_i \tag{9.3.78}$$

összefüggés. Ezért a továbbiakban az

$$|e_{i+1}| \le |y_{i+1} - \hat{y}_{i+1}| + |l_i| \tag{9.3.79}$$

egyenlőtlenség jobb oldalán lévő két tagra adunk felső becslést.

A lokális diszkretizációs hibafüggvényre az alábbi összefüggés érvényes:

$$\begin{aligned} l_i &= \hat{y}_{i+1} - u(t_{i+1}) = u(t_i) + h\Phi(h, t_i, u(t_i), \hat{y}_{i+1})) - u(t_{i+1}) = -u(t_{i+1}) + u(t_i) \\ &+ h\Phi(h, t_i, u(t_i), u(t_{i+1})) + h\left[\Phi(h, t_i, u(t_i), \hat{y}_{i+1})) - \Phi(h, t_i, u(t_i), u(t_{i+1}))\right] \\ &= g_i + h\left[\Phi(h, t_i, u(t_i), \hat{y}_{i+1})) - \Phi(h, t_i, u(t_i), u(t_{i+1}))\right]. \end{aligned} \tag{9.3.80}$$

Ezért, felhasználva a (9.3.75) feltételt,

$$|l_i| \le |g_i| + hL_4|\hat{y}_{i+1} - u(t_{i+1})| = |g_i| + hL_4|l_i| \tag{9.3.81}$$

Így (9.3.81) alapján érvényes az

$$|l_i| \le \frac{1}{1 - hL_4}|g_i| \tag{9.3.82}$$

egyenlőtlenség.

9.3.29. megjegyzés. A (9.3.82) egyenlőtlenség azt is mutatja, hogy egy p-ed rendben konzisztens módszer esetén a lokális diszkretizációs hiba is (legalább $p+1$-ed rendben) tart nullához. ◇

Térjünk át a (9.3.79) jobb oldali első tagjának becslésére.

$$\begin{aligned} |y_{i+1} - \hat{y}_{i+1}| &= |(y_i + h\Phi(h, t_i, y_i, y_{i+1})) - (u(t_i) + h\Phi(h, t_i, u(t_i), \hat{y}_{i+1}))| \\ &\le |e_i| + h|\Phi(h, t_i, y_i, y_{i+1}) - \Phi(h, t_i, u(t_i), \hat{y}_{i+1})| \\ &\le |e_i| + hL_3|y_i - u(t_i)| + hL_4|y_{i+1} - \hat{y}_{i+1}| = (1 + hL_3)|e_i| + hL_4|y_{i+1} - \hat{y}_{i+1}|. \end{aligned} \tag{9.3.83}$$

Így (9.3.83) alapján érvényes az

$$|y_{i+1} - \hat{y}_{i+1}| \leq \frac{1 + hL_3}{1 - hL_4}|e_i| \tag{9.3.84}$$

egyenlőtlenség.

A (9.3.82) és a (9.3.84) felhasználásával a (9.3.79) egyenlőtlenség átírható az

$$|e_{i+1}| \leq \frac{1 + hL_3}{1 - hL_4}|e_i| + \frac{1}{1 - hL_4}|g_i| \tag{9.3.85}$$

alakra. Vezessük be a

$$\mu = \mu(h) = \frac{1 + hL_3}{1 - hL_4}, \quad \chi = \chi(h) = \frac{1}{1 - hL_4} \tag{9.3.86}$$

jelöléseket.

9.3.30. megjegyzés. A (9.3.86) jelölések mellett

$$\mu = 1 + h\frac{L_3 + L_4}{1 - hL_4} \tag{9.3.87}$$

és így $\mu = 1 + \mathcal{O}(h)$. Ezért választhatók olyan h_0, μ_0 és λ_0 állandók[8], amelyek mellett

$$\mu = \mu(h) \leq 1 + \mu_0 h, \quad \chi = \chi(h) \leq \chi_0, \quad \forall h \in (0, h_0). \tag{9.3.88}$$

◇

A (9.3.86) jelöléssel (9.3.85) felírható az

$$|e_{i+1}| \leq \mu|e_i| + \chi|g_i| \tag{9.3.89}$$

alakban. Rekurzív módon alkalmazva a (9.3.89) relációt, a következő egyenlőtlenséget nyerjük:

$$\begin{aligned} |e_n| &\leq \mu|e_{n-1}| + \chi|g_{n-1}| \leq \mu\left[\mu|e_{n-2}| + \chi|g_{n-2}|\right] + \chi|g_{n-1}| \\ &= \mu^2|e_{n-2}| + \chi\left[\mu|g_{n-2}| + |g_{n-1}|\right] \leq \ldots \leq \mu^n|e_0| + \chi\sum_{i=0}^{n-1}\mu^i|g_{n-1-i}| \\ &\leq \mu^n\left[|e_0| + \chi\sum_{i=0}^{n-1}|g_{n-1-i}|\right]. \end{aligned} \tag{9.3.90}$$

A (9.3.88) összefüggés alapján, minden $h \in (0, h_0)$ esetén $\chi \leq \chi_0$ és

$$\mu^n \leq (1 + \mu_0 h)^n \leq \exp(\mu_0 hn) = \exp(\mu_0 t^\star). \tag{9.3.91}$$

Ezért (9.3.90) alapján érvényes az

$$|e_n| \leq \exp(\mu_0 t^\star)\left[|e_0| + \chi_0\sum_{i=0}^{n-1}|g_{n-1-i}|\right] \tag{9.3.92}$$

[8] Például a $h_0 = \frac{1}{2L_4}, \mu_0 = 2(L_3 + L_4)$ és $\chi_0 = 2$ egy alkalmas megválasztás.

becslés. Mivel feltettük, hogy a vizsgált Φ-módszer r-ed rendben konzisztens, ezért a (9.3.76) definíció alapján, megfelelően kis h esetén valamely $c_0 \geq 0$ állandóval fennáll a $|g_i| \leq c_0 h^{r+1}$ egyenlőtlenség. Ezért kis h esetén

$$\sum_{i=0}^{n-1} |g_{n-1-i}| \leq n c_0 h^{r+1} = c_0 t^\star h^r. \tag{9.3.93}$$

Összevetve a (9.3.92) és a (9.3.93) formulákat, az

$$|e_n| \leq \exp(\mu_0 t^\star) \left[|e_0| + c_1 h^r \right] \tag{9.3.94}$$

becslést nyerjük, ahol $c_1 = \chi_0 c_0 t^\star$=állandó. Mivel $e_0 = 0$, ezért a (9.3.94) alapján beláttuk a következő állítást.

9.3.31. tétel.
Tegyük fel, hogy a (9.3.71) képlettel definiált Φ numerikus módszer

- p-ed rendben konzisztens, és
- a módszert definiáló Φ függvényre érvényes a (9.3.75) Lipschitz-feltétel.

Ekkor a Φ-módszer p-ed rendben konvergens a $[0, T]$ intervallumon.

9.3.32. következmény. A 9.3.25. és a 9.3.26. megjegyzések alapján a θ-módszer $\theta = 0.5$ esetén másodrendben, egyébként pedig elsőrendben konvergens. Ezért tehát az explicit és az implicit Euler-módszer elsőrendben, a Crank–Nicolson-módszer pedig másodrendben konvergens. $\diamond$

9.3.33. megjegyzés. Az előzőekben a konvergenciát a (9.3.89) összefüggésből vezettük le, mégpedig a benne szereplő tagok két tulajdonságából:

- a módszer konzisztens, azaz $g_i = \mathcal{O}(h^{p+1})$ valamely p pozitív számmal, emellett a $\chi(h)$ függvény korlátos;
- a $\mu(h)$ függvényre a (9.3.88) nagyságrendi becslés teljesül. Ez a tulajdonság azt fejezi ki, hogy az egyik időrétegről a következő időrétegre való áttérésnél a lépésszámok növelésével (azaz h csökkenésével) a hiba csak korlátosan növekedhet. Ezt a tulajdonságot a numerikus módszer stabilitásának nevezzük.

A 9.3.31.. tétel leegyszerűsítve tehát azt mutatja, hogy a korrekt kitűzésű Cauchy-feladatokra a Φ-módszer konzisztenciája és stabilitása a konvergenciát biztosítja. A stabilitást a Lipschitz-féle feltétellel tudjuk biztosítani. $\diamond$

9.4. A Runge–Kutta típusú módszerek

Kettőnél magasabb rendű numerikus módszerek megkonstruálása a (9.3.1)-(9.3.2) Cauchy-feladatra az előzőekben ismertetett egylépéses módszerek segítségével akadályokba ütközik: az egyszerűbb módszerek (explicit Euler-módszer, implicit Euler-módszer, Crank–Nicolson-módszer) legfeljebb másodrendűek, a Taylor-módszerek viszont egy meglehetősen bonyolult előzetes analízist (a parciális deriváltak meghatározását) és azok kiértékelését igénylik. Ebben a részben

megmutatjuk, hogy a parciális deriváltak kiszámításának feladata – egy viszonylag egyszerű ötlet segítségével – megkerülhető.[9]

9.4.1. A másodrendű Runge–Kutta típusú módszerek

Tekintsük ismét a (9.3.1)-(9.3.2) Cauchy-feladatot. Először a Runge–Kutta típusú módszerek bevezetéséhez határozzunk meg egy, a Crank–Nicolson-módszertől különböző, másodrendű, egylépéses módszert.
Írjuk ki az $u(t)$ megoldás (9.3.6) alakú Taylor-sorának első tagjait a $t = t^\star + h$ pontban. Mivel a másodrendű konzisztenciát szeretnénk biztosítani, ezért a következő alakot írjuk fel:

$$u(t^\star + h) = u(t^\star) + hu'(t^\star) + \frac{h^2}{2!}u''(t^\star) + \mathcal{O}(h^3). \tag{9.4.1}$$

Felhasználva a (9.3.4) deriváltakat, bevezetve az

$$f = f(t^\star, u(t^\star)), \quad \partial_i f = \partial_i f(t^\star, u(t^\star)), \quad \partial_{ij} f = \partial_{ij} f(t^\star, u(t^\star)), \quad \text{stb.}$$

egyszerűsítő jelöléseket, (9.4.1) átírható az

$$\begin{aligned} u(t^\star + h) =& u(t^\star) + hf + \frac{h^2}{2!}(\partial_1 f + f\partial_2 f) + \mathcal{O}(h^3) \\ =& u(t^\star) + \frac{h}{2}f + \frac{h}{2}[f + h\partial_1 f + hf\partial_2 f] + \mathcal{O}(h^3) \end{aligned} \tag{9.4.2}$$

alakra. Mivel[10]

$$f(t^\star + h, u(t^\star) + hf(t^\star, u(t^\star)) = f + h\partial_1 f + hf\partial_2 f + \mathcal{O}(h^2), \tag{9.4.3}$$

ezért (9.4.2) felírható az

$$u(t^\star + h) = u(t^\star) + \frac{h}{2}f + \frac{h}{2}(f(t^\star + h, u(t^\star) + hf(t^\star, u(t^\star))) + \mathcal{O}(h^3) \tag{9.4.4}$$

alakban. Tehát egy ω_h rácsháló tetszőleges $t_i = t^\star$ pontjában felírva a (9.4.4) egyenlőséget, definiálhatjuk az

$$y_{i+1} = y_i + \frac{h}{2}f(t_i, y_i) + \frac{h}{2}f(t_{i+1}, y_i + hf(t_i, y_i)) \tag{9.4.5}$$

egylépéses, explicit numerikus módszert. Vezessük be a

$$k_1 = f(t_i, y_i); \quad k_2 = f(t_{i+1}, y_i + hf(t_i, y_i)) = f(t_i + h, y_i + hk_1) \tag{9.4.6}$$

jelöléseket. Ekkor a (9.4.5) módszer felírható

$$y_{i+1} = y_i + \frac{h}{2}(k_1 + k_2) \tag{9.4.7}$$

alakban. A (9.4.6)-(9.4.7) módszert *Heun-módszernek* nevezzük.

[9] Az ötlet Carl David Tolmé Runge (1856 - 1927) német matematikustól és fizikustól, illetve Martin Wilhelm Kutta (1867 - 1944) német matematikustól származik.

[10] Emlékeztetünk, hogy a kétváltozós $f : Q_T \to \mathbb{R}$ függvény (t, u) pont körüli elsőfokú Taylor-sorba fejtése tetszőleges $c_1, c_2 \in \mathbb{R}$ esetén $f(t + c_1 h, u + c_2 h) = f(t, u) + c_1 h\partial_1 f(t, u) + c_2 h\partial_2 f(t, u) + \mathcal{O}(h^2)$ alakú.

9.4.1. megjegyzés. Mivel (9.4.4) alapján

$$u(t^\star + h) - u(t^\star) - \frac{h}{2} f - \frac{h}{2}(f(t^\star + h, u(t^\star) + hf(t^\star, u(t^\star))) = \mathcal{O}(h^3), \tag{9.4.8}$$

ezért a pontos megoldás $\mathcal{O}(h^3)$ rendben elégíti ki a (9.4.5) képletet, azaz *a Heun-módszer másodrendű.* ◇

9.4.2. megjegyzés.
A Heun-módszer néhány részlete megtalálható a
http://math.fullerton.edu/mathews/n2003/Heun%27sMethodMod.html
linken. Az ugyanitt lévő animáción a módszer viselkedése az $u' = 1 - t\sqrt[3]{u}$ differenciálegyenletre is jól látható. ◇

Megadhatók-e egyéb másodrendű módszerek? A (9.4.4) összefüggés általánosítása a következő paraméteres alak:

$$u(t^\star + h) = u(t^\star) + \sigma_1 hf(t^\star, u(t^\star)) + \sigma_2 hf(t^\star + a_2 h, u(t^\star) + b_{21} hf(t^\star, u(t^\star))) + \mathcal{O}(h^3), \tag{9.4.9}$$

ahol σ_1, σ_2, a_2 és b_{21} egyelőre tetszőleges paraméterek. Felírva a $t = t_i$ pontban a (9.4.9) egyenletet, az

$$y_{i+1} = y_i + \sigma_1 hf(t_i, y_i) + \sigma_2 hf(t_i + a_2 h, y_i + b_{21} hf(t_i, y_i)) \tag{9.4.10}$$

egylépéses numerikus módszert kapjuk.

9.4.3. megjegyzés. A (9.4.12) általános alakban felírt módszer paramétereit célszerű csoportosítva a következő alakban felírni:

$$\begin{array}{c|cc} 0 & & \\ a_2 & b_{21} & \\ \hline & \sigma_1 & \sigma_2 \end{array} \tag{9.4.11}$$

◇

Fejtsük Taylor-sorba a (9.4.9) egyenlet jobb oldalát! Ekkor az

$$\begin{aligned} u(t^\star + h) =& u(t^\star) + \sigma_1 hf + \sigma_2 h[f + a_2 h\partial_1 f + b_{21} hf\partial_2 f] + \mathcal{O}(h^3) \\ =& u(t^\star) + (\sigma_1 + \sigma_2)hf + h^2[a_2\sigma_2\partial_1 f + \sigma_2 b_{21} f\partial_2 f] + \mathcal{O}(h^3) \end{aligned} \tag{9.4.12}$$

egyenlőséget kapjuk. A módszerek rendjére vonatkozó (9.4.1.) megjegyzést alkalmazva, a (9.4.2) és a (9.4.12) képletek összevetéséből azt kapjuk, hogy a (9.4.10) által meghatározott numerikus módszer pontosan akkor másodrendű, amikor

$$\begin{aligned} \sigma_1 + \sigma_2 &= 1 \\ a_2\sigma_2 &= 0.5 \\ b_{21}\sigma_2 &= 0.5. \end{aligned} \tag{9.4.13}$$

A (9.4.10) képlet egyszerű átírásával eredményeinket az alábbi tételben összegezhetjük.

9.4.4. tétel.
Tegyük fel, hogy a σ_1, σ_2, a_2 és b_{21} paraméterek megoldásai a (9.4.13) egyenletnek. Ekkor a

$$k_1 = f(t_i, y_i), \quad k_2 = f(t_i + a_2 h, y_i + h b_{21} k_1), \tag{9.4.14}$$

$$y_{i+1} = y_i + h(\sigma_1 k_1 + \sigma_2 k_2) \tag{9.4.15}$$

képletekkel definiált egylépéses explicit numerikus módszer másodrendű.

A (9.4.13) feltételeket kielégítő (9.4.14)-(9.4.15) módszert *másodrendű Runge–Kutta típusú módszernek* nevezzük és az RK2 szimbólummal jelöljük.
Vizsgáljuk meg az RK2 módszereket meghatározó (9.4.13) egyenletrendszert! Mivel a négy ismeretlenre három egyenletünk van, ezért a megoldása nem egyértelmű. Könnyen látható, hogy teszőleges $\sigma \neq 0$ esetén (9.4.13) megoldásai a következő alakúak:

$$\sigma_2 = \sigma, \quad \sigma_1 = 1 - \sigma, \quad a_2 = b_{21} = 0.5\sigma. \tag{9.4.16}$$

Tehát az RK2 módszerek egy egyparaméteres módszercsaládot alkotnak, amelynek paramétereit a (9.4.11) táblázat alapján

$$\begin{array}{c|cc} 0 & & \\ 0.5\sigma & 0.5\sigma & \\ \hline & 1-\sigma & \sigma \end{array} \tag{9.4.17}$$

szerint kell megválasztani.

9.4.5. megjegyzés. A $\sigma = 0.5$ értékhez tartozó RK2 módszer éppen a Heun-módszert eredményezi. Érdekes megválasztás a $\sigma = 1$. Ekkor $\sigma_1 = 0$, $\sigma_2 = 1$ és $a_2 = b_{21} = 0.5$ és így a származtatott numerikus módszer

$$k_1 = f(t_i, y_i), \quad k_2 = f(t_i + 0.5h, y_i + 0.5hk_1), \quad y_{i+1} = y_i + hk_2. \tag{9.4.18}$$

A (9.4.18) másodrendű módszert *javított explicit Euler-módszernek* nevezzük.
Vegyük észre, hogy a Heun-módszer és a javított explicit Euler-módszer is bevezethető a korábbiakban már ismertetett módszerek módosításával. Nevezetesen,

- Ha a Crank–Nicolson-módszer (9.3.61)

  $$y_{i+1} - y_i = \frac{h_i}{2}\left[f(t_i, y_i) + f(t_{i+1}, y_{i+1})\right]$$

  képletében az $f(t_{i+1}, y_{i+1})$ implicit tagban y_{i+1} helyébe az $\tilde{y}_{i+1} = y_i + hf(t_i, y_i)$ explicit Euler-módszerrel kiszámított értéket helyezzük, akkor éppen a Heun-módszert kapjuk.

- A javított explicit Euler-módszer esetén a $t = t_i + 0.5h = t_{i+0.5}$ felezőpontban explicit Euler-módszerrel kiszámoljuk az $u(t)$ pontos érték $\tilde{y}_{i+0.5} = y_i + 0.5hf(t_i, y_i)$ közelítését, és ezzel az értékkel meghatározott irányban egy újabb explicit Euler-módszert írunk fel a teljes intervallumra.

Tehát a fenti Runge-Kutta módszerek paramétereinek (9.4.17) szerinti megadása a következő. A Heun-módszer esetén

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & 0.5 & 0.5 \end{array} \tag{9.4.19}$$

alakban, míg a javított explicit Euler-módszer esetén

$$\begin{array}{c|cc} 0 & & \\ 0.5 & 0.5 & \\ \hline & 0 & 1 \end{array} \tag{9.4.20}$$

alakban adható meg. ◇

9.4.6. megjegyzés. Felmerülhet a kérdés: lehetséges-e a σ tetszőleges paramétert úgy megválasztani, hogy az RK2 módszer nemcsak másodrendű, hanem harmadrendű legyen? A válasz nemleges, amit a következő példa bizonyít. Legyen a (9.3.1)-(9.3.2) Cauchy-feladatban $f(t,u)=u$. Ezért a (9.3.1) differenciálegyenlet megoldására $u'(t)=u(t)$, amelyet deriválva $u''(t)=u'(t)$. Ezért $u''(t)=u'(t)=u(t)$. Másrészt, az f függvény definíciója miatt $f(t_i,y_i)=y_i$. Ezért a (9.4.10) módszer erre a feladatra az

$$\begin{aligned} y_{i+1} &= y_i+\sigma_1 h y_i+\sigma_2 h(y_i+b_{21}hf(t_i,y_i)) = y_i+\sigma_1 h y_i+\sigma_2 h(y_i+b_{21}hy_i) \\ &= y_i+hy_i[\sigma_1+\sigma_2+h\sigma_2 b_{21}] = y_i[1+(\sigma_1+\sigma_2)h+\sigma_2 b_{21}h^2] \end{aligned} \tag{9.4.21}$$

alakot ölti. Behelyettesítve a másodrendűséghez szükséges (9.4.16) értékeket, az RK2 módszer erre a feladatra az

$$y_{i+1}=y_i(1+h+\frac{h^2}{2}) \tag{9.4.22}$$

algoritmust eredményezi, amely független a σ szabad paramétertől. Helyettesítsük be az $u(t)$ pontos megoldást a (9.4.22) képletbe, azaz számítsuk ki a lokális approximációs hibát! Ekkor

$$g_i=u(t_{i+1})-u(t_i)(1+h+\frac{h^2}{2}). \tag{9.4.23}$$

Az $u(t_{i+1})$ kifejezés $t=t_i$ pontbeli sorbafejtése a deriváltakra vonatkozó $u''(t_i)=u'(t_i)=u(t_i)$ egyenlőség következtében $u(t_{i+1})=u(t_i)(1+h+h^2/2)+\mathcal{O}(h^3)$. Ezért tehát $g_i=\mathcal{O}(h^3)$ σ tetszőleges megválasztása mellett, azaz minden RK2 módszer legfeljebb másodrendű erre a feladatra. ◇

9.4.2. A magasabb rendű Runge–Kutta típusú módszerek

A gyakorlati számítások során az első illetve másodrendű módszerek segítségével reális idő alatt gyakran nem tudjuk biztosítani a szükséges pontosságot. Ezért a továbbiakban célunk a (9.4.10) módszerből kiindulva *magasabb rendben pontos* formulák előállítása.
Tekintsük ismételten a (9.4.14)-(9.4.15) alakban felírt módszert. Megmutattuk, hogy ez a módszer legfeljebb másodrendben pontos. (Nevezetesen, ha paramétereire teljesülnek a (9.4.13) feltételek.) Ezért, ha harmadrendű módszert szeretnénk készíteni, újabb paraméterek bevezetése és azok megfelelő megválasztása szükséges. Kiindulva a módszer (9.4.14)-(9.4.15) alakjából, kézenfekvő a következő általánosítás:

$$\begin{aligned} k_1 &= f(t_i,y_i), \\ k_2 &= f(t_i+a_2h, y_i+hb_{21}k_1), \\ k_3 &= f(t_i+a_3h, y_i+hb_{31}k_1+hb_{32}k_2) \end{aligned} \tag{9.4.24}$$

és ezután az új érték

$$y_{i+1}=y_i+h(\sigma_1k_1+\sigma_2k_2+\sigma_3k_3). \tag{9.4.25}$$

Ennek a módszernek a paramétereit a (9.4.11) szerinti táblázat segítségével az

$$\begin{array}{c|ccc} 0 & & & \\ a_2 & b_{21} & & \\ a_3 & b_{31} & b_{32} & \\ \hline & \sigma_1 & \sigma_2 & \sigma_3 \end{array} \tag{9.4.26}$$

alakban írhatjuk fel.

A (9.4.24) módszer harmadrendűségéhez – a másodrendűséghez hasonlóan – a lokális approximációs hiba vizsgálata szükséges. Egy meglehetősen hosszú (de matematikailag nem nehéz) számolás után azt kapjuk, hogy a módszer paramétereire a következő feltételek kikötése szükséges:

$$\begin{gathered} a_2 = b_{21}, \quad a_3 = b_{31} + b_{32}, \\ a_3(a_3 - a_2) - b_{32}a_2(2 - 3a_2) = 0, \quad \sigma_3 b_{32} a_2 = 1/6, \quad \sigma_2 a_2 + \sigma_3 a_3 = 1/2, \\ \sigma_1 + \sigma_2 + \sigma_3 = 1. \end{gathered} \tag{9.4.27}$$

Ez hat egyenletet jelent a nyolc ismeretlenre. A lehetséges megoldások közül kettőt emelünk ki.

- A

$$\begin{array}{c|ccc} 0 & & & \\ 1/3 & 1/3 & & \\ 2/3 & 0 & 2/3 & \\ \hline & 1/4 & 0 & 3/4 \end{array} \tag{9.4.28}$$

  táblázatban szereplő értékekkel definiált módszer gyakran szerepel az alkalmazásokban.

- Szintén harmadrendű a

$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 1 & -1 & 2 & \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \tag{9.4.29}$$

  módszer. Mivel ez a módszer az $f(t,u) = f(t)$ esetén $\mathcal{O}(h^5)$ rendben pontos, ezért alkalmazása abban az esetben ajánlatos, amikor a $\partial_2 f$ parciális derivált közel van a nullához.

A még magasabb ($p > 3$) pontosság eléréséhez a módszer további általánosítása szükséges. Ehhez fogalmazzuk meg az eddigi módszereinket általános alakban. Természetes általánosítás a következő. Legyen $m \geq 1$ egy adott egész szám. Definiáljuk a következő, ún. *m-lépcsős explicit Runge-Kutta típusú módszert*:

$$\begin{aligned} k_1 &= f(t_i, y_i), \\ k_2 &= f(t_i + a_2 h, y_i + h b_{21} k_1), \\ k_3 &= f(t_i + a_3 h, y_i + h b_{31} k_1 + h b_{32} k_2), \\ &\vdots \\ k_m &= f(t_i + a_m h, y_i + h b_{m1} k_1 + h b_{m2} k_2 + \ldots + h b_{m,m-1} k_{m-1}) \end{aligned} \tag{9.4.30}$$

$$y_{i+1} = y_i + h(\sigma_1 k_1 + \sigma_2 k_2 + \ldots + \sigma_m k_m). \tag{9.4.31}$$

A képletekben szereplő paraméterek rögzítése jelenti a módszer megadását. A korábbiakhoz hasonlóan, ismét összefoglalhatjuk egy táblázatban ezeket a paramétereket.

$$\begin{array}{c|cccc} 0 & & & & \\ a_2 & b_{21} & & & \\ a_3 & b_{31} & b_{32} & & \\ \vdots & \vdots & \vdots & & \\ a_m & b_{m1} & b_{m2} & \dots & b_{m,m-1} \\ \hline & \sigma_1 & \sigma_2 & \dots & \sigma_m \end{array} \tag{9.4.32}$$

Ennek kompakt felírása céljából vezessünk be új jelöléseket! Jelölje a továbbiakban $\boldsymbol{\sigma}, \mathbf{a} \in \mathbb{R}^{\boldsymbol{m}}$ a σ_i és a_i elemeiből álló oszlopvektorokat (ahol mindig $a_1 = 0$), továbbá $\mathbf{B} \in R^{m\times m}$ a b_{ij} elemekből felépített szigorúan alsó háromszögmátrixot, azaz

$$\mathbf{B}_{ij} = \begin{cases} b_{ij}, & \text{ha } i > j, \\ 0, & \text{ha } i \le j. \end{cases}$$

9.4.7. definíció.
Egy explicit Runge-Kutta típusú módszer

$$\begin{array}{c|c} \mathbf{a} & \mathbf{B} \\ \hline & \boldsymbol{\sigma}^\top \end{array} \tag{9.4.33}$$

alakban felírt paramétereinek táblázatát *Butcher-táblázatnak* nevezzük.

(A felírási mód ötlete J. Butchertől[11] ered, és a konkrét módszerek felírásánál a $\mathbf{B}$ mátrix nem nulla elemeit soroljuk csak fel a táblázatban. Ugyanakkor, mint azt a továbbiakban lálni fogjuk, ez a felírási mód alkalmazható tetszőleges $\mathbf{B}$ mátrixok esetén is.)
Valamely explicit Runge-Kutta típusú módszer konzisztenciarendjét viszonylag hosszadalmas számolással határozhatjuk meg: a (9.4.30)-(9.4.31) képletekben y_i helyébe $u(t_i)$-t helyettesítünk, majd kiszámítjuk az így nyert kifejezés és az y_{i+1} helyébe írt $u(t_i + h)$ kifejezés $t = t_i$ pontban felírt, megfelelő rendű Taylor-polinomja különbségének h szerinti rendjét. Ezen számítások elvégzésével nyerjük az explicit Runge-Kutta típusú módszerek p-ed rendű konzisztenciájának feltételeit.

9.4.8. megjegyzés. A másodrendűség (9.4.17) illetve a harmadrendűség (9.4.27) első feltételéből látható, hogy a $\mathbf{B}$ mátrix mindegyik sorának összege megegyezik az ugyanabban a sorban szereplő a_i együtthatóval. Ezért az $\mathbf{a}$ vektort az $\mathbf{a} = \mathbf{Be}$ összefüggésből határozzuk meg, ahol $\mathbf{e} = [1, 1, \dots, 1]^\top \in \mathbb{R}^m$ oszlopvektort jelöli. $\diamond$

Bevezetve az

$$\mathbf{a}^n = (a_1^n, a_2^n, \dots, a_m^n)^\top \in \mathbb{R}^m, \quad \mathbf{A} = \operatorname{diag}(a_1, a_2, \dots, a_m)^\top \in \mathbb{R}^{m\times m}$$

jelöléseket, felírható az explicit Runge-Kutta típusú módszer p-ed rendű konzisztenciájának feltétele a $\mathbf{B}$ mátrix elemeire illetve a $\boldsymbol{\sigma}$ vektorra:

[11] John Charles Butcher (1933 –) ma is aktív új-zélandi matematikus.

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

9.4.1. táblázat: Negyedrendű, explicit Runge-Kutta típusú módszer Butcher-táblázata.

rend (p)	feltétel
1	$\boldsymbol{\sigma}^\top \cdot \mathbf{e} = 1$
2	$\boldsymbol{\sigma}^\top \cdot \mathbf{a} = 1/2$
3	$\boldsymbol{\sigma}^\top \cdot (\mathbf{a}^2) = 1/3 \quad \boldsymbol{\sigma}^\top \cdot \mathbf{Ba} = 1/6$
4	$\boldsymbol{\sigma}^\top \cdot (\mathbf{a}^3) = 1/4 \quad \boldsymbol{\sigma}^\top \cdot \mathbf{ABa} = 1/8$ $\boldsymbol{\sigma}^\top \cdot \mathbf{B}(\mathbf{a}^2) = 1/12 \quad \boldsymbol{\sigma}^\top \cdot \mathbf{B}^2\mathbf{a} = 1/24,$

(9.4.34)

ahol a feltételeket kummulatívan kell érteni, azaz pl. $p = 2$ rendhez szükséges a $p = 1$ rend feltétele is. Így érvényes az alábbi

9.4.9. tétel.

A (9.4.33) Butcher-táblázatú explicit Runge-Kutta típusú módszer pontosan akkor konzisztens, amikor teljesülnek a

$$\mathbf{Be} = \mathbf{a}; \quad \boldsymbol{\sigma}^\top \cdot \mathbf{e} = 1 \tag{9.4.35}$$

feltételek, azaz

$$\sum_{k=1}^{m} b_{ik} = a_i \text{ minden } i = 1, 2, \ldots, m \text{ esetén, és emellett } \sum_{k=1}^{m} \sigma_k = 1. \tag{9.4.36}$$

A magasabb rendű explicit Runge-Kutta típusú módszerek közül a leggyakrabban a 9.4.1. táblázatban megadott negyedrendű módszert szokásos alkalmazni.

A módszer algorimusa (azaz a t_{i+1} pontbeli y_{i+1} közelítés meghatározása a már kiszámolt t_i pontbeli y_i közelítésből) a következő:

- A

$$\begin{aligned} k_1 &= f(t_i, y_i) \\ k_2 &= f(t_i + 0.5h, y_i + 0.5hk_1) \\ k_3 &= f(t_i + 0.5h, y_i + 0.5hk_2) \\ k_4 &= f(t_i + h, y_i + hk_3) \end{aligned} \tag{9.4.37}$$

  képletekkel rendre kiszámoljuk a k_1, k_2, k_3 és k_4 értékeket.

- Az

$$y_{i+1} = y_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \tag{9.4.38}$$

  képlettel meghatározzuk az új közelítést.

9.4.10. megjegyzés. Mi a kapcsolat az explicit Runge-Kutta típusú módszerek lépcsőszáma (m) és rendje (p) között? Azt láttuk, hogy az egylépcsős explicit Euler-módszer elsőrendű, a kétlépcsős Heun-módszermódszer másodrendű, a háromlépcsős (9.4.28)-(9.4.29) Butcher-táblázatú módszerek harmadrendűek, a négylépcsős (9.4.37)-(9.4.38) módszer pedig negyedrendben pontos. Tehát $m = 1, 2, 3, 4$ esetén a maximálisan elérhető konzisztenciarend megegyezik a lépcsőszámmal. Ugyanakkor, $m \geq 5$ esetén ez már nem érvényes, a rend alatta marad a lépcsőszámnak, azaz $p < m$. A közöttük lévő kapcsolat az első tíz lépcsőszámú explicit Runge-Kutta típusú módszerre a következő:

m	$1, 2, 3, 4$	$5, 6, 7$	$8, 9, 10$
$p(m)$	m	$m-1$	$m-2$

(9.4.39)

Megjegyezzük, hogy m növelésével a p és m közötti hézag növekszik. ⋄

9.4.11. megjegyzés. Az explicit Runge-Kutta típusú módszer néhány további részlete megtalálható a
http://math.fullerton.edu/mathews/n2003/RungeKuttaMod.html
linken. Az ugyanitt lévő animáción a módszer viselkedése az $u' = 1 - t\sqrt[3]{u}$ differenciálegyenletre is jól látható. ⋄

9.4.12. megjegyzés. Ebben a részben az explicit Runge-Kutta típusú módszerek konzisztenciájával foglalkoztunk, és a gyakorlat szempontjából fontosabb *konvergenciát* nem tárgyaltuk. Megmutatható, hogy mindegyik explicit Runge-Kutta típusú módszer (az f függvény lipschitzessége mellett) ún. *zéró-stabil* is, és ez a tulajdonság a p-ed rendű konzisztenciával együtt a p-ed rendű konvergenciát is biztosítja. Mivel ezen kérdés részletes tárgyalása meghaladja a jegyzet kereteit, ezért a részletek iránt érdeklődöknek javasoljuk az irodalomjegyzékben szereplő [1, 34] irodalmakat. ⋄

9.4.3. Az implicit Runge-Kutta típusú módszerek

Vegyük észre, hogy általánosan a Runge–Kutta típusú módszerek a következő módon definiálhatók. Legyen $\mathbf{B} \in \mathbb{R}^{m\times m}$ egy tetszőleges mátrix, $\boldsymbol{\sigma} \in \mathbb{R}^m$ egy adott vektor, és $\mathbf{a} = \mathbf{Be}$. Egy Runge–Kutta típusú módszer Butcher-táblázatát ezen elemekkel definiáljuk. Az olyan módszereket, amelyekre $\mathbf{B}$ nem szigorúan alsó háromszögmátrix, *implicit Runge-Kutta típusú módszernek* (IRK) nevezzük. Amikor $\mathbf{B}$ alsó (de nem szigorúan alsó) háromszögmátrix, akkor a módszert *diagonálisan implicit Runge-Kutta típusú módszernek* (DIRK) nevezzük. A DIRK-módszer esetén k_i értékének kiszámolása, az explicit Runge-Kutta típusú módszertől eltérően, egy (általában nemlineáris) egyenlet megoldását, míg az implicit Runge-Kutta típusú módszer esetén egy m ismeretlenes (általában nemlineáris) egyenletrendszer megoldását igényli. Ez a módszer alkalmazását bonyolultabbá teszi. Az implicit Runge-Kutta típusú módszerek fontosak és az explicit Runge-Kutta típusú módszerrel összehasonlítva a gyakorlatban többször használatosak. Ennek okai a következők. Egyrészt ugyanazon lépcsőszám mellett magasabb rendű konzisztencia (és így konvergencia) is elérhető, biztosítható. Másrészt, az explicit Runge-Kutta típusú módszerektől eltérően, a magasabb rendben pontos módszerek is jó kvalitatív tulajdonságokkal rendelkeznek.[12] (Ennek viszont, mint azt már említettük, a nagyobb számítási munka az "ára".) A továbbiakban röviden tárgyalunk néhány nevezetes implicit Runge-Kutta típusú módszert, megadva a módszer

[12] Ezek a jobb kvalitatív tulajdonság egy része, mint pl. az A-stabilitás, a következő 9.4.4. szakaszban kerül ismertetésre.

Butcher-táblázatát. Mint látni fogjuk, a numerikus módszer általános egylépéses módszer alakjában való megadása (azaz a Φ függvény meghatározása) ebben az esetben már jóval összetettebb feladat, mint az explicit módszerek esetén.[13]

- Legyen a Butcher-táblázat

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \tag{9.4.40}$$

alakú. Ez egy egylépcsős diagonálisan implicit Runge-Kutta típusú módszer, amely részletesen kiírva a következőt jelenti:

$$\begin{aligned} k_1 &= f(t_i + h, y_i + hk_1) \\ y_{i+1} &= y_i + hk_1. \end{aligned} \tag{9.4.41}$$

Tehát algoritmikus realizása a következőt jelenti: első lépésben megoldjuk a k_1 ismeretlenre az első egyenletet[14], majd a megoldást behelyettesítjük a második képletbe. Határozzuk meg a módszer Φ függvényét! Mivel a második képletből $hk_1 = y_{i+1} - y_i$, ezt behelyettesítve az első egyenletbe $k_1 = f(t_i + h, y_i + (y_{i+1} - y_i)) = f(t_i + h, y_{i+1})$. Ezért, ismét a második összefüggésből, $y_{i+1} = y_i + hf(t_i + h, y_{i+1})$, azaz $\Phi(h, t_i, y_i, y_{i+1}) = f(t_i + h, y_{i+1})$. Így a (9.4.40) Butcher-táblázatú módszer az implicit Euler-módszert jelenti. Így ez a módszer elsőrendű.

- Legyen a Butcher-táblázat

$$\begin{array}{c|c} 0.5 & 0.5 \\ \hline & 1 \end{array} \tag{9.4.42}$$

alakú. Ez szintén egy egylépcsős diagonálisan implicit Runge-Kutta típusú módszer, amely a következőt jelenti:

$$\begin{aligned} k_1 &= f(t_i + 0.5h, y_i + 0.5hk_1) \\ y_{i+1} &= y_i + hk_1. \end{aligned} \tag{9.4.43}$$

Határozzuk meg ennek a módszernek is a Φ függvényét! A második képletből $hk_1 = y_{i+1} - y_i$. Behelyettesítve az első egyenletbe $k_1 = f(t_i + 0.5h, y_i + 0.5(y_{i+1} - y_i)) = f(t_i + 0.5h, 0.5(y_i + y_{i+1}))$. Tehát $\Phi(h, t_i, y_i, y_{i+1}) = f(t_i + 0.5h, 0.5(y_i + y_{i+1}))$ és a megfelelő egylépéses módszer

$$y_{i+1} = y_i + hf(t_i + 0.5h, 0.5(y_i + y_{i+1})) \tag{9.4.44}$$

alakú. A (9.4.44) diagonálisan implicit Runge-Kutta típusú módszert *implicit középponti szabálynak* nevezzük. A módszer rendjét a

$$g_i = u(t_{i+1}) - u(t_i) - hf(t_i + 0.5h, 0.5(u(t_i) + u(t_{i+1})))$$

kifejezés nagyságrendjének meghatározásával nyerjük. A szokásos sorbafejtéssel a következőket kapjuk:

$$u(t_{i+1}) - u(t_i) = hu'(t_i) + \frac{h^2}{2}u''(t_i) + \mathcal{O}(h^3),$$

$$\begin{aligned} f(t_i + 0.5h, 0.5(u(t_i) + u(t_{i+1}))) &= f(t_i + 0.5h, u(t_i) + 0.5hu'(t_i) + \mathcal{O}(h^2)) \\ &= f(t_i, u(t_i)) + 0.5h\partial_1 f(t_i, u(t_i)) + 0.5hu'(t_i)\partial_2 f(t_i, u(t_i)) + \mathcal{O}(h^2) \\ &= f(t_i, u(t_i)) + \frac{h}{2}[\partial_1 f(t_i, u(t_i)) + f(t_i, u(t_i))\partial_2 f(t_i, u(t_i))] + \mathcal{O}(h^2). \end{aligned}$$

[13] Az explicit és implicit módszerek összehasonlítására, és az utóbbi előnyeire a 9.5.2. szakaszban még visszatérünk.

[14] A megoldáshoz valamely, már korábban ismertetett iterációs módszert (tipikusan a Newton-féle iterációt) alkalmazzuk.

Behelyettesítve ezeket az értékeket g_i kifejezésébe, és figyelembe véve a (9.3.4) második összefüggését, a $g_i = \mathcal{O}(h^3)$ nagyságrendet kapjuk. Tehát az implicit középponti szabály másodrendű.

- Tekintsük a következő Butcher-táblázatot!

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0.5 & 0.5 \\ \hline & 0.5 & 0.5 \end{array} \tag{9.4.45}$$

Ez egy kétlépcsős implicit Runge-Kutta típusú módszer, amely a következőt jelenti:

$$\begin{aligned} k_1 &= f(t_i, y_i) \\ k_2 &= f(t_i + h, y_i + 0.5k_1 + 0.5k_2) \\ y_{i+1} &= y_i + 0.5hk_1 + 0.5hk_2. \end{aligned} \tag{9.4.46}$$

A harmadik képletből $0.5h(k_1 + k_2) = y_{i+1} - y_i$. Ezt és az első összefüggést behelyettesítve a második egyenletbe, a $k_2 = f(t_i + h, y_i + (y_{i+1} - y_i)) = f(t_i + h, y_{i+1})$ összefüggést kapjuk. Tehát ezt a k_2 értéket, illetve az első összefüggésbeli k_1 értéket behelyettesítve a harmadik egyenletbe, $\Phi(h, t_i, y_i, y_{i+1}) = 0.5[f(t_i, y_i) + f(t_i + h, y_{i+1})]$. Ezért a (9.4.45) Butcher-táblázat a Crank–Nicolson-módszert jelenti.

A további módszereknek csak a Butcher-táblázatát ismertetjük.

- A Butcher-táblázatbeli **a** vektor koordinátáinak a numerikus integrálásnál már ismertetett Gauss-féle alappontokat választjuk:

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & 0.5 & 0.5 \end{array} \tag{9.4.47}$$

A (9.4.47) módszer egy kétlépcsős, implicit Runge-Kutta típusú módszer, amely negyedrendű.

- Tekintsük a következő, szintén kétlépcsős módszert:

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array} \tag{9.4.48}$$

A (9.4.48) módszert *kétlépcsős Radau-módszernek* nevezzük, amely egy harmadrendű implicit Runge-Kutta típusú módszer.

A fenti implicit Runge-Kutta típusú módszerek pontosságának vizsgálata során azt látjuk, hogy a módszerek pontossága (p) meghaladhatja a lépcsőszámot (m). Például az egylépcsős trapézszabály másodrendű, a kétlépcsős Gauss-féle alappontokkal rendelkező (9.4.47) módszer negyedrendű, a kétlépcsős Radau-módszer harmadrendű stb. Tehát az explicit Runge-Kutta típusú módszertől eltérően, a $p > m$ eset is lehetséges. (Ennek oka nyilván az, hogy az implicit Runge-Kutta típusú módszernél több szabadon megválasztható paraméter áll a rendelkezésünkre.) Megmutatható, hogy adott lépcsőszám esetén $p \leq 2m$.

9.4.13. megjegyzés. Azokat a módszereket, amelyekre $p = 2m$, *maximális pontosságú implicit Runge-Kutta típusú módszereknek* nevezzük. Ezért az implicit Euler-módszer, az implicit középponti szabály és a (9.4.47) Gauss-alappontos módszer egyaránt maximális pontosságú. ⋄

9.4.4. Az egylépéses módszerek egy tesztfeladaton

Az eddigiekben több numerikus módszert ismertettünk, amelyek a pontosságukban (rendjükben) illetve a megoldás kiszámításának módjában (explicit/implicit) különböztek egymástól. Ebben a részben egy modellfeladatra alkalmazva a módszereinket újabb tulajdonságokat állapítunk meg. Tekintsük a $t \in [0, \infty)$ intervallumon az

$$u' = \lambda u, \quad u(0) = 1 \tag{9.4.49}$$

ún. *tesztfeladatot* valamely rögzített $\lambda < 0$ szám esetén. (A (9.4.49) feladat pontos megoldása $u(t) = \exp(\lambda t)$.) Oldjuk meg néhány λ értékre a feladatot numerikusan az explicit és implicit Euler-módszerekkel, és számítsuk ki a hibát a $t = 1$ pontban! Eredményeinket a 9.4.2. táblázat tartalmazza.

	$\lambda = -9$		$\lambda = -99$		$\lambda = -999$	
h	EE	IE	EE	IE	EE	IE
0.1	$3.07e-01$	$1.20e-01$	$3.12e+09$	$9.17e-02$	$8.95e+19$	$9.93e-03$
0.01	$1.72e-02$	$1.60e-02$	$3.62e-01$	$1.31e-01$	$2.38e+95$	$9.09e-02$
0.001	$1.71e-03$	$1.60e-03$	$1.90e-02$	$1.75e-02$	$3.67e-01$	$1.32e-01$
0.0001	$1.66e-04$	$1.65e-04$	$1.78e-03$	$1.68e-03$	$1.92e-02$	$1.76e-02$
0.00001	$1.66e-05$	$1.66e-05$	$1.82e-04$	$1.18e-04$	$1.83e-03$	$1.83e-03$

9.4.2. táblázat: A különböző λ értékekkel kitűzött tesztfeladat numerikus megoldásának hibája a $t = 1$ pontban.

Figyeljük meg, hogy $\lambda = -9$ mellett mindkét módszer az elméletnek megfelelő módon viselkedik (azaz mindegyik h érték mellett a hiba elsőrendű). Ugyanakkor a $\lambda = -99$ és a $\lambda = -999$ esetekben ez már nem így van: a kezdeti h megválasztások mellett az explicit Euler-módszer nem közelíti a megoldást, az implicit Euler-módszer viszont jó közelítést ad. A h paraméter további csökkentésével viszont már mindkét módszer az elméletnek megfelelően viselkedik. Mindebből arra következtethetünk, hogy ezekre a feladatokra az implicit Euler-módszer h megválasztásától függetlenül mindig jól viselkedik, az explicit Euler-módszer viszont csak valamely $h_0 > 0$ melletti $h < h_0$ feltétel mellett alkalmazható. Ez utóbbi viszont problémát jelent: ha a negatív λ értékét tovább csökkentjük, akkor már csak olyan kis h_0 értékek mellett működik az explicit Euler-módszer, amelynél

1. h_0 közel van a legkisebb ábrázolható pozitív számhoz, ezért a számítógépes realizálás eleve nem lehetséges,

2. ha h_0 nagyobb ugyan a legkisebb ábrázolható pozitív számnál, de még mindig nagyon kicsi, akkor egy rögzített $t^\star$ időrétegen való $y_{n_{t^\star}}$ közelítés meghatározásához rendkívül sok lépés ($n_{t^\star} \approx t^\star/h_0$) végrehajtása szükséges. Ez egyrészt a számítások idejének megnövekedését, másrészt pedig a számítások során fellépő hibák felhalmozódásának a lehetőségét eredményezi.

A jelenség oka a következő. A (9.4.49) tesztfeladat megoldása az explicit Euler-módszerrel az

$$y_{i+1} = (1 + h\lambda)y_i, \quad i = 0, 1, \ldots, \quad y_0 = 1,$$

az implicit Euler-módszerrel pedig az

$$y_{i+1} = \frac{1}{1 - h\lambda} y_i, \quad i = 0, 1, \ldots, \quad y_0 = 1,$$

egylépéses iterációkat jelenti. Egységesen felírva, az egylépéses módszereknek a tesztfeladatra történő alkalmazása egy

$$y_{i+1} = R(h\lambda)y_i \tag{9.4.50}$$

iterációt jelent, ahol $R(h\lambda)$ az alkalmazott numerikus módszer által meghatározott ún. stabilitási függvénye. Mivel a tesztfeladat $u(t) = \exp(\lambda t)$ pontos megoldására

$$u(t_{i+1}) = \exp(h\lambda)u(t_i), \tag{9.4.51}$$

ezért az alkalmazott módszert az jellemzi, hogy (a $z = h\lambda$ jelölés bevezetésével) az $R(z)$ függvény milyen jól közelíti az $\exp(z)$ függvényt.

9.4.14. megjegyzés. Vegyük észre, hogy az $\exp(z) - R(z)$ eltérés a numerikus módszer képlethibáját jellemzi a tesztfeladaton! Ugyanis a képlethiba (9.3.73) definíciója, a (9.4.50) és a (9.4.51) összefüggések alapján

$$g_i(h) = u(t_{i+1}) - R(h\lambda)u(t_i) = (\exp(h\lambda) - R(h\lambda))u(t_i).$$

Mivel az $u(t)$ megoldás korlátos, ezért a konzisztencia rendje meghatározható az $(\exp(h\lambda) - R(h\lambda))$ rendjével, azaz p-ed rendű konzisztencia esetén $\exp(h\lambda) - R(h\lambda) = \mathcal{O}(h^{p+1})$. Mivel az explicit Euler-módszer esetén $R_{EE}(z) = 1 + z$, az implicit Euler-módszer esetén $R_{IE}(z) = 1/(1 - z)$, a Crank–Nicolson-módszer esetén pedig $R_{CN}(z) = (1 + z/2)/(1 - z/2)$, ezért könnyen ellenőrizhetően $\exp(z) - R_{EE}(z) = \mathcal{O}(h^2)$, $\exp(z) - R_{IE}(z) = \mathcal{O}(h^2)$, $\exp(z) - R_{CN}(z) = \mathcal{O}(h^3)$, amely természetesen összhangban áll az ezen módszerek rendjére vonatkozó korábbi megállapításainkkal. ◇

A tesztfeladatra tetszőleges $\lambda < 0$ esetén a pontos megoldás monoton csökkenő, és korlátos. Ezért nyilvánvalóan csak azok a numerikus megoldások tudják a pontos megoldást jól approximálni, amelyekre az előállított numerikus megoldás is rendelkezik ezekkel a tulajdonságokkal, , azaz a (9.4.50) módszerben az

$$|R(h\lambda)| \leq 1 \tag{9.4.52}$$

feltétel teljesül. Mivel

$$|R_{EE}(h\lambda)| \leq 1 \quad \Leftrightarrow h \leq 2/(-\lambda), \tag{9.4.53}$$

ezért az explicit Euler-módszerre a már említett $h \leq h_0$ feltételben $h_0 = 2/(-\lambda)$. Ugyanakkor az implicit Euler-módszer esetén

$$|R_{IE}(h\lambda)| \leq 1 \quad \text{minden } h > 0 \tag{9.4.54}$$

esetén teljesül. A (9.4.53) és a (9.4.54) összefüggések magyarázatot adnak arra, hogy a 9.4.2. táblázatban miért viselkednek ennyire eltérően az explicit és implicit Euler-módszerek bizonyos h értékek esetén.

Fontos megjegyeznünk, hogy valamely numerikus módszer konvergenciája a módszer $h \to 0$ melletti viselkedését jellemzi, azaz csak azt garantálja, hogy *elegendően kis h esetén* a numerikus megoldás közel kerül a pontos megoldáshoz. Ugyanakkor nem ad információt a megoldásról

valamely *rögzített rácshálón*. Ezért kiemelkedően fontosak azok az $R(z)$ stabilitási függvénnyel rendelkező numerikus módszerek, amelyek tetszőleges rögzített rácshálón is jól követik a pontos megoldást. Ezt a tulajdonságot a (9.4.49) tesztfeladat numerikus megoldásán ellenőrizzük, ahol megengedjük λ komplex értékét is.[15]

9.4.15. definíció.
Azon $z \in \mathbb{C}$ komplex számok halmazát, amelyekre az

$$|R(z)| \leq 1 \tag{9.4.55}$$

feltétel teljesül, a *numerikus módszer stabilitási tartományának* nevezzük. Azt mondjuk, hogy egy numerikus módszer *A-stabil*, ha a stabilitási tartománya tartalmazza a $\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Re}(z) < 0\} \subset \mathbb{C}$ komplex félsíkot.

(Itt lényegében a $(0, \infty)$ intervallumon való korlátosságot ellenőrizzük.) Az A-stabilitás tehát azt jelenti, hogy minden olyan $z = a + ib$ komplex számra, amelyre $a < 0$, érvényes az $|R(z)| \leq 1$ egyenlőtlenség. Könnyen megmutatható, hogy az implicit Euler-módszer A-stabil. (Az explicit Euler-módszer nyilván nem, hiszen, mint láttuk, (9.4.55) már az $\mathbb{R}^- \subset \mathbb{C}^-$ halmazon sem igaz.)

9.4.16. megjegyzés. Tekintsük a Crank–Nicolson-módszert, amely felírható

$$y_{i+1} = R_{CN}(h\lambda)y_i = \frac{1 + \lambda h/2}{1 - \lambda h/2} y_i$$

alakban. Mint megmutattuk, a módszer másodrendű, és könnyen láthatón tetszőleges $h > 0$ esetén minden valós $\lambda < 0$ esetén $|R_{CN}(h\lambda)| \leq 1$. [16] Ugyanakkor a tesztfeladaton tetszőleges $h > 0$ mellett mégsem viselkedik jól a módszer. Ugyanis $h > 2/(-\lambda)$ esetén $R_{CN}(h\lambda) \in (-1, 0)$, ezért az ilyen rácshálókon az y_i értékei lépésenként előjelet váltanak, azaz bár abszolút értékben csökkenek, de emellett oszcillálnak is, ami ellentmond a pontos megoldás szigorún monoton csökkenésének. $\diamond$

9.5. A többlépéses módszerek

Az eddigiekben az egylépéses módszereket vizsgáltuk, vagyis az olyan numerikus módszereket, amelyekkel a megoldásfüggvény valamely rácshálópontbeli közelítését az ezen pontot megelőző rácshálópontbeli közelítésének segítségével határozzuk meg. A továbbiakban ezt általánosítjuk: egy adott pontbeli közelítést m darab ($m \geq 1$) megelőző pontbeli érték segítségével határozzuk meg. Az ilyen módszereket *m-lépéses módszereknek* nevezzük. (A korábbi egylépéses módszereink a többlépéses módszerek speciális esetének tekinthetők az $m = 1$ megválasztással.)

Az alábbiakban két egyszerű példán bemutatjuk, hogy a (9.3.1)-(9.3.2) Cauchy-feladat megoldásának a megfelelő pontok körüli Taylor-sorba fejtésével hogyan származtathatók ilyen típusú módszerek.

[15] Ennek oka, hogy az $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{A} \in \mathbb{R}^{n \times n}$ lineáris rendszer Cauchy-feladata átírható n darab, a (9.4.49) alakú tesztfeladatra, ahol az egyes λ számok az $\mathbf{A}$ mátrix sajátértékei, amelyek ezért tehát komplex számok is lehetnek. Lásd a 9.6. szakaszt, illetve bővebb ismeretekért a jegyzékben található [1] hivatkozást.

[16] Megmutatható, hogy a módszer $R_{CN}(z) = \frac{1+z/2}{1-z/2}$ stabilitási függvényére $\mathrm{Re}(z) < 0$ esetén $|R_{CN}(z)| \leq 1$, azaz a módszer A-stabil is.

9.5.1. példa. Nyilván

$$\begin{aligned} u(t_{i-1}) =& u(t_i) - hu'(t_i) + \frac{h^2}{2}u''(t_i) + \mathcal{O}(h^3), \\ u(t_{i-2}) =& u(t_i) - 2hu'(t_i) + \frac{4h^2}{2}u''(t_i) + \mathcal{O}(h^3). \end{aligned} \tag{9.5.1}$$

Ezért

$$3u(t_i) - 4u(t_{i-1}) + u(t_{i-2}) = 2hu'(t_i) + \mathcal{O}(h^3) = 2hf(t_i, u(t_i)) + \mathcal{O}(h^3).$$

Így az $f_i = f(t_i, y_i)$ jelöléssel definiálható az

$$y_i - \frac{4}{3}y_{i-1} + \frac{1}{3}y_{i-2} = \frac{2}{3}hf_i, \quad i = 2, 3, \ldots \tag{9.5.2}$$

módszer. Láthatóan a (9.5.2) egy kétlépéses implicit módszer, amely másodrendben konzisztens.

◇

9.5.2. példa. Először fejtsük Taylor-sorba a megoldásfüggvényt, majd annak deriváltfüggvényét a t_{i-1} pont körül. Ekkor

$$\begin{aligned} u(t_i) =& u(t_{i-1}) + hu'(t_{i-1}) + \frac{h^2}{2}u''(t_{i-1}) + \mathcal{O}(h^3), \\ u'(t_{i-2}) =& u'(t_{i-1}) - hu''(t_{i-1}) + \mathcal{O}(h^2). \end{aligned} \tag{9.5.3}$$

Mivel a második összefüggésből $hu''(t_{i-1}) = u'(t_{i-1}) - u'(t_{i-2}) + \mathcal{O}(h^2)$, ezt behelyettesítve az első egyenletbe az

$$u(t_i) = u(t_{i-1}) + \frac{h}{2}[3u'(t_{i-1}) - u'(t_{i-2})] + \mathcal{O}(h^3)$$

összefüggést nyerjük. Ez alapján

$$y_i - y_{i-1} = h\left[\frac{3}{2}f_{i-1} - \frac{1}{2}f_{i-2}\right], \quad i = 2, 3, \ldots \tag{9.5.4}$$

egy kétlépéses explicit módszer, amely másodrendben konzisztens. ◇

9.5.1. A lineáris többlépéses módszer általános alakja és rendje

A fenti példáink alapján az m-lépéses módszerek általános alakban igy definiálhatók.

9.5.3. definíció.

Az adott $a_0, a_1, \ldots, a_m$ és $b_0, b_1, \ldots, b_m$ együtthatók melletti

$$a_0y_i + a_1y_{i-1} + \cdots + a_my_{i-m} = h[b_0f_i + b_1f_{i-1} + \ldots + b_mf_{i-m}], \quad i = m, m+1, \ldots, \tag{9.5.5}$$

iterációt *lineáris, m-lépéses módszernek* nevezzük.

A továbbiakban mindig feltesszük, hogy $a_0 \neq 0$, hiszen csak ebben az esetben lehetséges az ismert $y_{i-m}, y_{i-m+1}, \ldots, y_{i-1}$ értékekből y_i értékét meghatározni. Mivel $f_i = f(t_i, y_i)$, ezért a (9.5.5) módszer *explicit*, ha $b_0 = 0$, és *implicit*, amikor $b_0 \neq 0$. Egy lineáris többlépéses módszer definiálása az a_k és b_k paraméterek ($k = 0, 1, \ldots, m$) konkrét értékeinek a megadásával történik. Például, a (9.5.2) numerikus módszer esetén $m = 2$, $a_0 = 1$, $a_1 = -4/3$, $a_2 = 1/3$, és $b_0 = 2/3$, $b_1 = 0$, $b_2 = 0$. Ha két lineáris többlépéses módszer együtthatói (a_k, b_k) és $(a_k^\star, b_k^\star)$, és létezik olyan $\beta \neq 0$ állandó, amely mellett $a_k = \beta a_k^\star$ valamint $b_k = \beta b_k^\star$ minden $k = 0, 1, \ldots, m$ értékre, akkor adott kezdeti értékekből mindkét módszer ugyanazt az eredményt szolgáltatja. Ezért az ilyen módszereket nem különböztetjük meg egymástól. Tehát a lineáris többlépéses módszereket leíró $2(m+1)$ paraméter közül egyet előre rögzítenünk szükséges. Ez, konvenció szerint, az a_0 paraméter, és a továbbiakban mindig feltesszük, hogy

$$a_0 = 1. \tag{9.5.6}$$

Vizsgáljuk meg a (9.5.5) általános alakú lineáris többlépéses módszer konzisztenciájának feltételét, illetve lehetséges konzisztenciarendjét. A módszer lokális approximációs hibája

$$g_i(h) = \sum_{k=0}^{m} [a_k u(t_{i-k}) - h b_k f(t_{i-k}, u(t_{i-k}))]. \tag{9.5.7}$$

Mivel a (9.3.1) alapján $u'(t_{i-k}) = f(t_{i-k}, u(t_{i-k}))$, ezért

$$g_i(h) = \sum_{k=0}^{m} [a_k u(t_{i-k}) - h b_k u'(t_{i-k})]. \tag{9.5.8}$$

Fejtsük a $t = t_i$ pont körül Taylor-sorba p-ed illetve $p-1$-ed rendig a (9.5.8) jobb oldalán szereplő függvényeket! Ekkor

$$u(t_{i-k}) = u(t_i - kh) = u(t_i) - khu'(t_i) + \frac{1}{2!}k^2h^2u''(t_i) + \ldots + (-1)^p \frac{1}{p!}k^p h^p u^{(p)}(t_i) + \mathcal{O}(h^{p+1}),$$

$$u'(t_{i-k}) = u'(t_i - kh) = u'(t_i) - khu''(t_i) + \ldots + (-1)^{p-1} \frac{1}{(p-1)!}k^{p-1}h^{p-1}u^{(p)}(t_i) + \mathcal{O}(h^p).$$

Ezt behelyettesítve a (9.5.8) összefüggésbe, a lokális approximációs hibára a

$$g_i(h) = d_0 u(t_i) + h d_1 u'(t_i) + h^2 d_2 u''(t_i) + \ldots + h^p d_p u^p(t_i) + \mathcal{O}(h^{p+1}) \tag{9.5.9}$$

kifejezést kapjuk, ahol

$$\begin{aligned}
d_0 &= \sum_{k=0}^{m} a_k, \\
d_1 &= -\sum_{k=0}^{m} (k a_k + b_k) \\
d_2 &= \sum_{k=0}^{m} (\frac{1}{2}k^2 a_k + k b_k), \\
&\vdots \\
d_p &= (-1)^p \sum_{k=0}^{m} (\frac{1}{p!}k^p a_k + \frac{1}{(p-1)!}k^{p-1} b_k).
\end{aligned} \tag{9.5.10}$$

Egy lineáris többlépéses módszer pontosan akkor p-ed rendű, amikor $g_i(h) = \mathcal{O}(h^{p+1})$, azaz teljesülnek a $d_0 = d_1 = \ldots = d_p = 0$ feltételek. Ez a (9.5.10) alapján a következőt jelenti.

9.5.4. tétel.
A (9.5.5) lineáris többlépéses módszer p-ed rendű, ha a módszert definiáló paraméterekre teljesülnek az alábbi feltételek:

$$\begin{aligned} a_0 = 1, \quad \sum_{k=0}^{m} a_k = 0 \\ \frac{1}{j}\sum_{k=0}^{m} k^j a_k + \sum_{k=0}^{m} k^{j-1} b_k = 0, \quad j = 1, 2, \ldots, p. \end{aligned} \tag{9.5.11}$$

Ezek alapján közvetlenül megfogalmazható a konzisztencia feltétele is.

9.5.5. következmény. A (9.5.5) lineáris többlépéses módszer pontosan akkor konzisztens, amikor a módszert definiáló paraméterekre teljesülnek az

$$\begin{aligned} a_0 = 1, \quad \sum_{k=0}^{m} a_k = 0 \\ \sum_{k=0}^{m} k a_k + \sum_{k=0}^{m} b_k = 0 \end{aligned} \tag{9.5.12}$$

feltételek. ◇

9.5.6. megjegyzés. A (9.5.12) feltétel kiírva a következőt jelenti: a módszer akkor konzisztens, amikor

$$\begin{aligned} 1 + a_1 + \ldots + a_m = 0 \\ (a_1 + 2a_2 + \ldots + m a_m) + (b_0 + b_1 + \ldots + b_m) = 0. \end{aligned} \tag{9.5.13}$$

Továbbá $p \geq 2$ rendben pontos, ha (9.5.13) mellett teljesülnek az

$$\frac{1}{j}\sum_{k=1}^{m} k^j a_k + \sum_{k=1}^{m} k^{j-1} b_k = 0, \quad j = 2, 3, \ldots, p \tag{9.5.14}$$

feltételek. Ez például azt jelenti, hogy egy m-lépéses módszer másodrendűségéhez a konzisztencia mellett az

$$\frac{1}{2}\sum_{k=1}^{m} k^2 a_k + \sum_{k=1}^{m} k b_k = 0 \tag{9.5.15}$$

feltétel teljesülése szükséges. Könnyen ellenőrizhető, hogy a (9.5.2) numerikus módszer esetén (ahol $m = 2$, $a_0 = 1$, $a_1 = -4/3$, $a_2 = 1/3$, és $b_0 = 2/3$, $b_1 = 0$, $b_2 = 0$) ezek az összefüggések érvényesek. ◇

Milyen pontosságú lehet egy (9.5.5) alakú lineáris többlépéses módszer? Az általános alakban $2m + 1$ paraméter ($a_1, a_2, \ldots, a_m$ és $b_0, b_1, \ldots, b_m$) választható meg. Ugyanakkor ezeknek a paramétereknek a p-ed rendű pontossághoz $p + 1$ feltételt kell teljesíteniük. Így $p \leq 2m$. Ha a módszer explicit, azaz $b_0 = 0$, akkor eggyel kevesebb a szabadon megválasztható paraméterek száma. Összefoglalóan, érvényes a következő állítás.

9.5.7. tétel.
Egy m-lépéses implicit lineáris többlépéses módszer maximális rendje $2m$, az explicit lineáris többlépéses módszeré pedig $2m-1$.

9.5.8. megjegyzés. A lineáris többlépéses módszer egyértelműségét biztosító $a_0 = 1$ feltétel helyett megadható más feltétel is. Gyakori a

$$\sum_{k=0}^{m} b_k = 1 \tag{9.5.16}$$

feltétel megadása. Ez azt biztosítja, hogy (9.5.5)

$$\frac{a_0 y_i + a_1 y_{i-1} + \cdots + a_m y_{i-m}}{h} = b_0 f_i + b_1 f_{i-1} + \ldots + b_m f_{i-m}$$

alakjában a jobb oldalon szereplő kifejezés bármely konstans értékű f függvényt pontosan approximálja. A (9.5.12) alakból könnyen látható, hogy a (9.5.16) feltétel mellett a konzisztencia feltétele

$$a_0 = 1, \quad \sum_{k=0}^{m} a_k = 0 \tag{9.5.17}$$

$$\sum_{k=1}^{m} k a_k = -1. \tag{9.5.18}$$

A $p \geq 2$ rendűség feltétele a (9.5.14) alakból jól láthatóan a

$$\sum_{k=1}^{m} k^{j-1}(k a_k + j b_k) = 0, \quad j = 2, 3, \ldots, p \tag{9.5.19}$$

feltételek. Mint látható, ebben a felírásban $2m+2$ ismeretlenünk van, és $p+2$ feltételt tűztünk ki rájuk. Így (természetesen) $p \leq 2m$, és a $p = 2m$ maximális rend eléréséhez az együtthatókat az alábbi módon határozhatjuk meg.

1. Megoldjuk az $a_1, a_2, \ldots, a_m$ és $b_1, b_2, \ldots, b_m$ ismeretlenekre a $2m$ számú egyenletetből álló (9.5.18)–(9.5.19) rendszert.

- Ezután a (9.5.16) és a (9.5.17) feltételekből meghatározzuk az a_0 és b_0 együtthatókat az

$$a_0 = -\sum_{k=1}^{m} a_k = 1, \quad b_0 = 1 - \sum_{k=1}^{m} b_k \tag{9.5.20}$$

összefüggésekből.

◇

A lineáris többlépéses módszerek vizsgálatánál hasznos a következő két polinom bevezetése:

$$\varrho(\xi) = \sum_{k=0}^{m} a_k \xi^{m-k} \tag{9.5.21}$$

$$\sigma(\xi) = \sum_{k=0}^{m} b_k \xi^{m-k}. \tag{9.5.22}$$

9.5.9. definíció.
A (9.5.21) és (9.5.22) módon definiált $\varrho(\xi)$ és $\sigma(\xi)$ legfeljebb m-ed fokú polinomokat a (9.5.5) m-lépéses lineáris többlépéses módszer első illetve második karakterisztikus polinomjának nevezzük.

A (9.5.12) feltételből - némi számolás után - adódik a következő [1, 34].

9.5.10. tétel.
Egy lineáris többlépéses módszer pontosan akkor konzisztens, amikor a karakterisztikus polinomjaira érvenyesek a

$$\varrho(1) = 0, \quad \varrho'(1) = \sigma(1) \tag{9.5.23}$$

összefüggések.

9.5.2. A kezdeti értékek megválasztása és a módszer konvergenciája

Láttuk, hogy egy m-lépéses lineáris többlépéses módszer feltételezi a kezdeti közelítések ismeretét a rácsháló első m pontjában, azaz feltesszük, hogy $y_0, y_1, \ldots, y_{m-1}$ adottak. Ugyanakkor, a (9.3.2) kezdeti feltételből csak y_0 értéke ismert. Ezeket a közelítéseket egy megfelelő rendben pontos egylépéses módszerrel határozzuk meg. (Tipikusan valamely Runge-Kutta típusú módszerrel.) Itt ügyelnünk kell arra, hogy a kiválasztott módszer az alkalmazott lineáris többlépéses módszer rendjével egyezzék meg, hiszen ellenkező esetben a kezdeti közelítések alacsonyabb rendű meghatározásával elveszíthetjük a teljes módszer pontosságát.[17]

Térjünk át a lineáris többlépéses módszerek konvergenciájának kérdésére. Mint azt már az egylépéses módszereknél láttuk, a konzisztencia önmagában nem elégséges a konvergenciához. A továbbiakban, bizonyítás nélkül, megadunk olyan feltételt, amely mellett a konvergencia biztosított. (A részletek iránt érdeklődöknek javasoljuk az irodalomjegyzék [1, 34] hivatkozásait.)

Mint láttuk, konzisztens módszerek esetén $\xi = 1$ gyöke a $\varrho(\xi)$ karakterisztikus polinomnak. A következő definíció arra vonatkozik, hogy milyen egyéb gyökök lehetnek még.

9.5.11. definíció.
Azt mondjuk, hogy egy lineáris többlépéses módszer *kielégíti a gyökkritériumot*, ha a $\varrho(\xi) = 0$ karakterisztikus egyenlet $\xi_k \in \mathbb{C}$ $(k = 1, 2, \ldots, m)$ gyökeire $|\xi_k| \leq 1$, és a $|\xi_k| = 1$ tulajdonságú gyökök egyszeresek.

Mint azt a következő tétel mutatja, ez a feltétel a lineáris többlépéses módszer stabilitását jelenti.

9.5.12. tétel.
Ha egy lineáris többlépéses módszer konzisztens és érvényes rá a gyökkritérium, akkor konvergens is, azaz tetszőleges rögzített $t^\star \in (0, T)$ pontban $h \to 0$ esetén $y_n \to u(t^\star)$, ahol $nh = t^\star$.

[17] Megjegyezzük, hogy a modern programcsomagokban y_k $(k = 1, 2, \ldots m-1)$ értékét többnyire egy megfelelően megválasztott $k-1$-lépéses módszerrel számolják.

9.5.13. megjegyzés. A gyökkritérium teljesülésének szükségességét mutatja a következő példa. Tekintsük az

$$y_i + 4y_{i-1} - 5y_{i-2} = h(4f_{i-1} + 2f_{i-2}) \tag{9.5.24}$$

kétlépéses explicit módszert. Könnyen ellenőrizhetően ez a módszer maximálisan pontos, azaz konzisztenciarendje $p = 2m - 1 = 3$. Ugyanakkor első karakterisztikus polinomja $\varrho(\xi) = \xi^2 + 4\xi - 5 = (\xi - 1)(\xi + 5)$, tehát a gyökkritérium nem teljesül. Oldjuk meg az $u' = 0$ feladatot az $u(0) = 0$ kezdeti feltétellel. A feladat megoldása $u(t) = 0$. Legyen továbbá $y_0 = 0$ és $y_1 = \varepsilon$. (Ha valamilyen egylépéses módszerrel kiszámoljuk y_1 értékét, akkor az a módszer hibájának megfelelő rendben, de várhatón csak kissé fog eltérni a pontos megoldástól az első időpontban, így ε egy nullától különböző kis számnak tekinthető.) A fenti kétlépéses módszerrel számolva a kezdeti közelítésekre az alábbiakat kapjuk:

$$\begin{aligned} y_2 &= -4y_1 = -4\varepsilon, \\ y_3 &= -4y_2 + 5y_1 = 21\varepsilon, \\ y_4 &= -4y_3 + 5y_2 = -104\varepsilon, \quad \text{stb.} \end{aligned} \tag{9.5.25}$$

Láthatóan a numerikus megoldás nem marad korlátos, így a konvergencia sem lehetséges. ◇

Fontos megjegyeznünk, hogy a numerikus realizálások során a gyökkritérium általában nem elégséges a stabilitáshoz: vannak esetek, amikor a gyökkritérium teljesülése ellenére a számítások során kialakuló hibák miatt a módszerek nem szolgáltatnak megbízható eredményt. Az ilyen módszereknél az a probléma, hogy a karakterisztikus egyenletének több (bár csak egyszeres) egy abszolút értékű gyöke is van.

9.5.14. definíció.
Azt mondjuk, hogy egy lineáris többlépéses módszer *erősen stabil*, ha kielégíti a gyökkritériumot, és csak a $\xi = 1$ az egyetlen egy abszolút értékű gyöke.

Például az alábbi ún. Milne-módszerre

$$y_i - y_{i-2} = \frac{h}{3}(f_i + 4f_{i-1} + f_{i-2})$$

a gyökök $\xi_{1,2} = \pm 1$. Ezért a gyökkritérium érvényes, viszont nem lesz erősen stabil. Ezért ennek a módszernek a használata általában nem ajánlott. Mivel elsősorban az erősen stabil lineáris többlépéses módszerek használata a célszerű, ezért fontos megemlíteni G. Dahlquist eredményét, amely az ilyen típusú módszerek rendjéről szól.

9.5.15. tétel.
Egy erősen stabil m-lépéses lineáris többlépéses módszer legfeljebb $m + 1$-ed rendű lehet.

Az egylépéses módszereknél már bemutattuk, hogy egy módszer konvergenciája nem garantálja a megoldás adekvát viselkedését valamely rögzített rácshálón. (Csupán azt biztosítja, hogy a numerikus megoldás a "megfelően kicsi lépésközű rácshálón" közel van a pontos megoldáshoz. Ugyanakkor ez a lépésköz túlságosan kicsi is lehet.) Ennek elkerülése céljából definiáltuk az abszolút stabil módszereket. (Lásd a 9.4.15. definíciót.) Felmerülhet a kérdés: mi a helyzet a lineáris többlépéses módszerek abszolút stabilitásával? Sajnálatosan a válasz azt mutatja, hogy ezekre a módszerekre az A-stabilitást nehéz biztosítani. Ugyanis az ún. első és második Dahlquist-korlátok szerint

- az explicit lineáris többlépéses módszerek nem lehetnek A-stabilak;
- az A-stabil lineáris többlépéses módszerek rendje nem lehet kettőnél nagyobb.

Ezek a lineáris többlépéses módszerek alkalmazhatóságára nézve komoly korlátot jelentenek .

9.5.3. Adams-típusú módszerek

A lineáris többlépéses módszerek között kiemelkedő szerepet játszanak azok, amelyekre a (9.5.5) képletben

$$a_0 = 1, \quad a_1 = -1, \quad a_2 = a_3 = \ldots = a_m = 0. \tag{9.5.26}$$

Az ilyen módszereket *Adams-típusú módszereknek* (vagy röviden *Adams-módszereknek*) nevezzük.[18](Például a (9.5.4) módszer egy Adams-módszer.) Az Adams-típusú módszereknél megválasztható paraméterként a $b_0, b_1, \ldots, b_m$ értékei szolgálnak.

9.5.16. megjegyzés. Mint az ismeretes, a (9.2.6) kezdetiérték-feladat $u(t)$ megoldására érvényes a (9.3.3) azonosság a $[0, T]$ intervallumon. Így ezt integrálva a $[t_i, t_{i+1}]$ intervallumon a már ismert

$$u(t_{i+1}) - u(t_i) = \int_{t_i}^{t_{i+1}} f(t, u(t))\,\mathrm{d}t, \quad t \in [0, T] \tag{9.5.27}$$

$(i = 0, 1, \ldots, N-1$ tetszőleges) egyenlőséget nyerjük. A 9.3.2. szakaszban a jobb oldal numerikus kiintegráláshoz egyszerű formulákat alkalmaztunk. (V.ö. a (9.3.68), (9.3.69) és a (9.3.70) képleteket.) Az Adams-típusú módszereknél a numerikus integráláshoz egy több pontra támaszkodó kvadratúra-képletet alkalmazunk, és ezek határozzák meg a $b_0, b_1, \ldots, b_m$ értékeit. ◇

Alapvető különbség van a $b_0 = 0$ és a $b_0 \neq 0$ megválasztású Adams-módszer között: míg az első esetben a módszer explicit, addig a második esetben implicit.

9.5.17. definíció.
A $b_0 = 0$ megválasztású Adams-módszereket *Adams–Bashforth-módszernek*, a $b_0 \neq 0$ megválasztás melletti Adams-módszereket pedig *Adams–Moulton-módszernek* nevezzük.

Az Adams-típusú módszerek konzisztenciájának illetve p-ed rendűségének feltétele közvetlenül meghatározható a (9.5.13) és a (9.5.14) feltételekből.

9.5.18. tétel.
Egy Adams-típusú módszer pontosan akkor konzisztens, amikor

$$b_0 + b_1 + \ldots + b_m = 0. \tag{9.5.28}$$

Továbbá a módszer $p \geq 2$ rendben pontos, ha (9.5.28) mellett teljesülnek a

$$\sum_{k=1}^{m} k^{j-1} b_k = \frac{1}{j}, \quad j = 2, 3, \ldots, p \tag{9.5.29}$$

feltételek.

[18] John Couch Adams (1819 - 1892), angol matematikus és csillagász, Francis Bashforth (1819 – 1912) angol matematikus és fizikus, Forest Ray Moulton (1872 - 1952) USA-beli csillagász.

Az m-lépéses Adams-típusú módszerek maximális rendje is könnyen meghatározható: az Adams–Moulton-módszeré $p = m + 1$, az Adams–Bashforth-módszer pedig $p = m$. Egy tetszőleges Adams-módszer első karakterisztikus polinomja

$$\varrho(\xi) = \xi - 1. \tag{9.5.30}$$

Ennek egyetlen gyöke a $\xi = 1$, ezért ezekre a módszerekre mindig teljesül a gyökkritérium, és emellett a módszerek erősen stabilak is. Ezért érvényes a következő állítás.

9.5.19. tétel.
Az Adams-típusú módszerek a konzisztenciájukkal megegyező rendben konvergensek.

A következő 9.5.1. táblázatban hatodik rendig bezárólag felsoroljuk a maximális rendű Adams–Bashforth-módszerek b_k együtthatóit. (A könnyebb áttekinthetőség kedvéért nem törtként írjuk fel ezeket, és a harmadik oszlopban jelezzük, hogy b_k hányszorosa szerepel a táblázatban.)

p	m	b_k	1	2	3	4	5	6
1	1	b_k	1					
2	2	$2b_k$	3	-1				
3	3	$12b_k$	23	-16	5			
4	4	$24b_k$	55	-59	37	-9		
5	5	$720b_k$	1901	-2774	2616	-1274	251	
6	6	$1440b_k$	4277	-7923	9982	-7298	2877	-475

9.5.1. táblázat: A maximális rendű Adams–Bashforth-módszerek b_k együtthatói.

A táblázatban, az elméletünknek megfelelően, $p = m$. Az $m = 1$ az explicit Euler-módszert jelenti. Az $m = 2$ választás esetén a már ismert (9.5.4) módszert kapjuk. Az Adams–Bashforth-módszerek, mivel explicit módszerek, nem A-stabilak. (Lásd az első Dahlquist-féle korlátot.) Ráadásul az abszolút stabilitási tartományuk nagyon kicsi. Az $m = 1$ (vagyis az explicit Euler-módszer) esetben az abszolút stabilitási tartományuk az $|1+z| \leq 1$ tulajdonságú komplex számok, azaz a $(-1, 0)$ közepű, egységnyi sugarú kör a komplex számsíkon. (V.ö. (9.4.53).) Emellett m növelésével ez a tartomány tovább szűkül. Ezért ezek a módszerek nem alkalmasak az olyan feladatok megoldására, ahol az abszolút stabilitás szükséges. Ez motiválja az Adams–Moulton-módszerek alkalmazását.

A 9.5.2. táblázatban ismertetjük a maximális rendű Adams–Moulton-módszerek együtthatóit. A rendjük, az elméletnek megfelelően, $p = m + 1$.

Az első módszer ($m = 1$, $b_1 = 0$) az implicit Euler-módszert jelenti. Ez a módszer elsőrendű, így nem maximális rendű. A táblázat második módszere ($m = 1$, $\beta_1 \neq 0$) a (9.3.61) képletű Crank–Nicolson-módszert jelenti. Megjegyezzük, hogy az Adams–Moulton-módszerek stabilitási tartományai jóval nagyobbak, mint az azonos rendű Adams–Bashforth-módszereké. A fenti Adams–Moulton-módszerek közül csak az első kettő (azaz az implicit Euler-módszer és a Crank–Nicolson-módszer) A-stabil, a többi nem. (Ez következik a már említett második Dahlquist-féle korlátból.)

9.5.20. megjegyzés. Az Adams–Bashforth-módszer és az Adams–Moulton-módszer gyakran együttesen, egymással kombinálva kerülnek felhasználásra a következő módon. Először egy Adams–Bashforth-módszerrel kiszámolt $y_i^\star$ értékkel "megjósoljuk" a t_i időrétegen a közelítő értéket, majd

p	m	b_k	0	1	2	3	4	5
1	1	b_k	1					
2	1	$2b_k$	1	1				
3	2	$12b_k$	5	8	-1			
4	4	$24b_k$	9	19	-5	1		
5	5	$720b_k$	251	646	-264	106	-19	
6	6	$1440b_k$	475	1427	-798	482	-173	27

9.5.2. táblázat: A maximális rendű Adams–Moulton-módszerek b_k együtthatói.

az Adams–Moulton-módszerrel "javítjuk" a numerikus megoldást a következő módon: a jobb oldalon szereplő $b_0 f_i$ tagba f_i helyett $f_i^\star = f(t_i, y_i^\star)$ értékét rakjuk. Az így nyert módszert "jóslo-javító" (angolul: "predictor-corrector", PC) módszernek nevezzük, de szokásos az összefoglaló *Adams-Bashforth-Moulton-módszer* elnevezés is. Lényeges tulajdonsága, hogy ez a módszer már explicit.

A PC módszer néhány elméleti részlete megtalálható a
http://math.fullerton.edu/mathews/n2003/AdamsBashforthMod.html
linken, és konkrét formulákkal animáción is megfigyelhetjük a módszer viselkedését az $u' = 1 - t\sqrt[3]{u}$ differenciálegyenleten. ◇

9.5.4. Retrográd differencia módszerek

A lineáris többlépéses módszerek között, az Adams-típusú módszerek mellett fontosak azok az implicit módszerek, amelyekre a (9.5.5) képletben

$$b_0 \neq 0, \quad b_1 = b_2 = \ldots = b_m = 0. \tag{9.5.31}$$

Az ilyen módszereket *retrográd differencia módszereknek* nevezzük. A módszer konzisztenciáját és annak rendjét az $a_0, a_1, \ldots, a_m$ együtthatók alkalmas megválasztásával biztosítjuk. A retrográd differencia módszerek fontos tulajdonsága, hogy az $f(t, u)$ függvényt csak egy pontban (a (t_i, y_i) pontban) szükséges kiértékelnünk. Ez a módszer különösen alkalmas a speciális stabilitást igénylő feladatok megoldására. (Lásd a következőkben a merev rendszerek leírását.) A 9.5.6. megjegyzés alapján érvényes a következő állítás.

9.5.21. tétel.
Az m-lépéses

$$y_i + a_1 y_{i-1} + \cdots + a_m y_{i-m} = h b_0 f_i, \quad i = m, m+1, \ldots, \tag{9.5.32}$$

alakú retrográd differencia módszer pontosan akkor konzisztens, amikor

$$\begin{aligned} a_1 + \ldots + a_m &= -1 \\ a_1 + 2a_2 + \ldots + m a_m &= -b_0. \end{aligned} \tag{9.5.33}$$

Továbbá $p \geq 2$ rendben pontos, ha (9.5.33) mellett teljesülnek a

$$\sum_{k=1}^{m} k^j a_k = 0, \quad j = 2, 3, \ldots, p \tag{9.5.34}$$

feltételek.

Ez azt jelenti, hogy $p+1$ számú feltételünk van az $m+1$ darab $b_0, a_1, a_2, \ldots, a_m$ ismeretlenekre. Tehát a retrográd differencia módszer maximális rendje $p = m$. A következő 9.5.3. táblázatban megadjuk az első hat maximális rendű retrográd differencia módszer együtthatóit.

p	m	b_0	a_0	a_1	a_2	a_3	a_4	a_5	a_6
1	1	1	1	-1					
2	2	$\frac{2}{3}$	1	$-\frac{4}{3}$	$\frac{1}{3}$				
3	3	$\frac{6}{11}$	1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$			
4	4	$\frac{12}{25}$	1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$		
5	5	$\frac{60}{137}$	1	$-\frac{300}{137}$	$\frac{300}{137}$	$-\frac{200}{137}$	$\frac{75}{137}$	$-\frac{12}{137}$	
6	6	$\frac{60}{147}$	1	$-\frac{360}{147}$	$\frac{450}{147}$	$-\frac{400}{147}$	$\frac{225}{147}$	$-\frac{72}{147}$	$\frac{10}{147}$

9.5.3. táblázat: Az első hat maximális rendű retrográd differencia módszer együtthatói.

Az első módszer ($m = 1$) az implicit Euler-módszert jelenti, míg az $m = 2, 3, 4$ módszereket rendre *másod-, harmad- és negyedrendű Curtis-Hirschfeld-módszernek* nevezzük. Az első két módszer ($m = 1, 2$) A-stabil, a többi módszer – a második Dahlquist-féle korlátnak megfelelően– viszont nem. Ugyanakkor a stabilitási tartományuk igen nagy.[19] Megjegyezzük, hogy $m > 6$ esetén a retrográd differencia módszerek már a legalapvetőbb stabilitási tulajdonsággal (az ún. 0-stabilitással) sem rendelkeznek, ezért a hatodrendűnél magasabb módszereket nem alkalmazzák.

[19] Ha $m = 3, 4, 5, 6$, akkor a módszerek stabilitási tartománya ugyan nem tartalmazza a teljes $\mathbb{C}^-$ komplex félsíkot, de a valós negatív tengelyt tartalmazó, origóból $\pm\alpha$ szöggel kiinduló szögtartományt igen. Az ilyen tulajdonságú módszereket $A(\alpha)$-stabil módszereknek nevezzük. (Tehát az A-stabilitás az $\alpha = 90°$ fokos stabilitást jelenti.) Ugyanakkor ez a szektor m növekedésével beszűkül: míg $m = 3$ esetén $\alpha \simeq 86°$, (azaz "majdnem" A-stabil), addig $m = 6$ mellett már csak $\alpha \simeq 17,8°$.

9.6. A lineáris és a merev rendszerek numerikus megoldása

Eddig skaláris egyenletek megoldásával foglalkoztunk. Most tekintsük a

$$\begin{aligned} \frac{d\mathbf{u}}{dt} &= \mathbf{A}\mathbf{u}(t), \quad t \in (0,T], \\ \mathbf{u}(0) &= \mathbf{u}_0 \end{aligned} \tag{9.6.1}$$

feladatot, ahol $\mathbf{A} \in \mathbb{R}^{m\times m}$ adott mátrix, $\mathbf{u}_0 \in \mathbb{R}^m$ adott vektor, és $\mathbf{u} : [0,T] \to \mathbb{R}^m$ az ismeretlen függvény. A (9.6.1) feladatot egy *lineáris közönséges differenciálegyenlet–rendszer Cauchy–feladatának* nevezzük. Az egyszerűség kedvéért tegyük fel, hogy $\mathbf{A}$ diagonalizálható mátrix, azaz létezik olyan $\mathbf{S}$ reguláris mátrix, amellyel $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{\Lambda}$, ahol $\mathbf{\Lambda}$ az $\mathbf{A}$ mátrix sajátértékeit tartalmazó diagonális mátrix.[20] Vezessük be a $\mathbf{w}(t) = \mathbf{S}^{-1}\mathbf{u}(t)$ új ismeretlen függvényt! Ekkor a (9.6.1) Cauchy-feladat átírható a

$$\begin{aligned} \frac{d\mathbf{w}}{dt} &= \Lambda\mathbf{w}(t), \quad t \in (0,T], \\ \mathbf{w}(0) &= \mathbf{S}^{-1}\mathbf{u}_0 \end{aligned} \tag{9.6.2}$$

alakra.

Vegyük észre, hogy (9.6.2) egy olyan m-ismeretlenes rendszer, amely valójában szétesik m darab

$$w_k' = \lambda_k w_k, \quad w_k(0) = \text{adott}, \ k = 1,2,\ldots,m \tag{9.6.3}$$

skaláris feladatra, amelyet a korábbiakban (v.ö. (9.4.49)) már vizsgáltunk. Következésképpen a lineáris rendszerek numerikus megoldása visszavezetődik a skaláris feladatokra megfogalmazott módszerekre. Ez azt jelenti, hogy ezeket a módszereket a (9.6.3) feladatokra kell alkalmazni, ahol λ_k az $\mathbf{A}$ mátrix sajátértékei. Tehát egy numerikus módszer alkalmazása során azt kell megvizsgálnunk, hogy a módszer hogyan viselkedik a mátrix spektrumán a (9.6.3) tesztfeladatok esetén. Az explicit Euler-módszer esetén láttuk, hogy egyetlen egyenlenlet esetén ez az időlépcső megválasztására a (9.4.53) feltételt jelenti. Ezért $\lambda_k < 0$ esetén az

$$|R_{EE}(-h\lambda_k)| \le 1 \quad \Leftrightarrow h \le 2/(-\lambda_k), \ k = 1,2,\ldots m, \tag{9.6.4}$$

azaz a

$$h \le \frac{2}{\max_k(-\lambda_k)} \tag{9.6.5}$$

feltételt kapjuk.

9.6.1. megjegyzés. Nem foglalkozunk külön a magasabb rendű közönséges differenciálegyenletekkel, mert azok az ún. *átviteli elv* segítségével átírhatók elsőrendű rendszerré. Például, az

$$y^{(m)} + a_1 y^{(m-1)} + a_2 y^{(m-2)} + \ldots + a_{m-1}y' + a_m y = 0 \tag{9.6.6}$$

m-ed rendű, lineáris, homogén, közönséges differenciálegyenletet az

$$y(0) = c_1, \ y'(0) = c_2, \ \ldots, y^{(m-1)}(0) = c_m$$

[20] Emlékeztetünk rá (lásd az első szakaszt), hogy hasonlósági transzformációval diagonális alakra hozható mátrixokat *egyszerű struktúrájú mátrixoknak* is szokásos nevezni. Megjegyezzük, hogy egy mátrix pontosan akkor diagonalizálható, ha létezik m darab lineárisan független sajátvektora. A definícióban szereplő S hasonlósági mátrix választható ezen sajátvektorokból mint oszlopvektorokból összeállított mátrixként.

kezdeti feltételekkel a következő módon lehet átírni egy m-ismeretlenes lineáris rendszerré. Vezessük be az $u_1, u_2, \ldots, u_m$ új ismeretlen függvényeket:

$$\begin{aligned} u_1(t) &= y(t) \\ u_2(t) &= y'(t) = u_1'(t) \\ u_3(t) &= y''(t) = u_2'(t) \\ &\vdots \\ u_m(t) &= y^{m-1}(t) = u_{m-1}'(t). \end{aligned} \tag{9.6.7}$$

Deriválva az utolsó egyenletet, és felhasználva az egyenletünket, a bevezetett új függvényekkel ekkor az

$$u_m'(t) = y^m(t) = -a_1 u_m - a_2 u_{m-1} - \ldots - a_m u_1 \tag{9.6.8}$$

egyenletet kapjuk. Felhasználva a (9.6.7) egyenleteket és a (9.6.8) összefüggést, az $\mathbf{u}(t) : [0,T] \to \mathbb{R}^m$ $u_i(t)$ komponensű ismeretlen függvényre az

$$\mathbf{u}' = \mathbf{A}\mathbf{u}$$

elsőrendű, lineáris rendszert kapjuk, ahol $\mathbf{A} \in \mathbb{R}^{m\times m}$,

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ -a_m & -a_{m-1} & -a_{m-2} & \ldots & -a_2 & -a_1 \end{bmatrix}$$

adott mátrix. A megfelelő kezdeti feltétel $\mathbf{u}(0) = \mathbf{c}$, ahol $\mathbf{c} \in \mathbb{R}^m$ $c_1, c_2, \ldots c_m$ komponensű adott vektor. $\diamond$

Térjünk át a merev rendszerekre! Tekintsük példaként az $y'' + (\gamma+1)y' + \gamma y = 0$ másodrendű differenciálegyenletet az $y(0) = 1$ és $y'(0) = \gamma - 2$ kezdeti feltételekkel! Ekkor a megfelelő rendszerre

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\gamma & -(\gamma+1) \end{bmatrix},$$

a kezdeti vektor pedig $\mathbf{c} = [1, \gamma - 2]$. Az $\mathbf{A}$ mátrix karakterisztikus egyenlete

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \lambda^2 + (\gamma+1)\lambda + \gamma = 0.$$

Ezért az $\mathbf{A}$ mátrix sajátértékei $\lambda_1 = -1$ és $\lambda_2 = -\gamma$. A pontos megoldás tehát

$$\begin{aligned} u_1(t) &= 2\exp(-t) - \exp(-\gamma t) \\ u_2(t) &= -2\exp(-t) + \gamma\exp(-\gamma t). \end{aligned} \tag{9.6.9}$$

Ha az explicit Euler-módszert alkalmazzuk a feladat megoldására, akkor a lépésköz megválasztására $\gamma \geq 1$ esetén a $h < 2/\gamma$ feltételt kapjuk. Tegyük fel, hogy $\gamma >> 1$. Ekkor a megválasztható lépésköz igen kicsi, viszont, mint az a pontos megoldás (9.6.9) alakjából látható, az egyes megoldásokban szereplő $\exp(-\gamma t)$ függvények már nagyon kis t_0 mellett $t \geq t_0$ esetén nem játszanak szerepet a megoldásban, és a pontos megoldás gyakorlatilag az $u_1(t) \simeq -u_2(t) \simeq 2\exp(-t)$ lesz a $[t_0, T]$ intervallumon. Ez azt jelenti, hogy a teljes intervallumon nem szükséges a h lépésköz ilyen kicsi megválasztása. Az ilyen tulajdonságú rendszereket *merev*[21] *rendszernek* nevezzük. A merev feladatokat az alábbi módon szokásos definiálni.

[21] Gyakran a magyar terminológia is az angol "stiff" kifejezést használja.

9.6.2. definíció.

Azt mondjuk, hogy a (9.6.1) lineáris rendszer merev, ha a rendszer $\mathbf{A} \in \mathbb{R}^{m\times m}$ mátrixának λ_k $(k = 1, 2, \ldots, m)$ sajátértékei rendelkeznek az alábbi tulajdonságokkal.

1. $Re\lambda_k < 0$ minden $k = 1, 2, \ldots, m$ esetén. (Azaz a feladat Ljapunov-értelemben aszimptotikusan stabil.)
2. Az
$$S = \frac{\max_k |Re\lambda_k|}{\min_k |Re\lambda_k|} \tag{9.6.10}$$
módon definlált S merevségi mutató nagy, azaz $S >> 1$.

A merev rendszerek esetén tehát a pontos megoldás gyorsan és lassan lecsengő komponensekből adódik össze, és egy (általában kis) $t = t_0$ időpont után a megoldást szinte teljesen csak a lassan változó komponensek határozzák meg. Az ilyen feladatok numerikus kezelésére az explicit módszerek általában nem alkalmasak, és tipikusan az A-stabil módszerek használata javasolt. A korábbi módszereink közül különösen alkalmasak a retrográd differencia módszerek, ezen belül is az implicit Euler-módszer illetve az első két Curtis-Hirschfeld-módszer.

9.6.3. megjegyzés. Igazából nincs egységes és pontos definició a merev rendszerekre. Ez a fogalom valójában azt fejezi ki, az eredeti folytonos feladat nagyfokú stabilitással rendelkezik. Például az előző másodrendű feladatban γ értékétől (mint bemenő adattól) gyakorlatilag nem függ a megoldás, amely a szokásos stabilitási fogalomnál ("folytonosan függ a megoldás a bemenő adatoktól") jóval erősebb tulajdonság, azt is mondhatjuk, hogy túlságosan is stabil a feladat. Ezt a numerikus módszerek csak bizonyos feltételek mellett (egészen pontosan, az A-stabilitás mellett) képesek csak követni. Ezért a merev rendszerek numerikus megoldására az explicit módszerek gyakorlatilag nem alkalmazhatók. Megjegyezzük továbbá, -mint azt a példánk is mutatja- ezzel a tulajdonsággal már egy skaláris egyenlet is rendelkezhet. Például, tekintsük az

$$u' = ku,\ t > 0;\quad u(0) = u_0$$

feladatot, ahol k adott állandó, akkor ennek megoldása $u(t) = u_0 \exp(kt)$. Ha $k << 0$, akkor gyakoraltilak már kis t_0 esetén a $(0, t_0$ intervallumon $u(t)$ lecsökken u_0-ról a legkisebb ábrázolható pozitív számhoz, és ezután minden $t > t_0$ esetén $u(t) \sim 0$. Így a feladat az időintervallum nagy részén az u_0 ketzedeti állapottól függetlenül viselkedik. Erre a feladatra már viszonylag kis k esetén is az explicit Euler-módszer rosszul viselkedik. Például, $k = -15$ esetén, a $h = 1/4$ lépésközű numerikus megoldás kinő a végtelenbe; a $h = 1/8$ megválasztású módszer ugyan korlátos marad és követi is a pontos megoldást, de beoszcillál. Ha a Crank–Nicolson-módszert alkalmazzuk, akkor viszont a numerikus megoldás jól követi a pontos megoldást. Bővebben lásd a http://en.wikipedia.org/wiki/Stiff_equation linket, de saját program készítésével önállóan is ellenőrizhető a numerikus megoldások fenti viselkedése. (Lásd a következő 9.7. szakaszt.) ⋄

Befejezésül megemlítjük, hogy a merevség fogalma kiterjeszthető a nemlineáris egyenletekre, illetve a nemlineáris rendszerekre is. Ilyenkor a linearizált feladatra fogalmazzuk meg a feltételeket, azaz rendszerek esetén az $\mathbf{A}$ mátrix szerepét a rendszer Jacobi-mátrixa játsza.

9.7. A kezdetiérték-feladatok numerikus megoldása MATLAB segítségével

A Matlab segítségével viszonylag egyszerűen realizálhatók a numerikus eljárásaink. A MATLAB programrendszer rendelkezik saját, már elkészített és beépített programmal, de az egyszerűbbek önálló elkészítése sem nehéz. Például az explicit Euler-módszer megírása egy m-fájlban és a továbbiakban önálló függvényként való használata igen egyszerű.

Tekintsük azokat a lépéseket, amelyek ennek megvalósításához szükségesek.

- Első lépésben indítsuk el a MATLABot, majd az Editor segítségével készítsük el az alábbi m-fájlt!:

```
function[t,y] = expeuler(diffegy, t0, y0, h, N)
t=zeros(N+1,1);
y = zeros(N+1,1);
t(1) = t0;
y(1) = y0;
for i=1:N
t(i+1) = t(i) + h;
y(i+1) = y(i) + h * diffegy(t(i),y(i));
end
```

  A fenti programban az első sorban azt adtuk meg, hogyan tudjuk meghívni a módszerünket. (Mi most expeuler-nek neveztük el.) Utána soroljuk fel a megoldandó feladatot illetve az explicit Euler- módszert beazonosító bemenő paramétereket, a bal oldalon pedig a kimenő paramétereket. (Ezek tipikusan a kiszámított és a további célra felhasználni kívánt eredmények.) A mi esetünkben öt bemenő és kettő kimenő paraméter van. A két kimenő paraméter a diszkretizált idő vektor és az ezen pontokban kiszámolt numerikus közelítő megoldások vektora. Az első bemenő paraméter egy alfüggvény, aminek jelen esetben `diffegy` a neve, és ebben a függvényben írjuk meg a differenciálegyenletet specifikáló f függvényt. A második paraméter t0, ez a kezdeti időpontot jelöli, a harmadik a t0 pontbeli y0 kezdeti értéket jelenti. A következő paraméter h, amely az időintervallum diszkretizációs lépéstávolságát jelenti, majd végül N jelöli a megtett időlépések számát. A második és harmadik sorban vesszük fel a t és y vektorokat, amikben az értékeket először lenullázzuk, és a következő két sorban beállítjuk a kezdőértékeket. Utána következik lényegében a módszer algoritmusa: egy ciklus keretében először beállítjuk a t_i értékeket, majd kiszámoljuk a meredekséget, végül az y_i értékeket az explicit Euler-módszer képletének megfelelően.

- A fenti programmal még nem tudjuk közvetlenül az adott differenciálegyenlet numerikus megoldását előállítani, ehhez szükségünk van az f függvényt megadó "diffegy" alfüggvény megadására. Ha az

$$u'(t) = -u(t) + t + 1,$$

  ahol $t \in [0,1]$ és $u(0) = 1$ feladat[22] megoldását szeretnénk előállítani, akkor a diffegy nevű

[22] Emlékeztetünk rá, hogy ezen a feladaton teszteltük a numerikus módszereinket a 9.3.1. és a 9.4. fejezetekben. Ebben a részben a Runge-Kutta típusú módszerrel illetve lineáris többlépéses módszerrel is megoldjuk ezt a feladatot.

alfüggvény elkészítéséhez nyissunk egy új m-fájlt, amibe írjuk a következőket:

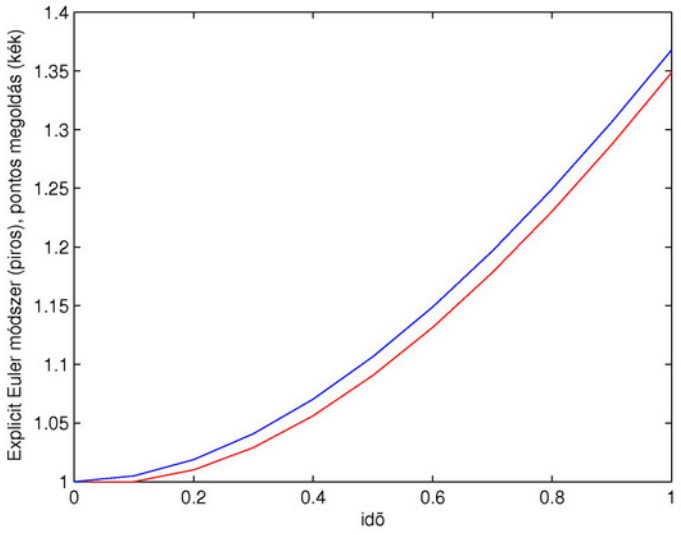
```
function dydt = diffegy(t,y)
dydt = -y + t + 1;
```

- Ha már megírtuk mindkét függvényt, futtassuk le a programunkat $h = 0.1$, $h = 0.01$ és $h = 0.001$ lépéstávolságokkal a $[0, 1]$ intervallumon a `[T1,Ye] = expeuler(@diffegy, 0, 1, 0.1, 10)` utasítással. Enter után megkapjuk a T1 és Ye vektorokat, amelyek a közelítések időbeli helyét és értékét tartalmazzák. Ha ki is szeretnénk rajzoltatni, akkor a `plot(T1,Ye)` paranccsal egy külön ablakban megjelenik a függvényünk. (A $h = 0.01$ és $h = 0.001$ lépésközökre értelemszerűen a `[T1,Ye] = expeuler(@diffegy, 0, 1, 0.01, 100)` illetve a `[T1,Ye] = expeuler(@diffegy, 0, 1, 0.001, 1000)` parancsokat kell beírni.)

Ha az explicit Euler-módszer hibájára vagyunk kíváncsiak, akkor a pontos megoldással szükséges a numerikus megoldásunkat összehasonlítani. A példánk pontos megoldása az

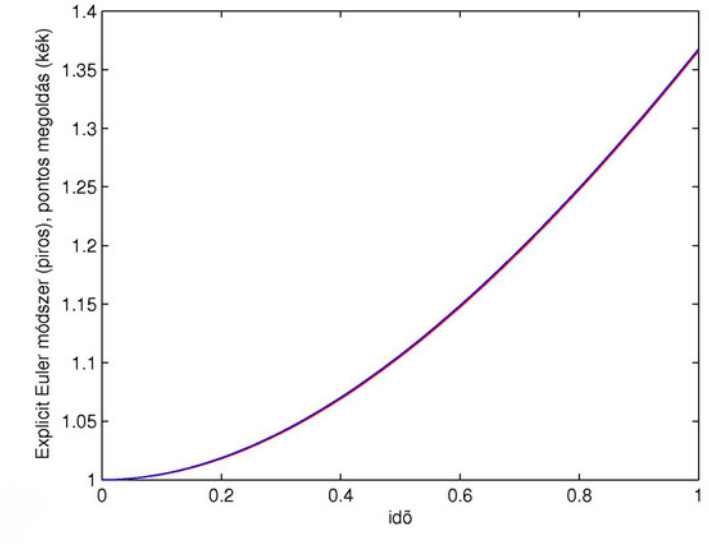
$$u(t) = e^{-t} + t$$

függvény. A 9.7.1.-9.7.3. ábrákon láthatjuk a módszer pontosságát a $h = 0.1$, $h = 0.01$ és $h = 0.001$ lépéstávolságokkal. Az elvártaknak megfelelően h csökkenésével a numerikus megoldás grafikonja közeledik a pontos megoldás grafikonjához.

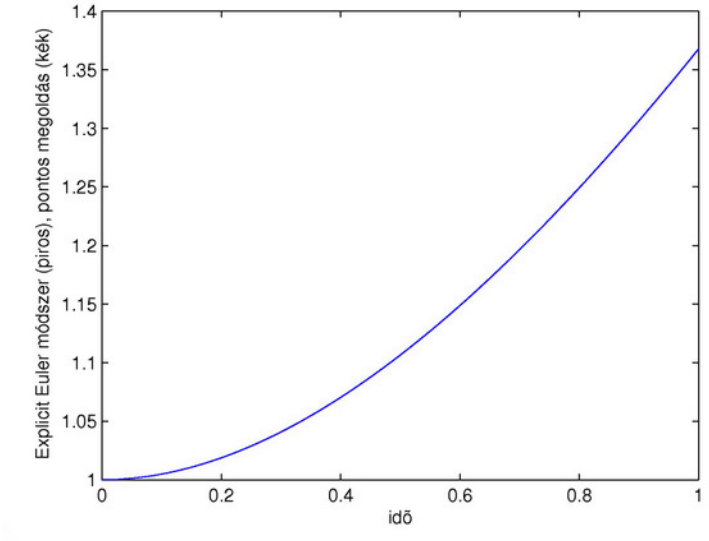


9.7.1. ábra: A $h = 0.1$ lépésközű explicit Euler-módszer

A differenciálegyenlet-rendszerek numerikus kezelésére készített programot egy egyszerű populációdinamikai modellen, az ún. Lotka-Volterra-modellen mutatjuk be. Ez az ún. ragadozó-zsákmány modell, amely két faj egyedei számának időbeli alakulását (fejlődését) írja le. Jelölje



9.7.2. ábra: A $h = 0.01$ lépésközű explicit Euler-módszer



9.7.3. ábra: A $h = 0.001$ lépésközű explicit Euler-módszer

$x(t)$ egy nyúlpopuláció méretét a t időpontban, míg $y(t)$ egy rókapopuláció méretét! Feltesszük, hogy ismerjük a populációk kezdeti méretét, azaz az $x(0)$ és $y(0)$ értékeket. Ekkor a matematikai modellünk felírható a következő differenciálegyenlet-rendszer segítségével:

$$x'(t) = ax(t) - bx(t)y(t)$$

$$y'(t) = cx(t)y(t) - dy(t),$$

ahol a a nyúlak növekedési rátája, d a rókapopuláció halálozási aránya. Mikor egy róka és egy nyúl találkozik, akkor a nyulak száma csökken. A nyulak számának csökkenési sebessége arányos azzal, hogy egy időegység alatt hány róka és nyúl találkozik, azaz az x(t)y(t) szorzattal. Ezt fejezik ki a b illetve c együtthatók. A modell kapcsán elvárásaink a következők. Ha nincs ragadozó, akkor feltesszük, hogy a zsákmány populáció nagysága exponenciális növekedésű, azaz $x' = ax$ dinamikájú, ahol a a zsákmány populáció növekedési rátája. Feltesszük továbbá, hogy zsákmány hiányában a ragadozók kihalnak az $y' = -dy$ egyenletdinamikája szerint, ahol d a ragadozó populáció halálozási aránya.

A továbbiakban a MATLAB segítségével numerikusan oldjuk meg a fenti egyenletet, és numerikus módszerként ismét az explicit Euler-módszert választjuk.

Tekintsük most egy olyan példát, amely általánosítása az előző modellnek abban az értelemben, hogy az együtthatók akár függvények is lehetnek.

$$x' = x - 2x^2 - xy$$

$$y' = -2y + 6xy,$$

legyen $x(0) = 1$, $y(0) = 0.1$.

Készítsük el az alábbi m-fájlt.

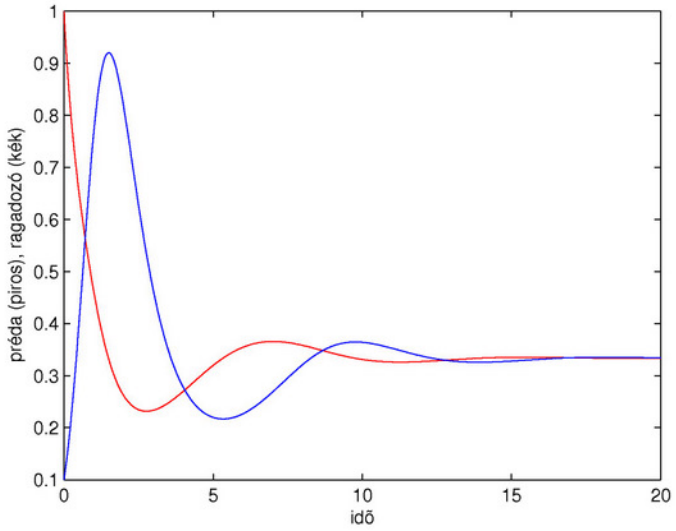
```
function LVexpeuler(x,y,T,N)
h=T/N;
for n=1:N
u=f(x,y); v=g(x,y);
x=x+h*u; y=y+h*v;
xhistory=[xhistory,x]; yhistory=[yhistory,y];
end
t=0:h:T;
plot(t,xhistory,'red', t, yhistory,'blue')
xlabel('idő'), ylabel('préda (piros), ragadozó (kék)')
```

```
   function U=f(x,y)
U=x-2*x.*x-x.*y;
function V=g(x,y)
V=-2*y+6*x.*y;
```

Itt bemenő paraméterként a következőket adjuk meg: x: a zsákmány kezdeti száma, y: a ragadozók kezdeti száma, T: a vizsgált időintervallum és N: az osztásrészek száma.

Az `LVexpeuler(1,0.1,20,1000)` utasítással számoljuk ki az eredményt. Futtatás után a 9.7.4. grafikont kapjuk.

Ahogy az elvárható, a populáció vagy kihal, vagy egyensúlyba kerül. A grafikonon is jól látszik, hogy a kezdeti időpontokban sok ragadozó van és nincsenek zsákmányállatok, ezután viszont a zsákmányállatok száma nő, és az idő múlásával egyensúlyba kerülnek a populációk méretei.



9.7.4. ábra: A Lotka-Volterra-modell megoldása explicit Euler-módszerrel

Megjegyezzük, hogy a többi egylépéses módszer MATLAB programja hasonlóan elkészíthető, de ezeknél lépésenként egy (általában nemlineáris) egyenlet megoldásának az algoritmusát is be kell építenünk.

Mint azt már említettük, léteznek a MATLAB programrendszerben beépített programok, amelyek alkalmasak akár nagy pontossággal és hatékonyan megoldani a kezdetiérték-feladatokat. Ilyen például a gyakran alkalmazott ODE45 rutin, ami egy ún. beágyazott Dormand-Prince-módszer. Ez egy egylépéses, váltakozó lépésközű módszer, amelynek paramétereit a 9.7.1. táblázat tartalmazza. [23] A módszer lényege, hogy kiszámol egy negyed- és egy ötödrendű Runge-Kutta-módszert, és úgy választ lépésközt, hogy a hiba a negyedrendű módszer hibája legyen. A következő táblázatban láthatjuk a módszer Butcher-tábláját. Az első σ sor a negyedrendű módszer, a második pedig az ötödrendű módszer súlyfüggvénye. Az ODE45 rutint ugyanolyan módon hívjuk meg, mint a korábban leírt, saját magunk által írt programokat. Nevezetesen, `[T1, Y45] = ode45(@diffegy,[kezdőidő, végidő])`. Itt is két kimenő paraméter van, az idő és a közelítő értékek vektora. A bemenő paraméterek rendre: az alfüggvény, ami leírja a differenciálegyenletünket, az idő vektor, hogy melyik időpontokban számoljuk ki a megoldást, és a kezdeti-érték vektor.

9.7.1. megjegyzés. Megjegyezzük, hogy megadható egy negyedik (opcionális) paraméter is, amelynek segítségével beállíthatjuk az integrálási paramétereket. Itt adhatjuk meg azt a paramétert is, amely a módszer pontosságát szabályozza. Tehát a rutin meghívásánál megadott, rácshálóra vonatkozó paraméter nem azt a rácshálót határozza meg, amin a numerikus módszerünket alkalmazzuk, hanem csak azokat a pontokat, ahol kiértékeljük a numerikus megoldást.

[23] A beágyazott Runge-Kutta-módszerek lényege, hogy két olyan különböző Runge–Kutta-módszert választunk, amelyek Butcher-táblázatában ugyanazon $\mathbf{a}$ vektor és $\mathbf{B}$ mátrix szerepel, viszont eltérőek a $\boldsymbol{\sigma}$ súlyfüggvények, és emiatt a módszerek rendje is különbözik. Erről, illetve a változó lépéshosszúság megválasztásáról olvashatunk a http://www.cs.elte.hu/blobs/diplomamunkak/bsc_alkmat/2010/molnar_viktoria.pdf linken található dolgozatban.

Ezért tehát ezen rácsháló h lépésközének csökkentése önmagában nem eredményezi a módszer pontosságának csökkenését. Ugyanakkor a gyakorlatban elegendő az alapbeállítás, és ezért az alapértelmezést csak indokolt esetekben célszerű megváltoztatni. Ezt az `odeset` rutinnal hajthatjuk végre, amelynek részletei a MATLAB help leírásában megtalálhatók. ⋄

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$\frac{-56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$\frac{-25360}{2187}$	$\frac{64448}{6561}$	$\frac{-212}{729}$			
1	$\frac{9017}{3168}$	$\frac{-355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$\frac{-5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$\frac{-2187}{6784}$	$\frac{11}{84}$	
	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$\frac{-92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$\frac{-2187}{6784}$	$\frac{11}{84}$	0

9.7.1. táblázat: A beágyazott Dormand-Prince RK-módszer paraméterei az ode45 rutinban.

Alkalmazzuk ODE45 módszert a szokásos tesztfeladatunkon, majd írassuk ki eredményeinket a $h = 0.1$ lépéstávolságú rácsháló pontjaiban. Eredményeinket a 9.7.2. táblázat tartalmazza.

Egy másik, ugyancsak gyakran használt beépített módszer az ODE23 ugyancsak beágyazott Runge-Kutta típusú módszer, amelyet Bogacki-Shampine-módszernek is nevezünk. Ennek

t_i	a pontos megoldás	a numerikus megoldás	a hiba
0	1.0000	1.0000	0
0.1000	1.0048	1.0048	$2.9737e-010$
0.2000	1.0187	1.0187	$5.3815e-010$
0.3000	1.0408	1.0408	$7.3041e-010$
0.4000	1.0703	1.0703	$8.8120e-010$
0.5000	1.1065	1.1065	$9.9668e-010$
0.6000	1.1488	1.1488	$1.0822e-009$
0.7000	1.1966	1.1966	$1.1424e-009$
0.8000	1.2493	1.2493	$1.1814e-009$
0.9000	1.3066	1.3066	$1.2026e-009$
1.0000	1.3679	1.3679	$1.2090e-009$

9.7.2. táblázat: Az ode45 eredményei a $h = 0.1$ lépésközű rácsháló pontjaiban.

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{3}{4}$	0	$\frac{3}{4}$		
1	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	
	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	0
	$\frac{7}{24}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{8}$

9.7.3. táblázat: A beágyazott Bogacki-Shampine RK-módszer paraméterei az ode23 rutinban

t_i	a pontos megoldás	a numerikus megoldás	a hiba
0	1.0000	1.0000	0
0.1000	1.0048	1.0048	$4.0847e-006$
0.2000	1.0187	1.0187	$7.3920e-006$
0.3000	1.0408	1.0408	$1.0033e-005$
0.4000	1.0703	1.0703	$1.2104e-005$
0.5000	1.1065	1.1065	$1.3690e-005$
0.6000	1.1488	1.1488	$1.4865e-005$
0.7000	1.1966	1.1966	$1.5692e-005$
0.8000	1.2493	1.2493	$1.6227e-005$
0.9000	1.3066	1.3066	$1.6518e-005$
1.0000	1.3679	1.3679	$1.6607e-005$

9.7.4. táblázat: Az ode23 eredményei a $h = 0.1$ lépésközű rácshálón

meghívása az ODE45 módszerrel megegyezően a `[T1, Y23] = ode23(@diffegy, [kezdőidő, végidő])` utasítás. A 9.7.3. táblázat tartalmazza a módszer Butcher-tábláját.

Ez egy explicit $(2, 3)$-típusú Runge-Kutta-módszer. Igazából akkor hatékony, amikor olcsón szeretnénk kevésbé pontos megoldást kapni. (Általában nem merev, vagy csak nagyon kis merevségű feladatokra alkalmazzuk.) Most is a szokásos tesztfeladatunkra futtassuk le az ODE23 algoritmust, majd írassuk ki eredményeinket a $h = 0.1$ lépéstávolságú rácsháló pontjaiban. Eredményeinket a 9.7.4. táblázat tartalmazza.

A többlépéses módszerek algoritmusai is megtalálhatók a MATLAB-ban. Ilyen az ODE113, amelynek a rendje 1-től 13-ig változhat és az Adams-Bashforth-Moulton-módszeren alapszik. Összehasonlítva az ODE45 módszerrel, kevésbé pontos, de kisebb számítási igényű, és különösen előnyös, amikor az f függvény kiértékelése költséges. Megjegyezzük, hogy ez a módszer is `[T1, Y113] = ode113(@diffegy,[kezdőidő, végidő])` szintaktikájú, és tipikusan nem merev feladatokra alkalmazzuk. A módszer pontosságát a szokásos tesztfeladatunkon a $h = 0.1$ lépéstávolságú rácshálón a 9.7.5. táblázat tartalmazza.

t_i	a pontos megoldás	a numerikus megoldás	a hiba
0	1.0000	1.0000	0
0.1000	1.0048	1.0048	$6.2300e-006$
0.2000	1.0187	1.0187	$1.8714e-005$
0.3000	1.0408	1.0408	$2.7885e-005$
0.4000	1.0703	1.0703	$2.1933e-005$
0.5000	1.1065	1.1065	$1.8889e-005$
0.6000	1.1488	1.1488	$1.7254e-005$
0.7000	1.1966	1.1966	$1.5668e-005$
0.8000	1.2493	1.2493	$1.4228e-005$
0.9000	1.3066	1.3066	$1.2872e-005$
1.0000	1.3679	1.3679	$1.1643e-005$

9.7.5. táblázat: Az ode113 eredményei a $h = 0.1$ lépésközű rácshálón.

Az alábbi táblázatban foglaljuk össze néhány olyan módszer pontosságát, amelyet ebben a szakaszban tárgyaltunk. (Módszereinket a szokásos (9.3.13) tesztfeladatra a $t^* = 1$ időpontban hasonlítjuk össze.)

módszer	e_{n^*}	
	$h_1 = 0.1$	$h_2 = 0.01$
explicit Euler	1.9201e-002	1.8471e-003
javított Euler	6.6154e-004	6.1775e-006
implicit Euler	2.1537e-002	1.8687e-003
ODE45	1.2090e-009	1.0903e-009
ODE23	1.6607e-005	1.5087e-005

Megjegyezzük, hogy a MATLAB saját programjai érdemben nem csökkentek a rácsháló finomodásával, amely a 9.7.1. megjegyzésben leírt okok miatt törvényszerű.

Befejezésül megjegyezzük, hogy merev feladatok megoldására is vannak rutinok a MATLAB-ban. Ilyenek a ODE15S (amelynek meghívása `[T1, Y15s] = ode15s(@diffegy, T1, 1)`), ODE23S (amelynek meghívása `[T1, Y23s] = ode23s(@diffegy, T1, 1)`). A ODE123T és az ODE123TB rutinok az enyhén merev rendszerek megoldására javasolhatók. Ezek a módszerek alacsony pontosságúak. Az ODE15S módszer a retrográd differencia módszeren (amelyet Gear-módszernek is szokásos nevezni) alapul (lásd a 9.5.4. szakaszt), és ez is változó lépéshosszúságú módszer. Különösen akkor javasolt ez a módszer, amikor az ODE45 módszer nagyon lassú vagy egyáltalán nem működik. Az ODE23S egy másodrendű, módosított, egylépéses ún. Rosenbrock-módszer. Mivel a módszer egylépéses, ezért általában hatékonyabb, mint az ODE15S módszer.
A módszerek részletei iránt érdeklődőknek javasoljuk a jegyzékben található [1, 34], illetve a MATLAB programok iránt érdeklődöknek a jegyzék [12, 33] irodalmait ajánljuk.

9.8. Feladatok

Közönséges differenciálegyenletek

9.8.1. feladat. Legyen f folytomos a $H = \{(t,u), t \in [-3,3], u \in [-4,4]\}$ halmazon, tovább $|f(t,u)| \leq 7$ a H halmazon Melyik az a legnagyobb intervallum, amelyen az

$$u' = f(t,u) \quad u(0) = u_0$$

feladatnak biztosan létezik megoldása?

9.8.2. feladat. Vizsgáljuk meg, hogy az

$$u' = 0.5 \sin u + t; \quad u(0) = 0$$

feladatnak létezik-e megoldása.

9.8.3. feladat. Mutassuk meg, hogy az

$$u' = te^{-u}, \quad u(0) = 0$$

egyenletnek $t_0 \geq 0$ és $u_0 = 0$ esetén mindig létezik egyértelmű megoldása. Állítsuk elő a megoldást!

9.8.4. feladat. Tekintsük az

$$u' = 1 + u^2, \quad u(0) = 0$$

egyenletet! Mutassuk meg, hogy nem minden $t > 0$ értékre létezik a megoldása! Magyarázzuk meg ennek okát!

9.8.5. feladat. Melyik az a legnagyobb intervallum, ahol a 9.8.4. példa feladatának létezik egyértelmű megoldása? Állítsuk elő ezt a megoldást!

9.8.6. feladat. Bizonyítsuk be, hogy az

$$u' = e^u + t^2, \quad u(0) = 0$$

feladatnak létezik egyetlen megoldása a $t \in [0, 0.351]$ intervallumon!

9.8.7. feladat. Bizonyítsuk be, hogy ha az

$$u' = f(t,u) \quad u(0) = u_0$$

feladatban f folytonos és korlátos a $H = \{(t,u), t \in [a,b], u \in \mathbb{R}\}$ halmazon, akkor a feladatnak létezik megoldása a $t \in [a,b]$ intervallumon!

Egylépéses numerikus módszerek

9.8.8. feladat. Mutassuk meg, hogy az $u(t) = t^2/4$ függvény megoldása az

$$u' = \sqrt{u}, \quad u(0) = 0$$

feladatnak! Az elsőrendű Taylor-sorba fejtéses numerikus módszerrel számoljuk ki a közelítő megoldását! Adjunk magyarázatot arra, hogy miért különbözik a numerikus megoldás a pontos megoldástól!

9.8.9. feladat. Számítsuk ki $u(0.1)$ közelítő értékét az

$$u' = -tu^2, \quad u(0) = 2$$

feladatra a másodrendű Taylor-sorba fejtéses numerikus módszerrel!

9.8.10. feladat. A Cauchy-feladat megoldásával jól kiszámíthaó néhány olyan határozott integrál, ahol a Newton-Leibniz szabály nem alkalmazható. Például az

$$\int_0^2 e^{-s^2} ds$$

integtál értéke meghatározható az

$$u' = e^{-t^2}, \quad u(0) = 2$$

feladat megoldásával a $[0, 2]$ intervallumon. A negyedrendű Taylor-sorba fejtéses numerikus módszerrel határozzuk meg az integrál közelítő értékét! (Az

$$\operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-s^2} ds$$

ún. "hiba-függvény" ("error-function") egy jól definiált folytonos függvény, melynek értékeit csak numerikusan tudjuk kiszámolni. Ezen értékeket általában táblázatokban szokás megadni. (Lásd részeletesebben a http://en.wikipedia.org/wiki/Error_function linket.) Innen a "pontos" értékre $u(2) \approx 0.8820813907$ adódik.)

9.8.11. feladat. Oldjuk meg analitikusan, majd az explicit Euler-módszerrel az $u' = tu^{1/3}$, $u(1) = 1$ feladatot a $[0, 5]$ intervallumon! Először kézi számolással a $h = 1$ lépésközzel, majd a EXPEULER programmal $h = 0.10, 0.05, 0.01$ lépéstávolságokkal! Vizsgáljuk meg a konvergenciát a $t = 1, 2, 3, 4, 5$ pontokban! Figyeljük meg a konvergenciarendet!

9.8.12. feladat. Írjunk programot az implicit Euler-módszerre és oldjuk meg a 9.8.11. feladat példáját a $h = 0.10, 0.05, 0.01$ lépéstávolságokkal. Vizsgáljuk meg a konvergenciát a $t = 1, 2, 3, 4, 5$ pontokban! Figyeljük meg a konvergenciarendet! (Ellenőrizzük, hogy a pontos megoldás az $u(t) = \ln(1 + t^2 + /2)$ függvény!)

9.8.13. feladat. Írjunk programot a θ-módszerre és oldjuk meg a 9.8.11. feladat példáját a $h = 0.10, 0.05, 0.01$ lépéstávolságokkal. Vizsgáljuk meg a konvergenciát a $t = 1, 2, 3, 4, 5$ pontokban! Figyeljük meg a konvergenciarendet a $t = 1$ pontban a $\theta = 0, 0.1, 0.2, \ldots, 0.9, 1$ értékek esetén.

9.8.14. feladat. Számítsuk ki MATLAB segítségével az explicit és az implicit Euler-módszerrel, illetve a Crank–Nicolson módszerrel a 9.8.3. feladat numerikus megoldását a $h = 0.10$, 0.05, 0.01, 0.001 lépéstávolságokkal a $t = 1$ pontban! Figyeljük meg a konvergenciarendet! (Ellenőrizzük, hogy a pontos megoldás az $u(t) = \ln(1 + t^2 + /2)$ függvény!)

9.8.15. feladat. Alkalmazzuk a negyedrendű RK-módszert az $u' = \lambda u$ tesztfeladatra! Mutassuk meg, hogy ekkor az $y_h(t)$ numerikus megoldást jelentő rácsfüggvényre az

$$y_h(t + h) = \left(1 + h\lambda + \frac{1}{2}h^2\lambda^2 + \frac{1}{6}h^3\lambda^3 + \frac{1}{24}h^4\lambda^4\right) y_h(t)$$

összefüggés érvényes! Mutassuk meg, hogy erre a feladatra a lokális approximációs hiba $\mathcal{O}(h^5)$!

9.8.16. feladat. A MATLAB program segítségével oldjuk meg a negyedrendű RK-módszerrel az

$$u' = e^{tu} + \cos(u - t), \quad u(1) = 3$$

feladatot! Használjuk a $h = 0.01$ lépésközt és állítsuk le a számítást a túlcsordulás elött.

9.8.17. feladat. Egy lőfegyverből felfelé lövünk. A golyó $v(t)$ sebességét a

$$v' = -32 - \frac{cv}{m}, \quad v(0) = 1$$

differenciálegyenlet írja le, ahol m a golyó tömege és c a légellenállást jellemző állandó. Legyen a példában $c/m = 2$. Valamelyik numerikus módszerrel határozzuk meg, hogy mennyi idő múlva éri el a golyó pályája legmagasabb pontját!

9.8.18. feladat. Tekintsük az $\mathbf{u}' = \mathbf{A}\mathbf{u} + \mathbf{b}$ közönséges differenciálegyenlet-rendszert, ahol $\mathbf{A} \in \mathbb{R}^{n\times n}$ egy adott kvadratikus mátrix és $\mathbf{b} \in \mathbb{R}^n$ egy adott vektor. Írjuk fel erre az egyenletrendszerre az explicit és implicit Eutel-sémákat és a Crank–Nicolson-sémát!

9.8.19. feladat. Legyen a 9.8.18. feladat jelöléseivel

$$\mathbf{A} = \begin{bmatrix} -10 & 3 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

és legyen a kezdeti feltétel: $y_1(0) = y_2(0) = 1$. Igazoljuk, hogy $y_1(x) = (e^{-10x} + 1)/2$, $y_2(x) = 1$ megoldása a differenciálegyenletnek! Oldjuk meg az egyenletet numerikusan a $[0, 4]$ intervallumon az explicit Euler-módszerrel és a (9.4.18) szerinti javított explicit Euler-módszerrel! Próbálkozzunk $h = 0.2$ körüli értékekkel!

Többlépéses numerikus módszerek

9.8.20. feladat. Határozzuk meg az egylépéses, elsőrendű Adams-Moulton formulát! Mi a kapcsolata a trapézszabállyal?

9.8.21. feladat. Ellenőrizzük a 9.5.2. táblázat szerinti negyedrendű Adams-Moulton módszer képletét, azaz hogy a

$$y_{n+1} = \frac{h}{24}\left(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}\right)$$

módszer valóban negyedrendű!

9.8.22. feladat. Ellenőrizzük a 9.5.1. táblázat szerinti negyedrendű Adams-Bashfort módszer képletét, azaz hogy a

$$y_{n+1} = \frac{h}{24}\left(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}\right)$$

módszer valóban negyedrendű!

9.8.23. feladat. Számítsuk ki $u(0.1)$ közelítő értékét az

$$u' = -tu^2, \quad u(0) = 2$$

feladatra a két- és háromlépéses Adams-Bashforth és Adams-Moulton formulákkal! (A kezdőértékekre válasszunk megfelelő RK-módszert!)

9.8.24. feladat. Tekintsük az $y_n - y_{n-2} = h\left(f_n - 3f_{n-1} + 4f_{n-2}\right)$ kétlépéses módszert! Vizsgáljuk meg a konzisztenciáját és a stabilitását!

9.8.25. feladat. Tekintsük az $y_n - 2y_{n-1} + y_{n-2} = h\left(f_n - f_{n-1}\right)$ kétlépéses módszert! Vizsgáljuk meg a módszer konvergenciáját!

Ellenőrző kérdések

1. Mit nevezünk közönséges differenciálegyenletnek?
2. Mi a szerepe a kezdeti feltételnek?
3. Mit nevezünk Lipschitz-féle tulajdonságnak és mi a kapcsolata a Cauchy-feladat megoldhatóságával?
4. Adja meg az explicit Euler-módszer algoritmusát!
5. Mi a Taylor-soros módszer? Sorolja fel a módszer előnyeit és hátrányait!
6. Mutassa meg, hogy az explicit Euler-módszer konvergens!
7. Definiálja az implicit Euler-módszert! Hasonlítsa össze az explicit Euler-moódszerrel!
8. Mit nevezünk θ-módszernek? Miért nevezetes a $\theta = 0.5$ megválasztású módszer?
9. Mikor nevezünk egy numerikus módszert konvergensnek? Mi a konzisztencia és a stabilitás? Mi a kapcsolat közöttük?
10. Mit nevezünk Runge-Kutta típusú módszernek?
11. Mi a Butcher-táblázat?
12. Mit nevezünk explicit, implicit és diagonálisan implicit Runge-Kutta típusú módszernek? Hasonlítsa össze ezek algoritmusait!
13. Mit nevezünk lineáris többlépéses módszernek?
14. Hogyan választhatók meg a megfelelő kezdeti közelítések a lineáris többlépéses módszerekre?
15. Mit nevezünk Adams-típusú módszernek? Milyen tulajdonságúak az Adams-Bashforth- és az Adams-Moulton-módszerek?
16. Mit nevezünk egy lineáris többlépéses módszer első illetve második karakterisztikus polinomjávak? Mi a kapcsolata a konzisztenciával?
17. Mit nevezünk merev feladatnak? Hogyan kezeljük ezeket numerikusan?
18. Mit nevezünk retrográd differencia módszernek és mire alkalmazhatjuk ezeket?
19. Az $u' = \lambda u$ tesztegyenletben miért engedjük meg a λ komplex értékét is?
20. Milyen MATLAB programokat ismer a kezdetiérték-feladatok megoldására?
21. Milyen alapon működnek a beépített MATLAB programok?

10. A közönséges differenciálegyenletek peremérték-feladatainak numerikus módszerei

Ebben a fejezetben a közönséges differenciálegyenletek peremérték-feladatainak elméleti összefoglalása után a feladatok numerikus megoldási módszereivel foglalkozunk. Ismertetjük a legtipikusabb módszereket: a belövéses módszert és a véges differenciák módszerét. A módszereket számítógépes (MATLAB segítségével készített) programikkal illusztráljuk.

10.1. Bevezetés

Láttuk, hogy a közönséges differenciálegyenletek megoldásának egyértelműségéhez kiegészítő feltételek megadása szükséges. Az előző fejezetben ezek a feltételek a megoldás valamely kezdeti ($t = 0$) időpontban való tulajdonságai voltak. Például, az

$$u'' = f(t, u, u') \tag{10.1.1}$$

másodrendű differenciálegyenlethez az

$$u(0) = u_0; \quad u'(0) = u'_0 \tag{10.1.2}$$

feltételeket adtuk meg, ahol u_0, u'_0 adott számok. Ugyanakkor gyakori eset, amikor a megoldást a $[0, T]$ korlátos időintervallumon vizsgáljuk, és a megoldást ismerjük ezen időintervallum mindkét végpontjában, vagyis a (10.1.1) feladat megoldására az

$$u(0) = u_0, \quad u(T) = u_1 \tag{10.1.3}$$

kiegészítő feltételeket adjuk meg.

10.1.1. példa. Tegyük fel, hogy egy rögzített földfelszíni pontból valamely irányba kilövünk egy ágyúgolyót. Jelölje $y(t)$ egy kilőtt golyó földtől mért magasságát, $x(t)$ pedig a kilövési ponttól mért vízszintes távolságát a $t \geq 0$ időpontban. Feltesszük, hogy a golyó mozgására csak a gravitáció hat, aminek következtében

- vízszintes (x) irányban állandó sebességgel halad a golyó;
- a függőleges (y) irányú mozgására csak a gravitáció hat.

Határozzuk meg, hogy milyen szögben kell kilőni a golyót ahhoz, hogy egy előre rögzített $x = L$ pontban érjen földet!

Ekkor Newton második törvénye szerint a mozgást leíró egyenletek

$$\begin{aligned} \ddot{x}(t) &= 0 \\ \ddot{y}(t) &= -g. \end{aligned} \tag{10.1.4}$$

Emellett $x(0) = 0$ és $y(0) = 0$. (A $t = 0$ kezdeti időpontban sem vízszintesen, sem függőlegesen nem távolodott el a golyó a kezdeti helyzetből.) Ezért a kezdeti feltétel figyelembevételével az első egyenlet megoldása $x(t) = vt$, ahol v az állandó vízszintes irányú sebesség. Innen $t = x/v$. Bevezetve az $y(t) = y(x/v) =: Y(x)$ függvényt, az összetett függvény deriválási szabálya alapján

$$\begin{aligned} \dot{y}(t) &= \frac{dY}{dx}\frac{dx}{dt} = \frac{dY}{dx}v, \\ \ddot{y}(t) &= v\frac{d^2Y}{dx^2}v = v^2\frac{d^2Y}{dx^2}. \end{aligned} \tag{10.1.5}$$

Ekkor tehát feltételeink alapján az ismeretlen új függvény az alábbi tulajdonságokkal rendelkezik:

$$\begin{aligned} Y''(x) &= -\frac{g}{v^2}, \quad x \in (0, L) \\ Y(0) &= 0, \quad Y(L) = 0. \end{aligned} \tag{10.1.6}$$

Mivel ez a differenciálegyenlet könnyen kiintegrálható, ezért a (10.1.6) feladat megoldása közvetlenül kiszámítható:

$$Y(x) = \frac{gx}{2v^2}(L - x).$$

Ezért a kilövés α szögét a

$$\mathrm{tg}\alpha = Y'(0) = \frac{gL}{2v^2}$$

összefüggésből határozhatjuk meg. $\diamond$

Megjegyezzük, hogy a fenti példában a t változóról x változóra való áttérést azt motiválta, hogy a (10.1.1)(10.1.3) peremérték-feladat az új ismeretlen függvényre nézve valamely korlátos *térbeli tartományon* lett kitűzve.[1]

10.1.2. definíció.

Az $u = u(t)$ $C^2[a, b]$-beli ismeretlen függvényre kitűzött

$$\begin{aligned} u'' &= f(t, u, u'), \quad t \in (a, b), \\ u(a) &= \alpha, \quad u(b) = \beta \end{aligned} \tag{10.1.7}$$

feladatot *peremérték-feladatnak* nevezzük.

Vegyük észre, hogy a definicióban egy skaláris egyenlet szerepel, ellentétben a kezdetiérték-feladatok megfogaémazásával, ahol általánosan vektor-skalár fv-re is definiáltuk a feladatot.

A fejezet felépítése a következő. A következő 10.2. szakaszban egy egyszerű kitűzésű feladaton mutatjuk be a véges differenciás numerikus megoldási módszert. Ezután rátérünk a téma bővebb kifejtésére és egyes részleteinek ismertetésére. Foglalkozunk az általános alakú folytonos feladat megoldhatóságával, majd ismertetjük az ún. belövéses módszert. Ezután a véges differenciás megoldási módszert részletezzük.

[1] A továbbiakban az ismeretlen függvényre az $y(t)$ (vagy az $u(t)$), illetve az $y(x)$ (vagy az $u(x)$) jelöléseket egyaránt használjuk.

10.2. Egy közönséges differenciálegyenlet peremérték-feladatának megoldása véges differenciákkal

Ezen szakasz célja, hogy általános bevezetést adjon a leggyakrabban használt véges differenciás numerikus módszerről illetve annak hátteréről. Ezt a szakaszt úgy állítottuk össze, hogy lényegében független a következő, mélyebb ismereteket nyújtó 10.5. szakasztól. A könnyebb olvashatóság érdekében a későbbi szakaszokban megismételjük azokat a fogalmakat, amelyeket ebben a szakaszban ismertetünk, tehát a 10.3.-10.6. szakaszok akár ezen szakasz olvasása nélkül is megérthetők[2].

10.2.1. A véges differenciás séma felépítése

Tekintsük a

$$\begin{aligned} -u'' + cu &= f, \quad x \in (0,l), \\ u(0) &= \alpha, \quad u(l) = \beta \end{aligned} \tag{10.2.1}$$

feladatot, ahol $c \geq 0$ állandó, f egy adott folytonos függvény. Mivel a feladat analitikus megoldását általános esetben nem tudjuk közvetlenül előállítani, ezért numerikus eljárást alkalmazunk. Ennek lényege a következő.

1. Definiálunk a $[0,l]$ intervallumon *rácshálókat*, nevezetesen az $\omega_h = \{x_i = ih,\ i = 1,2,\dots,N-1,\ h = l/N\ \}$ és az $\overline{\omega}_h = \{x_i = ih,\ i = 0,1,\dots,N,\ h = l/N\ \}$ rácshálókat. Jelölje $\gamma_h = \overline{\omega}_h \backslash \omega_h = \{x_0 = 0; x_N = l\}$ az ún. perempontokat.

2. Jelölje $\mathbb{F}(\overline{\omega}_h)$ és $\mathbb{F}(\omega_h)$ az $\overline{\omega}_h$ és az ω_h rácsokon értelmezett, $\mathbb{R}$-be képező függvények vektorterét.

3. Célunk olyan $y_h \in \mathbb{F}(\overline{\omega}_h)$ *rácsfüggvény* meghatározása, amely az $\overline{\omega}_h$ pontjaiban közel van a (10.2.1) feladat u megoldásához, és a rácsháló finomításával (azaz $h \to 0$ esetén) az eltérésük nullához tart.

A numerikus módszert az határozza meg, hogy milyen módon választjuk meg a keresett rácsfüggvényt. Kézenfekvő az alábbi ötlet. Tekintsük a (10.2.1) egyenletet az ω_h rácsháló ponjaiban! Ekkor a

$$-u''(x_i) + cu(x_i) = f(x_i), \quad x_i \in \omega_h \tag{10.2.2}$$

egyenlőségeket kapjuk. Mint azt a numerikus deriválásnál már megismertük, az x_i pontbeli első deriváltakat a

$$u'(x_i) \approx \frac{u(x_i+h) - u(x_i)}{h}, \quad u'(x_i) \approx \frac{u(x_i) - u(x_i-h)}{h} \tag{10.2.3}$$

módon, a második deriváltakat pedig a

$$u''(x_i) \approx \frac{u(x_i+h) - 2u(x_i) + u(x_i-h)}{h^2} \tag{10.2.4}$$

módon közelíthetjük. Mivel elvárásaink szerint a keresett rácsfüggvényre $y_h(x_i) \approx u(x_i)$, ezért (10.2.2) és (10.2.4) alapján felállíthatjuk a következő összefüggéseket:

$$-\frac{y_h(x_i+h) - 2y_h(x_i) + y_h(x_i-h)}{h^2} + cy_h(x_i) = f(x_i), \quad x_i \in \omega_h. \tag{10.2.5}$$

[2] Ebben a részben a független változót x betűvel jelöljük. Ennek oka, hogy a következő, a parciális differenciálegyenletekről szóló fejezetben több helyen is ezen jelölés mellett utalunk ezen szakasz eredményeire.

Mivel a perempontokban ismerjük a megoldást, ezért nyilvánvalóan

$$y_h(x_0) = \alpha, \quad y_h(x_N) = \beta. \tag{10.2.6}$$

A (10.2.5)-(10.2.6) feladat felírható az alábbi kompakt módon. Jelölje $L_h : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\overline{\omega}_h)$ azt az operátort, amely a következő módon rendeli hozzá a $w_h \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvényhez az $L_h w_h \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvényt:

$$(L_h w_h)(x_i) = \begin{cases} -\dfrac{w_h(x_i+h) - 2w_h(x_i) + w_h(x_i-h)}{h^2} + c w_h(x_i), & \text{ha } x_i \in \omega_h; \\ w_h(x_i), & \text{ha } x_i \in \gamma_h. \end{cases} \tag{10.2.7}$$

Legyen $b_h \in \mathbb{F}(\overline{\omega}_h)$ a következő rácsfüggvény:

$$b_h(x_i) = \begin{cases} f(x_i), & \text{ha } x_i \in \omega_h; \\ \alpha, & \text{ha } x_i = x_0, \\ \beta, & \text{ha } x_i = x_N. \end{cases} \tag{10.2.8}$$

Ezen jelölések mellett a (10.2.5)-(10.2.6) feladat nem más, mint azon $y_h \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvény meghatározása, amelyet az L_h operátor az adott $b_h \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvénybe képez le, azaz feladatunk az

$$L_h y_h = b_h \tag{10.2.9}$$

operátoregyenlet megoldása. Vezessük be az

$$y_h(x_i) = y_i, \ f(x_i) = f_i, \quad b_h(x_i) = b_i$$

jelöléseket! Ezen jelölésekkel a (10.2.5)-(10.2.6) feladat felírható a következő alakban:

$$\begin{aligned} -\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + c y_i &= f_i, \quad i = 1, 2, \ldots, N-1, \\ y_0 = \alpha, \quad y_N &= \beta. \end{aligned} \tag{10.2.10}$$

Ez egy $N+1$ ismeretlenes lineáris algebrai egyenletrendszert jelent, amely felírható

$$\mathbf{L}_h \mathbf{y}_h = \mathbf{b}_h \tag{10.2.11}$$

alakban, ahol $\mathbf{y}_h = [y_0, y_1, \ldots, y_N]^T$ az ismeretlen vektor, $\mathbf{b}_h = [\alpha, f_1, \ldots, f_{N-1}, \beta]^T$ adott vektor[3], és $\mathbf{L}_h$ a következő mátrix

$$\mathbf{L}_h = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ -\frac{1}{h^2} & \frac{2}{h^2}+c & -\frac{1}{h^2} & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{h^2} & \frac{2}{h^2}+c & -\frac{1}{h^2} & \ldots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 0 & -\frac{1}{h^2} & \frac{2}{h^2}+c & -\frac{1}{h^2} \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 1 \end{pmatrix} \tag{10.2.12}$$

10.2.2. A véges differenciás séma megoldhatósága és tulajdonságai

A továbbiakban alapvető fontosságú az $\mathbf{L}_h$ mátrix következő tulajdonsága.

[3] Az egyszerűbb jelölés kedvéért a vektorokat **vastagon**, a föléhúzás elhagyásával jelöljük.

10.2.1. tétel.
A (10.2.12) alakú $\mathbf{L}_h$ mátrix M-mátrix.

Bizonyítás. Mint az ismeretes (lásd 1.2.39. tétel), azt kell megmutatnunk, hogy létezik olyan $\mathbb{R}^{N+1}$-beli $\mathbf{g}_h > 0$ vektor, amelyre $\mathbf{L}_h\mathbf{g}_h > 0$. Legyen a $\mathbf{g}_h$ vektor i-edik eleme ($i = 0, 1, \ldots, N$) [4]

$$g_{h,i} = 1 + ih(l - ih). \tag{10.2.13}$$

Ekkor $\mathbf{g}_{h,i} \geq 1$ és $(\mathbf{L}_h\mathbf{g}_h)_0 = (\mathbf{L}_h\mathbf{g}_h)_N = 1$. Egyszerű behelyettesítéssel ellenőrizhető, hogy $i = 1, 2, \ldots, N-1$ esetén

$$-g_{h,i-1} + 2g_{h,i} - g_{h,i+1} = 2h^2.$$

Így

$$(\mathbf{L}_h\mathbf{g}_h)_i = 2 + c(1 + ih(l - ih)), \quad i = 1, 2, \ldots, N-1.$$

Ez azt jelenti, hogy $(\mathbf{L}_h\mathbf{g}_h)_i \geq 1$ minden $i = 0, 1, 2, \ldots, N-1, N$ index esetén. Összefoglalóan: az $\mathbf{e} = [1, 1, \ldots, 1]^\top \in \mathbb{R}^{N+1}$ jelöléssel

$$\mathbf{g}_h \geq \mathbf{e} > 0, \text{ és } \mathbf{L}_h\mathbf{g}_h \geq \mathbf{e} > 0. \tag{10.2.14}$$

A (10.2.14) reláció bebizonyítja az állításunkat. ■

10.2.2. következmény. Az $\mathbf{L}_h$ mátrix minden $h > 0$ mellett invertálható, $\mathbf{L}_h^{-1} \geq 0$, és az 1.2.40. tétel valamint (10.2.14) következtében inverzének maximumnormája felülről becsülhető az alábbi módon.

$$\|\mathbf{L}_h^{-1}\|_\infty \leq \frac{\|\mathbf{g}_h\|_\infty}{\min_i (\mathbf{L}_h\mathbf{g}_h)_i} = \frac{\max_i g_{h,i}}{1}. \tag{10.2.15}$$

A számtani-mértani közepek közötti összefüggés alapján

$$ih(l - ih) \leq \left(\frac{ih + (l - ih)}{2}\right)^2 = \frac{l^2}{4},$$

és így $g_{h,i} \leq 1 + l^2/4$. Ezért (10.2.15) alapján érvényes az

$$\|\mathbf{L}_h^{-1}\|_\infty \leq K := \frac{l^2 + 4}{4} \tag{10.2.16}$$

becslés. ◇

10.2.3. A véges differenciás módszer konvergenciája

Legyen $P_h : C([0, l]) \to \mathbb{F}(\overline{\omega}_h)$ *projekciós* operátor, azaz $(P_h u)(x_i) = u(x_i)$ minden $x_i \in \overline{\omega}_h$ pontban. Jelölje $e_h \in \mathbb{F}(\overline{\omega}_h)$ az $e_h = y_h - P_h u$ egyenlőséggel definiált ún. *hibafüggvényt*. Ekkor tehát

$$e_h(x_i) = y_h(x_i) - u(x_i) = y_i - u(x_i). \tag{10.2.17}$$

[4] Az egyszerűbb jelölés kedvéért a koordináták indexelését nullától indítjuk.

10.2.3. definíció.

Az L_h rácsoperátorral meghatározott numerikus módszert *a maximumnormában konvergensnek* nevezzük, ha

$$\lim_{h\to 0} \|e_h\|_\infty = 0. \tag{10.2.18}$$

Ha $\|e_h\|_\infty = \mathcal{O}(h^p)$ valamely $p \geq 1$ egész számmal, akkor a módszert *p-ed rendben konvergensnek* nevezzük.

A továbbiakban megmutatjuk, hogy az előző pontban definiált (10.2.9) numerikus módszer konvergens, továbbá meghatározzuk konvergenciájának rendjét is.
A (10.2.17) összefüggésből, az $e_h(x_i) = e_i$ egyszerűsítő jelöléssel $y_i = e_i + u(x_i)$. Ezt behelyettesítve a (10.2.10) sémába a következő egyenletrendszert nyerjük:

$$\begin{gathered} -\frac{e_{i+1} - 2e_i + e_{i-1}}{h^2} + ce_i = \Psi_i^h, \quad i = 1, 2, \ldots, N-1, \\ e_0 = 0, \quad e_N = 0, \end{gathered} \tag{10.2.19}$$

ahol

$$\Psi_i^h = f_i + \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} - cu(x_i). \tag{10.2.20}$$

Mivel $f_i = f(x_i) = -u''(x_i) + cu(x_i)$, ezért

$$\Psi_i^h = \left(\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} - u''(x_i) \right) + \underbrace{(cu(x_i) - cu(x_i))}_{=0}. \tag{10.2.21}$$

Mint ismeretes, ekkor

$$\Psi_i^h = \mathcal{O}(h^2). \tag{10.2.22}$$

(A vezető konstans kiírásával $\Psi_i^h = (M_4/12)\, h^2 + \mathcal{O}(h^4)$, ahol $M_4 = \max_{[0,l]} |u^{(4)}|$.) Jelölje $\Psi^h \in \mathbb{F}(\overline{\omega}_h)$ azt a rácsfüggvényt, amelyre $\Psi^h(x_0) = \Psi^h(x_N) = 0$, míg az ω_h rácsháló pontjaiban a (10.2.21) szerint definiált Ψ_i^h értékeket veszi fel. Ekkor $\|\Psi^h\|_\infty = \mathcal{O}(h^2)$. Vegyük észre, hogy a (10.2.19) hibaegyenlet felírható

$$\mathbf{L}_h \mathbf{e}_h = \Psi^h \tag{10.2.23}$$

alakban, ahol $\mathbf{L}_h$ a (10.2.12) alakú mátrix, $\mathbf{e}_h$ pedig az e_h hibafüggvénynek megfelelő $\mathbb{R}^{N+1}$-beli vektor. Mivel $\mathbf{L}_h$ reguláris, ezért $\mathbf{e}_h = \mathbf{L}_h^{-1}\Psi^h$. Innen, felhasználva a (10.2.16) és a (10.2.22) egyenlőtlenségeket,

$$\|\mathbf{e}_h\|_\infty \leq \|\mathbf{L}_h^{-1}\|_\infty \|\Psi^h\|_\infty \leq K \cdot \mathcal{O}(h^2) = \mathcal{O}(h^2). \tag{10.2.24}$$

Ezért $\lim_{h\to 0} \|\mathbf{e}_h\|_\infty = 0$. Ezzel beláttuk az alábbi állítást.

10.2.4. tétel.

Tegyük fel, hogy a (10.2.1) feladat $u(x)$ megoldása négyszer folytonosan differenciálható. Ekkor a (10.2.5)-(10.2.6) (avagy a vele ekvivalens (10.2.9) operátoregyenlet) által előállított véges differenciás numerikus megoldás a maximumnormában másodrenben konvergál az $u(x)$ megoldáshoz.

10.2.5. megjegyzés. A 10.2.4. tétel szerint a közelítés másodrendben konvergens, emellett a (10.2.24) becslésben szereplő $\mathcal{O}(h^2) = const. \cdot h^2$ kifejezésben az állandó értéke $\frac{M_4(l^2+4)}{48}$. $\diamond$

A (10.2.21) összefüggés alapján $\mathbf{\Psi}_i^h$ jelentése a következő: megmutatja, hogy az $u(x)$ pontos megoldás rácspontbeli értékei milyen pontosan elégítik ki egy $x_i \in \overline{\omega}_h$ rácspontban a numerikus módszer sémáját. Az előző fejezetben láttuk, hogy a finomodó rácshálók sorozatán ez a numerikus megoldás adott pontbeli viselkedésére adhat választ, ugyanakkor az egész intervallumon való viselkedésére nem. Míg a kezdetiérték-feladatok esetén a pontbeli illetve az egész intervallumon való viselkedés jellemzésére egyaránt kerestük a választ, addig a peremérték-feladatok esetén csak a második eset a tipikus.[5] Tehát arra vagyunk kíváncsiak, hogy a numerikus megoldás a $[0, l]$ intervallumon hogyan közelíti a (10.2.1) feladat $u(x)$ megoldásfüggvényét. Ehhez megvizsgáljuk, hogy az $\overline{\omega}_h$ rácsháló pontjai összességében hogyan viselkednek a $\mathbf{\Psi}_i^h$ értékek. Ezért a módszer pontbeli approximációs tulajdonságát a $\mathbf{\Psi}_i^h$ koordinátájú $\mathbf{\Psi}^h$ vektorral jellemezzük.

10.2.6. definíció.

Azt mondjuk, hogy a numerikus módszer *konzisztens* a maximum normában, ha $\lim_{h\to 0} \|\mathbf{\Psi}^h\|_\infty = 0$. Ha $\|\mathbf{\Psi}^h\|_\infty = \mathcal{O}(h^p)$ $(p \geq 1)$, akkor a módszert p-ed rendben konzisztensnek nevezzük.

Korábbi számításaink alapján tehát a (10.2.9) módszer a maximumnormában másodrendben konzisztens.

Mint azt a 10.2.4. tétel bizonyítása is mutatja, általában a konzisztencia önmagában nem elegendő a konvergencia bizonyításához. Ehhez egy másik tulajdonság is szükséges.

10.2.7. definíció.

Azt mondjuk, hogy a numerikus módszer a maximumnormában *stabil*, ha a módszert leíró $\mathbf{L}_h$ lineáris operátorok (mátrixok) mindegyike invertálható, és megadható olyan $K > 0$, h-tól független állandó, amelyre

$$\|\mathbf{L}_h^{-1}\|_\infty \leq K. \tag{10.2.25}$$

10.2.8. megjegyzés. Gyakran a fenti tulajdonságok csak megfelelően kis h értékek mellett érvényesek, azaz csak valamely $h_0 > 0$ szám melletti $h < h_0$ mellett teljesülnek a megkövetelt tulajdonságok. Ebben az esetben a módszert *feltételesen stabilnak* nevezzük. Ha h megválasztására nincs korlát, azaz az invertálhatóság és a (10.2.25) tulajdonság minden $h > 0$ szám esetén érvényes, akkor a sémát *feltétel nélkül stabilnak* nevezzük. ◇

Közvetlenül belátható a numerikus módszerek egyik alaptétele. (Ezt az állítást később bebizonyítjuk, és több alkalommal is alkalmazzuk.)

10.2.9. tétel.

Egy konzisztens és stabil numerikus módszer konvergens, és a konvergencia rendje megegyezik a konzisztencia rendjével.

10.2.4. Összefoglalás

Foglaljuk össze a szakasz eddigi eredményeit!

[5] Ennek egyik oka, hogy a peremfeltételeket, amelyek az intervallum mindkét végpontjában adottak, fel kell használnunk a numerikus megoldás meghatározásához. Így nem szoritkozhatunk egy rögzített $x^\star \in [0, l]$ pont esetén csak a $[0, x^\star]$ intervallumon generált rácshálósorozatokra.

Jelölje L azt a $C^2[0,l]$-beli függvényeken értelmezett operátort, amely a következő módon hat:

$$(Lv)(x) = \begin{cases} -\dfrac{d^2v}{dx^2}(x) + c(x)v(x), & \text{ha } x \in (0,l); \\ v(x), & \text{ha } x \in \{0;l\}, \end{cases} \tag{10.2.26}$$

továbbá $\tilde{f}$ azt a $[0,l]$ intervallumon értelmezett függvényt, amelyre

$$\tilde{f}(x) = \begin{cases} f(x), & \text{ha } x \in (0,l); \\ \mu_1, & \text{ha } x = 0, \\ \mu_2, & \text{ha } x = l. \end{cases} \tag{10.2.27}$$

Feltesszük, hogy $c(x)$ és $f(x)$ adott folytonos függvények, μ_1 és μ_2 adott számok. Feladatunk az

$$Lu = \tilde{f} \tag{10.2.28}$$

operátoregyenlet megoldása Ehhez az $\{\omega_h\} \subset [0,l]$ rácshálósorozat mindegyik tagján definiálunk egy

$$L_h y_h = b_h \tag{10.2.29}$$

alakú feladatot, ahol $b_h \in \mathbb{F}(\overline{\omega}_h)$ adott, $y_h \in \mathbb{F}(\overline{\omega}_h)$ pedig ismeretlen rácsfüggvény, és $L_h : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\overline{\omega}_h)$ a numerikus módszert leíró, (10.2.7) alakú lineáris operátor. Ekkor a $P_h : C[0,l] \to \mathbb{F}(\overline{\omega}_h)$ projekciós operátor segítségével megmutattuk a következőket.

- A diszkrét feladat másodrendben konzistens, ami azt jelenti, hogy az L_h operátor h^2 pontossággal approximálja az L operátort a (10.2.28) feladat pontos megoldásán. Ez azt jelenti, hogy teljesül az
$$(L_h(P_h u))(x_i) - (Lu)(x_i) = \mathcal{O}(h^2) \tag{10.2.30}$$
reláció minden $x_i \in \overline{\omega}_h$ pontban, és az $\mathcal{O}(h^2) = const. \cdot h^2$ előállításban *const.* független az x_i pont megválasztásától, azaz univerzális állandó a teljes intervallumon.
- A diszkrét feladat stabil, azaz L_h reguláris és létezik olyan $K \geq 0$ állandó, amelyre (10.2.25) érvényes.

Megmutattuk, hogy ezen tulajdonság mellett érvényes a konvergencia, azaz $e_h = P_h u - y_h \in \mathbb{F}(\overline{\omega}_h)$ hibafüggvény a maximumnormában nullához tart. Emellett a konvergencia másodrendű: $\|e_h\|_\infty = \mathcal{O}(h^2)$.

10.3. A közönséges differenciálegyenletek peremérték-feladatának megoldhatósága

A továbbiakban azt vizsgáljuk meg, hogy a (10.1.7) peremérték-feladatnak milyen feltételek mellett létezik egyértelmű megoldása. Emlékeztetünk, hogy az előző fejezetben ezt a kérdést a kezdetiérték-feladatokra is megvizsgáltuk, és megmutattuk, hogy a differenciálegyenlet jobb oldalán szereplő f függvény megválasztása határozza meg a megoldhatóságot és annak egyértelműségét, függetlenül a kezdeti feltétel (avagy magasabb rendű differenciálegyenletek esetén, a kezdeti feltételek) megválasztásától. Az alábbi példa megmutatja, hogy a peremérték-feladatok esetén ez megváltozik, az f függvény mellett bizonyos esetekben a peremfeltételek is kihatnak a megoldás létezésére és annak egyértelműségére.

10.3.1. példa. Legyen a (10.1.3) feladatban $f(t,u,u') = -u$, tehát vizsgáljuk az

$$u'' = -u \tag{10.3.1}$$

egyenletet. Mint ismeretes, (10.3.1) általános megoldása $u(t) = C_1 \sin t + C_2 \cos t$, ahol C_1 és C_2 állandók. Ez utóbbiak meghatározására szolgálnak a (10.1.7) feladatban a $t = a$ és $t = b$ pontokban megadott peremfeltételek.

- Legyen először $a = 0$, $b = \pi/2$, $\alpha = 3$ és $\beta = 7$. Könnyen láthatóan ekkor az egyértelmű megoldás az $u(t) = 7\sin t + 3\cos t$ függvény.
- Legyen most $a = 0$, $b = \pi$, $\alpha = 3$ és $\beta = 7$. (Tehát csak b értékét változtattuk meg.) Behelyettesítve az általános megoldásba láthatóan nincs olyan C_1 és C_2 értékpár, amely mellett ez a peremfeltétel teljesül.

◇

Megjegyezzük, hogy van olyan példa is, amelyben létezik ugyan a peremérték-feladatnak megoldása, de az nem egyértelmű. Például, az

$$\begin{aligned} u'' &= -\exp(u+1), \quad t \in (0,1), \\ u(0) &= 0, \quad u(1) = 0 \end{aligned} \tag{10.3.2}$$

feladatnak megoldása az

$$u(t) = -2\ln \frac{\cosh[(t-0.5)\theta/2]}{\cosh(\theta/4)}$$

függvény[6], ahol θ a $\theta = \sqrt{2e}\cos(\theta/4)$ egyenlet megoldása. Mivel ez utóbbi egyenletnek két megoldása van, ezért a (10.3.2) feladatnak két megoldása létezik.

Az alábbi tétel egy elégséges feltételt ad az egyértemű megoldás létezésére.

10.3.2. tétel.
Tegyük fel, hogy a $T := \{(t,s_1,s_2) : t \in [a,b],\ s_1,s_2 \in \mathbb{R}\}$ jelöléssel a (10.1.7) feladat $f : \mathbb{R}^3 \to \mathbb{R}$ függvényére teljesülnek a következők:

1. $f \in C(T)$,
2. $\partial_2 f, \partial_3 f \in C(T)$,
3. $\partial_2 f > 0$ T-n,
4. létezik olyan $M \geq 0$, amelyre $|\partial_3 f| \leq M$ T-n.

Ekkor a (10.1.7) peremérték-feladatnak létezik egyértelmű megoldása.

A 10.3.2. tétel fontos következménye az alábbi állítás.

10.3.3. következmény. Legyen f lineáris, azaz tekintsük az

$$\begin{aligned} u'' = f(t,u,u') &\equiv p(t)u' + q(t)u + r(t), \quad t \in [a,b], \\ u(a) &= \alpha, \quad u(b) = \beta \end{aligned} \tag{10.3.3}$$

[6] A fenti képletben szereplő cosh a koszinusz hiperbolikus függvényt jelenti, amelyet időnként a ch szimbolummal is szokásos jelölni.

feladatot, ahol $p, q, r \in C[a,b]$ adott folytonos függvények. Ha $q(t) > 0 \quad [a,b]$-n, akkor a (10.3.3) lineáris peremérték-feladatnak létezik egyértelmű megoldása. $\diamond$

Mint azt a 10.3.1. példa is mutatja, a $q(t) > 0$ feltétel nem hagyható el, hiszen ezt a feltételt kivéve a példában szereplő $f(t, s_1, s_2) = -s_1$ függvényre a 10.3.3. következményben szereplő valamennyi feltétel teljesül. Ugyanakkor, mint azt beláttuk, nem létezik megoldás.

Az előző fejezetben megmutattuk, hogy a magasabbrendű differenciálegyenletek átírhatók elsőrendű rendszerek alakjára. Alkalmazva ezt az ún. átviteli elvet, a (10.1.7) peremérték-feladat egyenlete is átírható kétismeretlenes rendszerré. Vezessük be az $\mathbf{u} : [a,b] \to \mathbb{R}^2$, $\mathbf{u}(t) = (u_1(t), u_2(t))$ függvényt a következő módon: $u_1(t) = u(t)$ és $u_2(t) = u'(t)$. Ekkor a feladatunk felírható

$$\begin{aligned} &u_1' = u_2, \quad u_2' = f(t, u_1, u_2), \\ &u_1(a) = \alpha, \quad u_1(b) = \beta \end{aligned} \tag{10.3.4}$$

alakban. Természetesen a (10.3.4) feladat ebben a formában nem oldható meg, hiszen csak az u_1 függvényre ismerünk kiegészítő feltételeket.

Megjegyezzük, hogy (10.3.4) egyenlete speciális alakja a következő általános alakban felírt egyenletnek:

$$\mathbf{u}' = \mathbf{f}(t, \mathbf{u}), \quad t \in [a,b] \tag{10.3.5}$$

ahol $\mathbf{f} : \mathbb{R}^3 \to \mathbb{R}^2$ adott függvény. Ugyanis, ha a (10.3.5) feladatban

$$\mathbf{f}(t, \mathbf{u}) = \mathbf{f}(t, u_1, u_2) = \begin{bmatrix} u_2 \\ f(t, u_1, u_2) \end{bmatrix}, \quad t \in [a,b] \tag{10.3.6}$$

alakban választjuk meg az $\mathbf{f}$ függvényt, akkor éppen a (10.3.4) egyenletrendszerét kapjuk.

10.3.1. A lineáris peremérték-feladat megoldhatósága

Írjuk fel a (10.3.3) lineáris peremérték-feladatot elsőrendű rendszer alakjában! Könnyen láthatóan az egyenlet

$$\mathbf{u}' = \mathbf{A}(t)\mathbf{u} + \mathbf{r}(t) \tag{10.3.7}$$

alakú, ahol

$$\mathbf{A}(t) = \begin{pmatrix} 0 & 1 \\ q(t) & p(t) \end{pmatrix}, \quad \mathbf{r}(t) = \begin{pmatrix} 0 \\ r(t) \end{pmatrix}. \tag{10.3.8}$$

A peremfeltételek felírásához vezessük be a

$$\mathbf{B}_a = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{B}_b = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{10.3.9}$$

jelöléseket. Ekkor a (10.3.3) feladat peremfeltétele

$$\mathbf{B}_a\mathbf{u}(a) + \mathbf{B}_b\mathbf{u}(b) = \mathbf{v} \tag{10.3.10}$$

alakban írható fel. A továbbiakban előállítjuk a (10.3.7)–(10.3.10) feladat megoldását.

A (10.3.7) egyenlet általános megoldása felírható a következő módon. Legyen $\mathbf{Y}(t) \in \mathbb{R}^{m \times m}$ az egyenlet alapmegoldása (más néven fundamentális mátrixa), vagyis az

$$\begin{aligned} &\mathbf{Y}'(t) = \mathbf{A}(t)\mathbf{Y}(t), \quad t \in [a,b] \\ &\mathbf{Y}(a) = \mathbf{I} \end{aligned} \tag{10.3.11}$$

Cauchy-feladat megoldása, ahol $\mathbf{I} \in \mathbb{R}^{m\times m}$ az egységmátrixot jelöli. Ekkor a (10.3.7) egyenlet általános megoldása

$$\mathbf{u}(t) = \mathbf{Y}(t)\left(\mathbf{c} + \int_a^t \mathbf{Y}^{-1}(s)\mathbf{r}(s)ds\right), \tag{10.3.12}$$

ahol $\mathbf{c} \in \mathbb{R}^m$ egy tetszőleges vektor. Célunk $\mathbf{c}$ olyan megválasztása, amely mellett a (10.3.12) szerint definiált $\mathbf{u}(t)$ függvény kielégíti a (10.3.10) peremfeltételeket. Behelyettesítve a (10.3.12) képletet a (10.3.10) feltételbe, a

$$\mathbf{B}_a\mathbf{u}(a) + \mathbf{B}_b\mathbf{u}(b) = \mathbf{v} = \mathbf{B}_a\mathbf{Y}(a)\mathbf{c} + \mathbf{B}_b\mathbf{Y}(b)\left(\mathbf{c} + \int_a^b \mathbf{Y}^{-1}(s)\mathbf{r}(s)ds\right) \tag{10.3.13}$$

feltételt kapjuk. Ebből a $\mathbf{c}$ vektorra rendezve és az $\mathbf{Y}(a) = \mathbf{I}$ feltételt figyelembe véve a

$$\left(\mathbf{B}_a + \mathbf{B}_b\mathbf{Y}(b)\right)\mathbf{c} = \mathbf{v} - \mathbf{B}_b\mathbf{Y}(b)\int_a^b \mathbf{Y}^{-1}(s)\mathbf{r}(s)ds \tag{10.3.14}$$

egyenletet nyerjük. Ezért a

$$\mathbf{Q} = \mathbf{B}_a + \mathbf{B}_b\mathbf{Y}(b) \tag{10.3.15}$$

jelöléssel a $\mathbf{c}$ vektorra a

$$\mathbf{Q}\mathbf{c} = \mathbf{v} - \mathbf{B}_b\mathbf{Y}(b)\int_a^b \mathbf{Y}^{-1}(s)\mathbf{r}(s)ds \tag{10.3.16}$$

feladatot kapjuk. Ezért érvényes az alábbi állítás.

10.3.4. tétel.
A (10.3.7)–(10.3.10) lineáris peremérték-feladatnak pontosan akkor létezik egyértelmű megoldása, amikor a (10.3.15) alakú $\mathbf{Q}$ mátrix reguláris. Emellett a megoldás (10.3.12) alakú, ahol

$$\mathbf{c} = \mathbf{Q}^{-1}\left(\mathbf{v} - \mathbf{B}_b\mathbf{Y}(b)\int_a^b \mathbf{Y}^{-1}(s)\mathbf{r}(s)ds\right). \tag{10.3.17}$$

10.3.5. példa. Vizsáljuk meg az

$$\begin{aligned} u'' &= -u, \quad t \in (0,b) \\ u(0) &= \alpha, \quad u(b) = \beta \end{aligned} \tag{10.3.18}$$

feladat megoldhatóságát! Mivel erre a feladatra

$$\mathbf{A}(t) = \mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{10.3.19}$$

ezért az alapmegoldása az

$$\mathbf{Y}(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \tag{10.3.20}$$

alakú mátrix. Ezért a $\mathbf{B}_a$ és $\mathbf{B}_b$ mátrixok (10.3.9) definíciója alapján a (10.3.15) szerinti $\mathbf{Q}$ mátrix

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 \\ \cos b & \sin b \end{pmatrix} \tag{10.3.21}$$

alakú. Ez a $\mathbf{Q}$ mátrix pontosan akkor szinguláris, amikor $b = j\pi$, ahol $j \in \mathbb{N}$ tetszőleges pozitív egész szám. Tehát pontosan akkor létezik a (10.3.18) feladatnak egyértelmű megoldása, amikor a b végpont a π nem egész számú többszöröse. (Ez az eredmény egyben megválaszolja a 10.3.1. példa eredményét, hogy miért volt egyértelmű megoldás $b = \pi/2$ esetén, és miért nem létezik megoldás a $b = \pi$ megválasztás mellett.) ◇

10.3.6. megjegyzés. Vezessük be a

$$\mathbf{\Phi}(t) = \mathbf{Y}(t)\mathbf{Q}^{-1} \tag{10.3.22}$$

függvényt! Ekkor

$$\mathbf{\Phi}'(t) = \mathbf{Y}'(t)\mathbf{Q}^{-1} = \mathbf{A}(t)\mathbf{Y}(t)\mathbf{Q}^{-1} = \mathbf{A}(t)\mathbf{\Phi}(t).$$

Másrészt

$$\mathbf{B}_a\mathbf{\Phi}(a) + \mathbf{B}_b\mathbf{\Phi}(b) = \left[\mathbf{B}_a\mathbf{Y}(a) + \mathbf{B}_b\mathbf{Y}(b)\right]\mathbf{Q}^{-1} = \mathbf{I}.$$

Ezért a (10.3.7)–(10.3.10) lineáris peremérték-feladat esetén (10.3.22) alapján definiált $\mathbf{\Phi}(t)$ függvény az alapmegoldás, és segítségével a megoldás a bemenő függvényekből közvetlenül felírható:

$$\mathbf{u}(t) = \mathbf{\Phi}(t)\mathbf{v} + \int_a^t \mathbf{G}(t,s)\mathbf{r}(s)ds, \tag{10.3.23}$$

ahol[7]

$$\mathbf{G}(t,s) = \begin{cases} \mathbf{\Phi}(t)\mathbf{B}_a\mathbf{\Phi}(a)\mathbf{\Phi}^{-1}(s), & \text{ha } s \le t; \\ -\mathbf{\Phi}(t)\mathbf{B}_b\mathbf{\Phi}(b)\mathbf{\Phi}^{-1}(s), & \text{ha s >t.} \end{cases} \tag{10.3.24}$$

◇

10.4. A peremérték-feladat numerikus megoldása Cauchy-feladatra való visszavezetéssel

Bár az előző pontban megmutattuk, hogy egy peremérték-feladat megoldása képletek segítségével előállítható (vö.(10.3.23)), ez a felírás formális: konkrét feladatok esetén sem az alapmegoldás, sem a $\mathbf{G}(t,s)$ függvény általában nem határozható meg. Ezért a Cauchy-feladatokhoz hasonlóan a peremérték-feladatok esetén is valamely numerikus megoldás alkalmazása szükséges.

Az ismertetésre kerülő numerikus módszereket két csoportba sorolhatjuk.

- A peremérték-feladat megoldását visszavezetjük Cauchy-féle kezdetiérték-feladatokra, és ezek megoldására a korábban ismertetett numerikus módszerek valamelyikét alkalmazzuk.
- Közvetlenül diszkretizáljuk a peremérték-feladatot.

Ebben a pontban az első módszerrel foglalkozunk.

[7] Ezt a $\mathbf{G}(t,s)$ függvényt a feladat Green-függvényének nevezzük.

10.4.1. A belövéses módszer

Tekintsük az

$$\begin{aligned} \mathbf{u}' &= \mathbf{f}(t,\mathbf{u}), \quad t\in[a,b] \\ u_1(a) &= \alpha, \quad u_1(b)=\beta \end{aligned} \tag{10.4.1}$$

előző szakaszban már felírt általános alakú peremérték-feladatot! A *belövéses módszer*[8] lényege, hogy a (10.4.1) feladat helyett az

$$\begin{aligned} \mathbf{u}' &= \mathbf{f}(t,\mathbf{u}), \quad t\in[a,b] \\ \mathbf{u}(a) &= \mathbf{c}, \end{aligned} \tag{10.4.2}$$

kezdetiérték-feladatot oldjuk meg. Mivel

$$\mathbf{u}(a) = \begin{bmatrix} u_1(a) \\ u_2(a) \end{bmatrix}, \tag{10.4.3}$$

ezért a $t=a$ pontbeli kezdeti állapotot leíró $\mathbf{c}\in\mathbb{R}^2$ vektor első komponense a (10.4.1) peremfeltételéből ismert, viszont a második komponensét ($u_2(a)=u'(a)$) nem ismerjük. Legyen $c=u_2(a)$. A belövéses módszer lényege, hogy a

$$\mathbf{c} = \begin{bmatrix} \alpha \\ c \end{bmatrix} \tag{10.4.4}$$

kezdetiérték megválasztásában a c állandót úgy határozzuk meg, hogy a (10.4.2) feladat megoldása a $t=b$ pontban az előírtaknak megfelelően a β értéket vegye fel. Ez a következőt jelenti. Jelölje $\mathbf{u}(t,c)$ a (10.4.1) feladatnak a (c paraméter megválasztásától függő) megoldását. Feladatunk a c paraméter olyan megválasztása, amely mellett

$$u_1(b,c)=\beta. \tag{10.4.5}$$

Ezt a

$$h(c)=u_1(b,c)-\beta \tag{10.4.6}$$

függvény bevezetésével a

$$h(c)=0 \tag{10.4.7}$$

egyenlet megoldásával kaphatjuk meg. A (10.4.6) egyenlet egy nemlineáris egyenlet, amelyben $h:\mathbb{R}\to\mathbb{R}$ ("valós-valós") típusú függvény. Megoldására több ismert módszer is alkalmazható.

Az egyik legegyszerűbb módszer az *intervallum-felező módszer*, azaz keresünk olyan c_1 és c_2 értékeket, amelyekre $h(c_1)h(c_2)<0$, és ekkor a (c_1,c_2) intervallumon lévő gyököt az intervallum folyamatos felezésével határozhatjuk meg.

A belövéses módszer intervallum-felező módszeres algoritmusa tehát a következő.

1. Rögzítünk valamely c értéket.
2. A (10.4.4) kezdeti vektorral megoldjuk az $[a,b]$ intervallumon a (10.4.1) kezdetiérték-feladatot, alkalmazva az előző fejezetben ismertetett valamely numerikus módszert. (Például egy Runge-Kutta típusú módszert.)
3. Az 1. lépésben két olyan c értéket keresünk, amelyekre a 2. lépésben kiszámolt $t=b$ pontbeli h értékek ellentétes előjelűek lesznek. Legyenek ezek c_1 és c_2.

[8] A módszer angol elnevezése: *shooting method*.

4. A $c = 0.5(c_1 + c_2)$ értékkel újra számoljuk a 2. lépést.

5. Az így nyert megoldás $t = b$ pontbeli értékének előjele alapján újra megválasztjuk c_1 vagy c_2 értékét, és folytatjuk az eljárást.

6. Amikor a két végpont távolsága kisebb, mint egy előre megadott $\varepsilon > 0$ érték, akkor befejezzük az iterációt.

A belövéses módszer során a (10.4.6) egyenlet megoldása szükséges, amihez az előzőekben a felező eljárást alkalmaztuk. Ennél hatékonyabb (bár még mindig csak elsőrendű) a *húrmódszer*. Ennek algoritmusa könnyen felírható az előzőek alapján. Ugyanakkor a felezőmódszer és a húrmódszer közös hátránya a kezdeti ellentétes előjelű alappontok meghatározása. Ha a *szelőmódszert* alkalmazzuk, akkor erre nincs szükség. Ezzel a módszerrel algoritmusunk a következő:

1. Legyenek $c^{(0)}$ és $c^{(1)}$ tetszőleges értékek.

2. A (10.4.4) kezdeti vektor képletébe először $c^{(0)}$-t, majd $c^{(1)}$-t helyettesítve, megoldjuk az $[a, b]$ intervallumon a (10.4.1) kezdetiérték-feladatot, alkalmazva az előző fejezetben ismertetett valamely numerikus módszert.

3. Ekkor $h(c^{(0)}) = u(b, c^{(0)})$ és $h(c^{(1)}) = u(b, c^{(1)})$.

4. Lineáris közelítést alkalmazva a $(c^{(0)}, h(c^{(0)}))$ és a $(c^{(1)}, h(c^{(1)}))$ pontok között, meghatározzuk azt a $c^{(2)}$ pontot, ahol ez a közelítés nulla értéket vesz fel:

$$c^{(2)} = c^{(1)} - \left(\frac{c^{(1)} - c^{(0)}}{h(c^{(1)}) - h(c^{(0)})} \right) h(c^{(1)}). \tag{10.4.8}$$

5. A (10.4.8) képlettel kiszámolt $c^{(2)}$ értékkel meghatározzuk a (10.4.4) alakú kezdeti vektort, és a Cauchy-feladat numerikus megoldásával meghatározzuk a $h(c^{(2)})$ értéket.

6. A továbbikban egy kétlépéses iterációt építünk fel a fenti lépések analógiájaként: $k = 2, 3, \ldots$ értékekre a már kiszámolt $(c^{(k-2)}, h(c^{(k-2)}))$ és a $(c^{(k-1)}, h(c^{(k-1)}))$ pontok ismeretében meghatározzuk a

$$c^{(k)} = c^{(k-1)} - \left(\frac{c^{(k-1)} - c^{(k-2)}}{h(c^{(k-1)}) - h(c^{(k-2)})} \right) h(c^{(k-1)}) \tag{10.4.9}$$

közelítést, majd ennek ismeretében a $h(c^{(k)})$ értéket.

7. Amikor $|h(c^{(k)})|$ értéke kisebb, mint egy előre megadott $\varepsilon > 0$ érték, akkor befejezzük az iterációt.

10.4.1. megjegyzés. A felezőmódszerrel és a húrmódszerrel összehasonlítva a szelőmódszer előnye tehát, hogy az iteráció elindításához nem szükséges két olyan kezdeti közelítés megkeresése, amely mellett a h ellentétes előjelet vesz fel ezekben a pontokban, hanem indulhatunk két tetszőleges $c^{(0)}$ és $c^{(1)}$ közelítésből. Ugyanakkor tetszőleges feladatok esetén a módszer konvergenciája a'priori nem biztosítható. ⋄

A fenti algoritmus utolsó lépése után szokásos még egy javító lépést tenni. (Ennek oka, hogy ha $h(c^{(k-1)})$ és $h(c^{(k)})$ nagyon közel vannak a nullához, akkor az algoritmusunk instabil. Ezért

ε értékét nem célszerű túlságosan kicsinek választani.) Ezt a következő módon hajtjuk végre. Elkészítjük a

$$\begin{array}{|c|c|c|c|} \hline h(c^{(0)}) & h(c^{(1)}) & \dots & h(c^{(n)}) \\ \hline c^{(0)} & c^{(1)} & \dots & c^{(n)} \\ \hline \end{array} \tag{10.4.10}$$

táblázatot és egy n-ed fokú interpolációs polinomot fektetünk ezekre az adatokra, azaz meghatározzuk azt a $p(t)$ n-ed fokú polinomot, amelyre $p(h(c^{(k)})) = c^{(k)}$ minden $k = 0, 1, \dots n$ esetén. Tehát a p polinom a h^{-1} függvényt (a h függvény inverzét) interpolálja. Ekkor a h függvény gyökéhez való $c^{(n+1)}$ közelítést a $p(0) = c^{(n+1)}$ összefüggés alapján határozhatjuk meg.

Az egyenlet megoldására lényegesen hatékonyabb módszer a másodrendű *Newton-módszer*, amelynek során a

$$c^{(s+1)} = c^{(s)} - \frac{h(c^{(s)})}{h'(c^{(s)})}, \quad s = 0, 1, \dots \tag{10.4.11}$$

iterációval felépített sorozattal közelítjük a keresett gyököt. (A (10.4.11) iterációban a $c^{(0)}$ megfelelő megválasztása valamely elsőrendű módszerrel nyerhető.) A (10.4.11) képlet alkalmazásánál alapvető probléma $h'(c^{(s)})$ értékének meghatározása.

Ehhez térjünk vissza az eredeti (10.1.7) feladatra! A belövéses módszer alkalmazása azt jelenti, hogy keressük valamely $c \in \mathbb{R}$ mellett az

$$\begin{aligned} u'' &= f(t, u, u'), \quad t \in (a, b), \\ u(a) &= \alpha, \quad u'(a) = c \end{aligned} \tag{10.4.12}$$

feladat megoldását. Legyen ez az $u(t, c)$ függvény és ekkor

$$\begin{aligned} u''(t, c) &= f(t, u(t, c), u'(t, c)), \quad t \in (a, b), \\ u(a, c) &= \alpha, \quad u'(a, c) = c. \end{aligned} \tag{10.4.13}$$

Deriváljuk a (10.4.13) azonosságot a c paraméter szerint! Ekkor

$$\begin{gathered} \frac{\partial u''}{\partial c}(t, c) = \partial_2 f(t, u(t, c), u'(t, c)) \frac{\partial u}{\partial c}(t, c) + \partial_3 f(t, u(t, c), u'(t, c)) \frac{\partial u'}{\partial c}(t, c), \\ \frac{\partial u(a, c)}{\partial c} = 0, \quad \frac{\partial u'(a, c)}{\partial c} = 1. \end{gathered} \tag{10.4.14}$$

Vezessük be a

$$w(t) := \frac{\partial u}{\partial c}(t, c)$$

új függvényt! A (10.4.14) összefüggések alapján erre a függvényre:

$$\begin{gathered} w''(t) = \partial_2 f(t, u(t, c), u'(t, c)) w(t) + \partial_3 f(t, u(t, c), u'(t, c)) w'(t), \\ w(a) = 0, \quad w'(a) = 1. \end{gathered} \tag{10.4.15}$$

A (10.4.15) reláció látszólag az ismeretlen $w(t)$ függvényre egy Cauchy-feladat[9]. Ugyanakkor valójában nem az, hiszen a $\partial_2 f$ és $\partial_3 f$ függvényeket az ismeretlen $(t, u(t, c), u'(t, c))$ pontban nem tudjuk kiértékelni. Ha hozzávesszük a (10.4.15) relációhoz a (10.4.13) egyenlőséget, akkor egy négyismeretlenes elsőrendű közönséges differenciálegyenlet-rendszer Cauchy-feladatát kapjuk! Nevezetesen, bevezetve a

$$v_1(t) = u(t, c), \quad v_2(t) = u'(t, c), \quad v_3(t) = w(t), \quad v_4(t) = w'(t) \tag{10.4.16}$$

[9] Ezt a feladatot az irodalomban *első variációs egyenletnek* szokásos nevezni.

függvényeket, ezekre a függvényekre a fenti összefüggések alapján a következő kapcsolat áll fenn:

$$\begin{aligned} v_1'(t) &= v_2(t), \quad v_2'(t) = f(t, v_1(t), v_2(t)), \\ v_3'(t) &= v_4(t), \quad v_4'(t) = \partial_2 f(t, v_1(t), v_2(t)) v_3(t) + \partial_3 f(t, v_1(t), v_2(t)) v_4(t). \end{aligned} \tag{10.4.17}$$

Emellett

$$v_1(a) = \alpha, \quad v_2(a) = c, \quad v_3(a) = 0, \quad v_4(a) = 1. \tag{10.4.18}$$

Ezért az $[a, b]$ intervallumon megoldva a

$$\begin{aligned} v_1' &= v_2, \quad v_2' = f(t, v_1, v_2) \\ v_3' &= v_4, \quad v_4' = \partial_2 f(t, v_1, v_2) v_3 + \partial_3 f(t, v_1, v_2) v_4, \\ v_1(a) &= \alpha, \quad v_2(a) = c, \quad v_3(a) = 0, \quad v_4(a) = 1 \end{aligned} \tag{10.4.19}$$

Cauchy-feladatot, meghatározhatjuk az ismeretlen függvényeinket. Mivel

$$v_3(b) = w(b) = \frac{\partial u}{\partial c}(b, c), \tag{10.4.20}$$

ezért (10.4.6) miatt $h'(c) = v_3(b)$, azaz a (10.4.20) szerinti Newton-iteráció jobb oldalának nevezője kiszámítható.

A belövéses módszer Newton-módszeres algoritmusa tehát a következő.

1. Rögzítünk valamely $c^{(0)}$ értéket.
2. A $c = c^{(0)}$ megválasztással az $[a, b]$ intervallumon megoldjuk a (10.4.19) rendszerre kitűzött kezdetiérték-feladatot az előző fejezetben ismertetett valamely numerikus módszerrel.
3. A megoldások ismertében kiszámoljuk a $h(c^{(0)}) = v_1(b)$ és $h'(c^{(0)}) = v_3(b)$ értékeket.
4. A (10.4.11) képlet alapján meghatározzuk a $c^{(1)}$ közelítést.
5. Folytatjuk az eljárást a 2. lépéstől kezdve.
6. Amikor $|h(c^{(k)})|$ értéke kisebb, mint egy előre megadott $\varepsilon > 0$ érték, akkor befejezzük az iterációt.

10.4.2. Lineáris peremérték-feladatok numerikus megoldása

Ebben a szakaszban a (10.3.3) módon bevezetett

$$\begin{aligned} u'' &= p(t)u' + q(t)u + r(t), \quad t \in [a, b], \\ u(a) &= \alpha, \quad u(b) = \beta \end{aligned} \tag{10.4.21}$$

lineáris peremérték-feladat numerikus megoldásával foglalkozunk. Megmutatjuk, hogy ha az általános alakú (10.1.7) feladatban az f függvényt a (10.4.21) szerinti speciális módon adjuk meg, akkor a numerikus tárgyalás is lényegesen egyszerűbbé válik.

Tegyük fel, hogy a (10.4.21) feladatot a belövéses módszer segítségével megoldjuk a $c = c_1$ és a $c = c_2$ megválasztással. Jelölje $u_1(t) = u(t, c_1)$ és $u_2(t) = u(t, c_2)$ a c_1 és c_2 értékekhez tartozó két Cauchy-feladat megoldását. Ekkor tehát

$$\begin{aligned} u_1''(t) &= p(t)u_1'(t) + q(t)u_1(t) + r(t), \quad t \in [a, b], \\ u_1(a) &= \alpha, \quad u_1'(a) = c_1, \end{aligned} \tag{10.4.22}$$

$$\begin{aligned} u_2''(t) =& p(t)u_2'(t) + q(t)u_2(t) + r(t), \quad t \in [a,b], \\ & u_2(a) = \alpha, \quad u_2'(a) = c_2. \end{aligned} \tag{10.4.23}$$

Legyen továbbá

$$w(t) = \lambda u_1(t) + (1-\lambda)u_2(t) \tag{10.4.24}$$

egy új függvény, ahol $\lambda \in \mathbb{R}$ valamely, egyelőre tetszőleges paraméter. Mivel a (10.4.21) feladat lineáris, ezért az egyenlet két megoldásának (10.4.24) szerinti lineáris kombinációja is megoldás, azaz

$$w''(t) = p(t)w'(t) + q(t)w(t) + r(t), \quad t \in [a,b]. \tag{10.4.25}$$

Másrészt $w(a) = \alpha$ és $w(b) = \lambda u_1(b) + (1-\lambda)u_2(b)$. Ezért válasszuk meg a λ paramétert úgy, hogy teljesüljön a $w(b) = \beta$ feltétel, azaz $\lambda u_1(b) + (1-\lambda)u_2(b) = \beta$. Innen λ értékét a

$$\lambda = \frac{\beta - u_2(b)}{u_1(b) - u_2(b)} \tag{10.4.26}$$

egyenlőségből határozhatjuk meg. Ezért a (10.4.26) értékű λ melletti (10.4.24) alakú $w(t)$ függvény lesz az eredeti (10.4.21) lineáris peremérték-feladat megoldása.

A fenti megoldás előállításának számítógépes realizálása a következő módon valósítható meg. Tekintsünk a következő két Cauchy-feladatot:

$$\begin{aligned} u'' =& p(t)u' + q(t)u + r(t), \quad t \in [a,b], \\ & u(a) = \alpha, \quad u'(a) = 0, \end{aligned} \tag{10.4.27}$$

valamint

$$\begin{aligned} u'' =& p(t)u' + q(t)u + r(t), \quad t \in [a,b], \\ & u(a) = \alpha, \quad u'(a) = 1. \end{aligned} \tag{10.4.28}$$

Jelölje (10.4.27) megoldását $u_1(t)$, a (10.4.28) feladat megoldását pedig $u_2(t)$. (Ezeket a (10.4.27) és a (10.4.28) feladatok elsőrendű rendszerre való visszavezetésével megkaphatjuk az $[a,b]$ intervallumon, az előző fejezetben ismertetett numerikus módszerek valamelyikével.) Ezután a (10.4.26) képlet alapján meghatározzuk a λ paraméter értékét, és végezetül a (10.4.21) feladat megoldását $u_1(t)$ és $u_2(t)$ előzőekben már kiszámolt közelítéseiből a (10.4.24) összefüggés alapján határozzuk meg.

10.4.2. megjegyzés. Felmerülhet a kérdés: vajon mindig kifejezhető-e λ a (10.4.26) képlettel? Érvényes a következő állítás [20].

10.4.3. tétel.

Tegyük fel, hogy a (10.4.21) lineáris peremérték-feladatnak létezik egyértelmű megoldása. Ekkor

- vagy a (10.4.22) tulajdonságú $u_1(t)$ függvény ez a megoldás,
- vagy $u_1(b) - u_2(b) \neq 0$, és így a (10.4.26) képletből λ meghatározható.

Tehát korrekt kitűzésű feladatok esetén a numerikus megoldás is meghatározható. ◇

A (10.4.21) lineáris peremérték-feladat egyértelműen létező megoldását elő tudjuk állítani nemcsak a (10.4.27) és a (10.4.28) Cauchy-feladatok lineáris kombinációjaként, hanem más Cauchy-feladatok megoldásainak kombinációjaként is. Tekintsünk a következő két Cauchy-feladatot:

$$\begin{aligned} u'' =& p(t)u' + q(t)u + r(t), \quad t \in [a,b], \\ & u(a) = \alpha, \quad u'(a) = 0, \end{aligned} \tag{10.4.29}$$

valamint

$$\begin{aligned} u'' =& p(t)u' + q(t)u + r(t), \quad t \in [a,b], \\ & u(a) = 0, \quad u'(a) = 1. \end{aligned} \tag{10.4.30}$$

Jelölje ismét (10.4.29) megoldását $u_1(t)$, a (10.4.30) feladat megoldását pedig $u_2(t)$. Könnyen láthatóan ekkor a

$$w(t) = u_1(t) + \frac{\beta - u_1(b)}{u_2(b)} u_2(t)$$

függvény megoldása a (10.4.21) feladatnak. Így elegendő numerikusan megoldani a (10.4.29) és a (10.4.30) Cauchy-feladatokat.

Összefoglalóan tehát megállapíthatjuk, hogy a lineáris peremérték-feladatok numerikus megoldása során ténylegesen nem alkalmazzuk a belövéses módszert, mivel a megoldást elő tudjuk állítani két, egyszerű struktúrájú, másodrendű közönséges differenciálegyenlet Cauchy-feladatának megoldásával. Ez, összehasonlítva az általános esettel, a számítási költségek tekintetében jelentős megtakarítást jelent: míg a lineáris esetben összesen kétszer kell valamely egylépéses kezdetiérték-feladatot megoldó numerikus módszert alkalmazni, addig az általános esetben annyiszor, ahány iterációs lépést hajtunk végre a (10.4.7) egyenlet közelítő megoldására.

10.5. A peremérték-feladat numerikus megoldása véges differenciák módszerével

Az előző szakaszban láttuk, hogy a belövéses módszer alkalmazása meglehetősen munkaigényes: minden egyes függvénykiértékelés egy Cauchy-feladat numerikus megoldását igényli. Ebben a szakaszban más megközelítést alkalmazunk: a belövéses módszertől eltérően, a véges differenciás módszer alkalmazása során nem egy kezdetiérték-feladatra való visszavezetéssel oldjuk meg az

$$\begin{aligned} u'' =& f(t,u,u'), \quad t \in (a,b), \\ & u(a) = \alpha, \quad u(b) = \beta \end{aligned} \tag{10.5.1}$$

feladatot, hanem az időintervallumon rácshálót határozunk meg, és a feladatban szereplő deriváltakat ezen rácshálón approximáljuk. Ezután ennek segítségével állítjuk elő a közelítő megoldást.

10.5.1. Véges differenciás approximáció

A továbbiakban felhasználjuk a megfelelően sima függvényekre jól ismert, és a numerikus deriválás során már ismertetett alábbi összefüggéseket:

$$u'(t) = \frac{u(t+h) - u(t)}{h} - \frac{1}{2} h u''(\zeta_1) \tag{10.5.2}$$

$$u'(t) = \frac{u(t) - u(t-h)}{h} + \frac{1}{2} h u''(\zeta_2) \tag{10.5.3}$$

$$u'(t) = \frac{u(t+h) - u(t-h)}{2h} - \frac{1}{6} h^2 u'''(\zeta_3) \tag{10.5.4}$$

$$u''(t) = \frac{u(t+h) - 2u(t) + u(t-h)}{h^2} - \frac{1}{12} h^2 u^{(4)}(\zeta_4), \tag{10.5.5}$$

ahol $\zeta_i \in (t, t+h)$ $(i = 1,2,3,4)$ valamely rögzített szám.

Ez adja azt az ötletet, hogy a (megfelelően sima) $u(t)$ függvény egy rögzített $t^\star$ pontbeli deriváltját a $t^\star$ pontot környezetében lévő rácspontbeli helyettesítési értékekkel a következő módon közelítsük:

$$u'(t^\star) \simeq \frac{u(t^\star+h)-u(t^\star)}{h},\ u'(t^\star) \simeq \frac{u(t^\star)-u(t^\star-h)}{h},\ u'(t^\star) \simeq \frac{u(t^\star+h)-u(t^\star-h)}{2h},$$
$$u''(t^\star) \simeq \frac{u(t^\star+h)-u(t^\star)+u(t^\star-h)}{h^2}. \tag{10.5.6}$$

10.5.2. Az általános alakú peremérték-feladat megoldása a véges differenciák módszerével

Jelöljünk ki az $[a,b]$ intervallumon egy rácsháló t! Az egyszerűség kedvéért legyen ez az

$$\overline{\omega}_h = \{t_i = a+ih,\quad i=0,1\ldots,N+1,\ h=(b-a)/(N+1)\} \tag{10.5.7}$$

ekvidisztáns rácsháló, ahol h a *rácsháló lépésköze*. Jelölje ω_h a *rácsháló belső pontjait*, azaz

$$\omega_h = \{t_i = a+ih,\quad i=1,2,\ldots,N,\ h=(b-a)/(N+1)\}, \tag{10.5.8}$$

továbbá γ_h a rácsháló határpontjait, azaz

$$\gamma_h = \{t_0 = a,\ t_{N+1}=b\}. \tag{10.5.9}$$

Tegyük fel, hogy a (10.5.1) feladatnak létezik egyéretelmű $u(t)$ megoldása, azaz $u(t)\in C^2[a,b]$ olyan függvény[10], amelyre

$$\begin{aligned} u''(t) =& f(t,u(t),u'(t)),\quad t\in(a,b),\\ u(a)=\alpha,&\quad u(b)=\beta. \end{aligned} \tag{10.5.10}$$

Ezért az $\overline{\omega}_h$ rácsháló pontjaiban felírva a fenti összefüggést, az

$$\begin{aligned} u''(t_i) =& f(t_i,u(t_i),u'(t_i)),\quad i=1,2,\ldots,N\\ u(t_0)=\alpha,&\quad u(t_{N+1})=\beta \end{aligned} \tag{10.5.11}$$

összefüggések állnak fenn. Alkalmazva a (10.5.11) azonosságban a deriváltakra nyert (10.5.6) szerinti közelítéseket, az ismeretlen y_i értékekre (amelyek szándékaink szerint $u(t_i)$ közelítései), az alábbi egyenleteket nyerjük:

$$\begin{aligned} \frac{y_{i+1}-2y_i+y_{i-1}}{h^2} =& f(t_i,y_i,\frac{y_{i+1}-y_i}{h}),\quad i=1,2,\ldots,N\\ y_0=\alpha,&\quad y_{N+1}=\beta. \end{aligned} \tag{10.5.12}$$

Megoldva a fenti $N+2$ ismeretlenes egyenletrendszert az $y_0,y_1,\ldots,y_{N+1}$ ismeretlenekre, meghatározhatjuk a véges differenciás közelítéseket. Ugyanakkor, valójában nem tudjuk a választ a következő kérdésekre.

- A (10.5.12) rendszernek létezik-e egyértelmű megoldása?
- Ha igen, hogyan oldható meg a rendszer hatékonyan?

[10] Mint az a (10.5.10) összefüggésből látszik, elegendő az $u(t)\in C^2(a,b)\cap C[a,b]$ simasági feltétel, de a gyakorlati esetekben ez az egyszerűbben ellenőrizhető feltétel nem jelent tényleges megszorítást.

- Közel van-e y_i értéke $u(t_i)$ értékéhez?
- Ha sűrítjük a rácshálót (azaz h nullához tart), akkor az $[a,b]$ intervallumon a közelítő megoldások sorozata tart-e a pontos megoldáshoz?

10.5.1. megjegyzés. A (10.5.12) diszkretizációban az első deriváltat a (10.5.6) első formulájával helyettesítettük. Ha a második formulával helyettesítjük, akkor az

$$\begin{aligned} \frac{y_{i+1}-2y_i+y_{i-1}}{h^2} &= f(t_i, y_i, \frac{y_i-y_{i-1}}{h}), \quad i=1,2,\dots,N \\ y_0=\alpha, &\quad y_{N+1}=\beta, \end{aligned} \tag{10.5.13}$$

ha pedig a harmadikkal, akkor az

$$\begin{aligned} \frac{y_{i+1}-2y_i+y_{i-1}}{h^2} &= f(t_i, y_i, \frac{y_{i+1}-y_{i-1}}{2h}), \quad i=1,2,\dots,N \\ y_0=\alpha, &\quad y_{N+1}=\beta \end{aligned} \tag{10.5.14}$$

feladatot kapjuk. ◇

Vezessük be a következő jelöléseket:

$$y_{x,i}=\frac{y_{i+1}-y_i}{h}; \quad y_{\bar{x},i}=\frac{y_i-y_{i-1}}{h}; \quad y_{x^\circ,i}=\frac{y_{i+1}-y_{i-1}}{2h}, \tag{10.5.15}$$

amelyeket rendre jobb, bal és középponti differenciáknak nevezünk. A (10.5.2) és a (10.5.3) alapján nyilvánvalóan a jobb és bal oldali differenciák elsőrendben, míg a (10.5.4) következtében a középponti differencia másodrendben approximálja az első deriváltat a $t=t_i$ pontban.
Könnyen látható, hogy a t_i pontbeli második derivált véges differenciás közelítése

$$y_{\bar{x}x,i}:=\frac{y_{x,i}-y_{\bar{x},i}}{h}=\frac{y_{i+1}-2y_i+y_{i-1}}{h^2} \tag{10.5.16}$$

alakú, és ezért a (10.5.5) összefüggés miatt $y_{\bar{x}x,i}$ másodrendben approximálja az $u''(t)$ függvényt a $t=t_i$ pontban. Ezen jelölésekkel az egyes sémák a következő alakot öltik:

- A (10.5.12) séma:

$$\begin{aligned} y_{\bar{x}x,i} &= f(t_i, y_i, y_{x,i}), \quad i=1,2,\dots,N \\ y_0=\alpha, &\quad y_{N+1}=\beta. \end{aligned} \tag{10.5.17}$$

- A (10.5.13) séma:

$$\begin{aligned} y_{\bar{x}x,i} &= f(t_i, y_i, y_{\bar{x},i}), \quad i=1,2,\dots,N \\ y_0=\alpha, &\quad y_{N+1}=\beta. \end{aligned} \tag{10.5.18}$$

- A (10.5.14) séma:

$$\begin{aligned} y_{\bar{x}x,i} &= f(t_i, y_i, y_{x^\circ,i}), \quad i=1,2,\dots,N \\ y_0=\alpha, &\quad y_{N+1}=\beta. \end{aligned} \tag{10.5.19}$$

10.5.3. A lineáris peremérték-feladatok approximációja a véges differenciák módszerével

Ebben a szakaszban választ adunk az előzőekben megfogalmazott kérdésekre a

$$\begin{aligned} u'' =& p(t)u' + q(t)u + r(t), \quad t \in [a,b], \\ & u(a) = \alpha, \quad u(b) = \beta \end{aligned} \tag{10.5.20}$$

lineáris peremérték-feladat véges differenciás numerikus megoldására. Mint azt megmutattuk, az egyértelmű megoldhatósághoz feltételezzük, hogy $p, q, r \in C[a,b]$ és $q(t) > 0$ az $[a,b]$ intervallumon, azaz $\min_{[a,b]} q := q_{\min} > 0$. Vezessük be a $p_i = p(t_i), q_i = q(t_i), r_i = r(t_i)$ jelöléseket!

10.5.2. tétel.
A (10.5.20) lineáris peremérték-feladat (10.5.17)-(10.5.19) véges differenciás diszkretizációja

$$\begin{aligned} a_i y_{i-1} + d_i y_i + c_i y_{i+1} &= -r_i, \quad i = 1, 2, \ldots, N \\ y_0 = \alpha, \quad & y_{N+1} = \beta \end{aligned} \tag{10.5.21}$$

alakú lineáris algebrai egyenletrendszer, ahol mindhárom diszkretizációra, megfelelően kis h esetén, érvényes a

$$|d_i| - |a_i| - |c_i| = q_i \tag{10.5.22}$$

egyenlőség.

Bizonyítás. A (10.5.20) feladatra a (10.5.17) diszkretizáció az

$$\begin{aligned} \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} &= p_i \frac{y_{i+1} - y_i}{h} + q_i y_i + r_i, \quad i = 1, 2, \ldots, N, \\ y_0 = \alpha, \quad & y_{N+1} = \beta \end{aligned} \tag{10.5.23}$$

lineáris algebrai egyenletrendszert jelenti, ami az

$$a_i = -\frac{1}{h^2}, \; d_i = \frac{2}{h^2} + q_i - \frac{1}{h} p_i, \; c_i = -\frac{1}{h^2} + \frac{1}{h} p_i \tag{10.5.24}$$

megválasztással a (10.5.21) feladatot jelenti. Megfelelően kis h esetén ezért

$$|d_i| - |a_i| - |c_i| = \frac{1}{h^2} \left((2 + h^2 q_i - h p_i) - 1 - (1 - h p_i) \right) = q_i. \tag{10.5.25}$$

Könnyen látható, hogy a (10.5.18) diszkretizáció esetén

$$a_i = -\frac{1}{h^2} - \frac{1}{h} p_i, \; d_i = \frac{2}{h^2} + q_i + \frac{1}{h} p_i, \; c_i = -\frac{1}{h^2}, \tag{10.5.26}$$

míg a (10.5.19) diszkretizáció esetén

$$a_i = -\frac{1}{h^2} - \frac{1}{2h} p_i, \; d_i = \frac{2}{h^2} + q_i, \; c_i = -\frac{1}{h^2} + \frac{1}{2h} p_i, \tag{10.5.27}$$

és mindkét esetben alkalmasan megválasztott kis h értékek mellett (10.5.22) érvényes. ■

10.5.3. következmény. Mivel a (10.5.21) rendszer felírható

$$\begin{aligned} d_1 y_1 + c_1 y_2 &= -r_1 - a_1 \alpha, \\ a_i y_{i-1} + d_i y_i + c_i y_{i+1} &= -r_i, \quad i = 2, 3, \ldots, N-1, \\ a_N y_{N-1} + d_N y_N &= -r_N - c_N \beta \end{aligned} \tag{10.5.28}$$

alakban, ezért a diszkretizált feladat valójában egy N ismeretlenes lineáris algebrai egyenletrendszert jelent, amelynek együtthatómátrixa tridiagonális, szigorúan diagonálisan domináns mátrix. ◇

Jelölje e_i a $t = t_i$ pontban a pontos és a közelítő megoldás eltérését (azaz $e_i = y_i - u_i$), és $\mathbf{e}_h$ az ω_h rácsháló pontjaiban értelmezett hibafüggvényt (azaz $\mathbf{e}_h(t_i) = e_i$)!

10.5.4. definíció.

Valamely numerikus módszer *konvergenciája* azt jelenti, hogy az általa generált numerikus megoldásra a rácsháló finomításával a hibafüggvény nullához tart, azaz

$$\lim_{h \to 0} \mathbf{e}_h = 0. \tag{10.5.29}$$

Ha $\mathbf{e}_h = \mathcal{O}(h^p)$, akkor a numerikus módszert p-ed rendű módszernek nevezzük.

10.5.5. megjegyzés. A 10.5.29 képletben az $\mathbf{e}_h$ vektorsorozat konvergenciáját természetesen normában értjük, azaz az $\mathbf{e}_h \in \mathbb{R}^{N+2}$ vektorsorozat (ahol $h = (b-a)/(N+1)$) az $\|\cdot\|_{\mathbb{R}^{N+2}}$ normában tart nullához. ◇

A következő állításban a fenti véges differenciás közelítések konvergenciájával és a konvergencia rendjével foglalkozunk.

10.5.6. tétel.

A (10.5.24), (10.5.26) és (10.5.27) megválasztású (10.5.21) alakú véges differenciás diszkretizációk lineáris peremérték-feladat esetén konvergensek, emellett

1. a (10.5.27) megválasztás esetén a séma másodrendben,
2. a (10.5.24) és a (10.5.26) megválasztás esetén pedig a sémák elsőrendben

konvergálnak a (10.5.20) feladat megoldásához.

Bizonyítás. Mivel a két állítás bizonyítása lényegében megegyezik, a továbbiakban csak az elsőt mutatjuk meg.
Felhasználva a (10.5.4) és a (10.5.5) approximációs tulajdonságokat, a (10.5.20) feladat a $t = t_i$ pontban így írható fel:

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - \frac{1}{12}h^2 u^{(4)}(\zeta_4^{(i)}) = p_i \left(\frac{u_{i+1} - u_{i-1}}{2h} - \frac{1}{6}h^2 u'''(\zeta_3^{(i)}) \right) + q_i u_i + r_i. \tag{10.5.30}$$

Átrendezve a (10.5.30) egyenlőséget a következő alakot kapjuk:

$$\left(-\frac{1}{h^2} + \frac{p_i}{2h} \right) u_{i+1} + \left(\frac{2}{h^2} + q_i \right) u_i + \left(-\frac{1}{h^2} - \frac{p_i}{2h} \right) u_{i-1} = -r_i - h^2 g_i, \tag{10.5.31}$$

ahol

$$g_i = \frac{1}{12}u^{(4)}(\zeta_4^{(i)}) - \frac{p_i}{6}u'''(\zeta_3^{(i)}).$$

Ezért a (10.5.27) jelöléseivel (10.5.31) felírható az

$$a_i u_{i-1} + d_i u_i + c_i u_{i+1} = -r_i - h^2 g_i \tag{10.5.32}$$

alakban. Mivel a numerikus megoldásra érvényes a (10.5.21) összefüggés, (ahol az együtthatók (10.5.27) szerintiek), a (10.5.32) egyenlőségből kivonva a (10.5.21) egyenleteket, az $\mathbf{e}_h$ hibavektor koordinátái kielégítik az

$$\begin{gathered} a_i e_{i-1} + d_i e_i + c_i e_{i+1} = -h^2 g_i, \quad i = 1, 2, \ldots, N, \\ e_0 = e_{N+1} = 0 \end{gathered} \tag{10.5.33}$$

alakú lineáris algebrai egyenletrendszert. Az i-edik egyenletet átrendezve:

$$d_i e_i = -a_i e_{i-1} - c_i e_{i+1} - h^2 g_i, \tag{10.5.34}$$

azaz abszolút értékben becsülve a

$$|d_i||e_i| \le |a_i||e_{i-1}| + |c_i||e_{i+1}| + h^2|g_i| \le (|a_i| + |c_i|)\,\|\mathbf{e}_h\|_\infty + h^2\|\mathbf{g}\|_\infty \tag{10.5.35}$$

egyenlőtlenséget nyerjük, ahol $\mathbf{g}$ a g_i koordinátájú vektor. Jelölje i_0 azt az indexet, amelyre $\|\mathbf{e}_h\|_\infty = |e_{i_0}|$. (Nyilvánvalóan $i_0 \neq 0$ és $i_0 \neq N+1$, mivel $e_0 = e_{N+1} = 0$.) Mivel (10.5.35) minden $i = 1, 2, \ldots, N$ értékre érvényes, ezért az $i = i_0$ indexre is, azaz

$$|d_{i_0}|\|\mathbf{e}_h\|_\infty \le (|a_{i_0}| + |c_{i_0}|)\,\|\mathbf{e}_h\|_\infty + h^2\|\mathbf{g}\|_\infty. \tag{10.5.36}$$

Átrendezve a (10.5.36) egyenlőtlenséget, a

$$(|d_{i_0}| - |a_{i_0}| - |c_{i_0}|)\,\|\mathbf{e}_h\|_\infty \le h^2\|\mathbf{g}\|_\infty \tag{10.5.37}$$

egyenlőtlenséghez jutunk. Ez a (10.5.22) tulajdonság figyelembevételével a

$$q_{i_0}\|\mathbf{e}_h\|_\infty \le h^2\|\mathbf{g}\|_\infty \tag{10.5.38}$$

becslést jelenti. A $\min_{[a,b]} q := q_{\min} > 0$ feltétel következtében

$$\|\mathbf{e}_h\|_\infty \le h^2 \frac{\|\mathbf{g}\|_\infty}{q_{\min}} \le \tilde{C}h^2, \tag{10.5.39}$$

ahol

$$\tilde{C} = \left(\frac{M_4}{12} + \frac{p_{\max} M_3}{6}\right)/q_{\min}, \quad M_j = \max_{[a,b]} |u^{(j)}|, \quad p_{\max} = \max_{[a,b]} |p|.$$

Mindez azt jelenti, hogy $h \to 0$ esetén a hiba másodrendben tart nullához. ■

A 10.5.6. tétel arra az esetre vonatkozik, amikor a (10.5.20) feladatban $q(t) > 0$ az $[a, b]$ intervallumon. Így nem ad választ a (10.1.6) feladat $p = q = 0$ speciális esetére.

Tekintsük tehát az

$$\begin{gathered} u'' = r(t), \quad t \in (0, l), \\ u(0) = \alpha, \quad u(l) = \beta \end{gathered} \tag{10.5.40}$$

peremérték-feladatot. (Az egyszerűség kedvéért az intervallumot l hosszúságúnak tekintjük, és a bal oldali végpontot az origóba helyezzük.) Ebben az esetben tehát az $\overline{\omega}_h$ rácsháló

$$\overline{\omega}_h = \{t_i = ih, \quad i = 0, 1, \ldots, N+1, \ h = l/(N+1)\} \tag{10.5.41}$$

alakú. Ekkor a (10.5.21) rendszer mindegyik diszkretizácó esetén a (10.5.28) feladatot eredményezi, ahol

$$a_i = -\frac{1}{h^2}, \ d_i = \frac{2}{h^2}, \ c_i = -\frac{1}{h^2}. \tag{10.5.42}$$

Ezért a diszkrét feladat az

$$\mathbf{A}_h \mathbf{y}_h = \mathbf{b}_h \tag{10.5.43}$$

alakú lineáris algebrai egyenletrendszert jelenti, ahol az $\mathbf{A}_h \in \mathbb{R}^{N\times N}$ mátrix

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & 0 & -1 & 2 \end{pmatrix} \tag{10.5.44}$$

alakú, és a $\mathbf{b}_h \in \mathbb{R}^N$ vektorra pedig

$$\mathbf{b}_h = \begin{bmatrix} -r(t_1) + \alpha/h^2 \\ -r(t_i),\ i = 2,3,\dots,N-1 \\ -r(t_N) + \beta/h^2 \end{bmatrix}. \tag{10.5.45}$$

Először megmutatjuk, hogy a fenti diszkretizáció korrekt.

10.5.7. tétel.
A (10.5.43) feladatnak tetszőleges $h > 0$ esetén létezik egyértelmű megoldása.

Bizonyítás. Azt kell megmutatnunk, hogy bármely $h > 0$ mellett az $\mathbf{A}_h$ mátrix reguláris, azaz $\lambda = 0$ nem sajátértéke.
Határozzuk meg az $\mathbf{A}_h$ mátrix N darab $\lambda^{(k)}$ $(k = 1,2,\dots,N)$ sajátértékét! Mivel $\mathbf{A}_h$ szimmetrikus, ezért mindegyik λ_k valós, és valamely nem nulla $\mathbf{v}_k \in \mathbb{R}^N$ vektor mellett $\mathbf{A}_h\mathbf{v}_k = \lambda_k\mathbf{v}_k$, azaz (ideiglenesen elhagyva a k index jelölését) a $v_0 = v_{N+1} = 0$ értékekkel a $\mathbf{v}$ sajátvektor v_i koordinátáira teljesül a

$$\frac{-v_{i-1} + 2v_i - v_{i+1}}{h^2} = \lambda v_i, \quad i = 1,2,\dots,N \tag{10.5.46}$$

egyenlet. Ezzel a

$$\begin{gathered} v_{i-1} - 2(1 - 0.5\lambda h^2)v_i + v_{i+1} = 0, \quad i = 1,2,\dots,N, \\ v_0 = 0, \quad v_{N+1} = 0 \end{gathered} \tag{10.5.47}$$

feladatot kapjuk. Keressük a (10.5.47) feladat megoldását a

$$v_i = \sin(pt_i) \quad i = 0,1,\dots,N+1 \tag{10.5.48}$$

alakban, ahol $p \in \mathbb{R}$ egy egyelőre tetszőleges szám. Ekkor a (10.5.47) első egyenletébe behelyettesítve:

$$\sin(p(t_i - h)) - 2(1 - 0.5\lambda h^2)\sin(pt_i) + \sin(p(t_i + h)) = 0, \quad i = 1,2,\dots,N. \tag{10.5.49}$$

Felhasználva a közismert $\sin(p(t_i - h)) + \sin(p(t_i + h)) = 2\sin(pt_i)\cos(ph)$ azonosságot, a (10.5.49) összefüggés a

$$2\sin(pt_i)\cos(ph) - 2(1 - 0.5\lambda h^2)\sin(pt_i) = 0, \quad i = 1,2,\dots,N, \tag{10.5.50}$$

azaz a

$$\big(2\cos(ph) - 2(1 - 0.5\lambda h^2)\big)\sin(pt_i) = 0, \quad i = 1,2,\dots,N \tag{10.5.51}$$

feltételt jelenti. Mivel $\sin(pt_i) = v_i$ és a $\mathbf{v}$ sajátvektor nem nulla, ezért legalább egy i indexre $v_i \neq 0$. Ezért (10.5.51) következtében

$$2\cos(ph) - 2(1 - 0.5\lambda h^2) = 0. \tag{10.5.52}$$

Innen

$$\lambda = \frac{2}{h^2}(1 - \cos(ph)) = \frac{4}{h^2}\sin^2\frac{ph}{2}. \tag{10.5.53}$$

A p paramétert úgy választjuk meg, hogy a (10.5.47) második feltétele is teljesüljön, azaz a (10.5.48) megválasztás következtében $v_0 \equiv \sin(p \cdot 0) = 0$ és $v_{N+1} \equiv \sin(p \cdot l) = 0$ legyen. Az első feltétel nyilvánvalóan minden p esetén teljesül, míg a második egyenletből $pl = k\pi$, azaz $p = p_k = k\pi/l$. Ezt behelyettesítve a (10.5.53) összefüggésbe megkapjuk az $\mathbf{A}_h$ mátrix sajátértékeit:

$$\lambda_k = \frac{4}{h^2}\sin^2\frac{k\pi h}{2l}, \quad k = 1, 2, \ldots, N. \tag{10.5.54}$$

A (10.5.48) összefüggésből a $\mathbf{v}_k$ sajátvektorok is felírhatók:

$$\mathbf{v}_k = \left(\sin\frac{k\pi i h}{l}\right)_{i=1}^{N}, \quad k = 1, 2, \ldots, N. \tag{10.5.55}$$

Mivel λ_k függ h-tól, alkalmazzuk a $\lambda_k =: \lambda_k(h)$ jelölést! Könnyen látható, hogy $\lambda_k(h) > 0$ minden $k = 1, 2, \ldots, N$ esetén, és emellett a legkisebb sajátérték a $k = 1$ indexhez tartozó érték. Így tetszőlegesen rögzített h (azaz rögzített felosztás) esetén

$$\min_{k=1,2,\ldots,N} \lambda_k(h) = \lambda_1(h) = \frac{4}{h^2}\sin^2\frac{\pi h}{2l}. \tag{10.5.56}$$

Vezessük be az

$$s = \frac{\pi h}{2l} \tag{10.5.57}$$

új változót! Mivel $h \leq l/2$, ezért $s \in (0, \pi/4]$. Ekkor a legkisebb sajátérték felírható

$$\lambda_1(s) = \frac{\pi^2}{l^2}\left(\sin s/s\right)^2, \quad s \in (0, \pi/4] \tag{10.5.58}$$

alakban. Használjuk fel, hogy a $\sin x/x$ függvény a $(0, \pi/4]$ intervallumon monoton csökken.[11] Ezért $\min_{s\in(0,\pi/4]} \lambda_1(s) = \lambda_1(\pi/4)$, azaz

$$\inf_{h>0} \lambda_1(h) = \lambda_1\left(\frac{l}{2}\right) = \frac{16}{l^2}\sin^2\frac{\pi}{4} = \frac{8}{l^2}. \tag{10.5.59}$$

Így minden $h > 0$ esetén

$$\lambda_1(h) \geq \frac{8}{l^2}, \tag{10.5.60}$$

vagyis az $\mathbf{A}_h$ mátrix legkisebb sajátértéke minden $h > 0$ esetén a $\delta := 8/l^2 > 0$ számnál nagyobb. Ezzel beláttuk állításunkat. ■

10.5.8. következmény. A fenti tételből és a bizonyításából közvetlenül következnek az alábbi állítások.

[11] Ez az állítás elemi függvényvizsgálattal önállóan is könnyen belátható.

1. Az $\mathbf{A}_h$ mátrix invertálható, és $\|\mathbf{A}_h^{-1}\|_2 \leq 1/\delta = l^2/8$.

2. Az $\mathbf{A}_h$ mátrix szimmetrikus, szigorúan pozitív definit.

3. A (10.5.55) formulából látható, hogy $\mathbf{v}_1 > 0$. Mivel $\lambda_1(h) > 0$ is igaz, ezért az $\mathbf{A}_h\mathbf{v}_1 = \lambda_1(h)\mathbf{v}_1$ összefüggés alapján igaz, hogy a $\mathbf{v}_1 > 0$ vektorral $\mathbf{A}_h\mathbf{v}_1 > 0$, azaz $\mathbf{A}_h$ M-mátrix. (További részletekért lásd a 10.5.5. pontot.)

◇

10.5.9. megjegyzés. A fenti tétel bizonyításához a klasszikus algebrai megközelítést alkalmazzuk, és nem használtuk a korábbi (az M-mátrixra vonatkozó) korábbi eredményeinket, amelyből a regularitás szintén következik. Ezen megközelítés egyik fontos eredménye a 10.5.8. következményben szereplő becslés a $\|\mathbf{A}_h^{-1}\|_2$ normára, hiszen ennek segítségével az ezen normabeli konvergencia és annak sebessége rendjében szereplő állandó nagysága is megmutatható. (Lásd a 10.5.10. tételt.) ◇

Eredményünket felhasználva közvetlenül belátható a (10.5.28), (10.5.42) numerikus módszer konvergenciája az $\overline{\omega}_h := \{t_0, t_1, \ldots, t_{N+1}\}$, h lépésközű ekvidisztáns rácshálón értelmezett y_h rácsfüggvények

$$\|y_h\|_{2,h}^2 := h \sum_{i=0}^{N+1} y_i^2 \tag{10.5.61}$$

alakú normájában.[12]

Megjegyezzük, hogy ekkor a megfelelő mátrixnormára

$$\|\mathbf{A}\|_{2,h} = \sup_{\mathbf{v}\in\mathbb{R}^K} \frac{\|\mathbf{A}\mathbf{v}\|_{2,h}}{\|\mathbf{v}\|_{2,h}} = \sup_{\mathbf{v}\in\mathbb{R}^K} \frac{h\|\mathbf{A}\mathbf{v}\|_2}{h\|\mathbf{v}\|_2} = \|\mathbf{A}\|_2,$$

tehát az 1. fejezetben definiált szokásos 2-es (vagy euklideszi) normával számolható.

Tekintsük tehát az

$$\begin{gathered} \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = r_i, \quad i = 1, 2, \ldots, N \\ y_0 = \alpha, \quad y_{N+1} = \beta \end{gathered} \tag{10.5.62}$$

feladatot, és mutassuk meg, hogy a megoldása a $\|\cdot\|_{2,h}$ normában $h \to 0$ estén tart a (10.5.40) peremérték-feladat megoldásához. Jelölje ismét $e_i = y_i - u(t_i)$ a t_i pontbeli hibát! Ekkor $y_i = e_i + u(t_i)$, és ezt behelyettesítve a (10.5.62) sémába a hibafüggvényre a

$$\begin{gathered} \frac{-e_{i-1} + 2e_i - e_{i+1}}{h^2} = -r_i + \frac{u(t_{i-1}) - 2u(t_i) + u(t_{i+1})}{h^2} \equiv \psi_i, \quad i = 1, 2, \ldots, N \\ e_0 = 0, \quad e_{N+1} = 0 \end{gathered} \tag{10.5.63}$$

feladatot kapjuk. A (10.5.63) feladat felírható

$$\mathbf{A}_h\mathbf{e}_h = \mathbf{\Psi}^h \tag{10.5.64}$$

[12] Könnyen látható, hogy az $a = t_0$ és $b = t_{N+1}$ következtében ez a norma az $L_2[a,b]$ tér normájának diszkrét analógja, azaz ha u az $[a,b]$ intervallumon értelmezett, négyzetesen integrálható függvény, és $y_h(x_i) = u(x_i)$, akkor $\lim_{h\to 0} \|y_h\|_{2,h}^2 = \|u\|_{L_2[a,b]}^2 \equiv \int_a^b u^2(x)dx$.

alakú lineáris algebrai egyenletrendszerként, ahol az $\mathbf{A}_h \in \mathbb{R}^{N\times N}$ mátrix a (10.5.44) alakú, a $\mathbf{\Psi}^h \in \mathbb{R}^N$ vektorra pedig $\mathbf{\Psi}^h)_i = \psi_i$, $i = 1, 2, \ldots, N$. A (10.5.64) egyenletből

$$\mathbf{e}_h = \mathbf{A}_h^{-1}\mathbf{\Psi}^h, \tag{10.5.65}$$

azaz

$$\|\mathbf{e}_h\|_{2,h} \le \|\mathbf{A}_h^{-1}\|_{2,h}\|\mathbf{\Psi}^h\|_{2,h} = \|\mathbf{A}_h^{-1}\|_2\|\mathbf{\Psi}^h\|_{2,h}. \tag{10.5.66}$$

Mivel (10.5.5) következtében $\psi_i = \mathcal{O}(h^2)$, ezért

$$\|\mathbf{\Psi}^h\|_{2,h}^2 = h\sum_{i=1}^{N}\psi_i^2 = h\sum_{i=1}^{N}\mathcal{O}(h^4) = hN\mathcal{O}(h^4) = \mathcal{O}(h^4),$$

mivel $hN = const.$ Ezért tehát $\|\mathbf{\Psi}^h\|_{2,h} = \mathcal{O}(h^2)$. Másrészt a 10.5.8. következmény szerint $\|\mathbf{A}_h^{-1}\|_2 \le l^2/8$. Így (10.5.66) alapján

$$\|\mathbf{e}_h\|_{2,h} = \mathcal{O}(h^2). \tag{10.5.67}$$

Ezzel beláttuk az alábbi állítást.

10.5.10. tétel.
A (10.5.62) séma megoldása $h \to 0$ esetén másodrendben tart a $\|\cdot\|_{2,h}$ normában a (10.5.40) peremérték-feladat megoldásához.

10.5.4. A lineáris peremérték-feladatok numerikus megoldásának általános vizsgálata

Ebben a szakaszban ismételten a lineáris peremérték-feladatok numerikus megoldásával és a közelítések konvergenciájával foglalkozunk, de most az eredményeinket általánosan fogalmazzuk meg. Megmutatjuk, hogy ezen általános megközelítésből speciális esetként az előző szakaszban leírt eredményeink hogyan nyerhetők.

Jelölje L_0 azt az operátort, amely a $C^2[a,b]$-beli függvényeken van értelmezve, és

$$L_0u = -u'' + p(t)u' + q(t)u, \tag{10.5.68}$$

valamint B_1, B_2 azokat a szintén $C^2[a,b]$-beli függvényeken értelmezett operátorokat, amelyekre

$$B_1u = u(a), \quad B_2u = u(b). \tag{10.5.69}$$

Ekkor $L_0 : C^2[a,b] \to C[a,b]$, $B_1, B_2 : C^2[a,b] \to \mathbb{R}$ típusú lineáris operátorok. Vezessük be az

$$\mathbb{F}_1[a,b] = C^2[a,b], \quad \mathbb{F}_2[a,b] = C[a,b]\times\mathbb{R}\times\mathbb{R} \tag{10.5.70}$$

jelöléseket! Legyen L az az operátort, amely az $\mathbb{F}_1[a,b]$ halmazon van értelmezve, és

$$Lv = \begin{bmatrix} L_0v \\ B_1v \\ B_2v \end{bmatrix}. \tag{10.5.71}$$

Tehát L egy $\mathbb{F}_1[a,b] \to \mathbb{F}_2[a,b]$ lineáris operátor, amely egy $v \in C^2[a,b]$ függvényhez az

$$Lv = \begin{bmatrix} -v'' + p(t)v' + q(t)v \\ v(a) \\ v(b) \end{bmatrix} \in \mathbb{F}_2[a,b] \tag{10.5.72}$$

hármast rendeli hozzá.
Ekkor a (10.5.20) feladat felírható a következő módon. Az

$$f := \begin{bmatrix} -r(t) \\ \alpha \\ \beta \end{bmatrix} \in \mathbb{F}_2[a,b] \tag{10.5.73}$$

jelölés mellett keressük azon $u \in F_1[a,b]$ elemet (azaz $C^2[a,b]$-beli függvényt), amelyre

$$Lu = f. \tag{10.5.74}$$

Jelölje $\mathbb{F}(\overline{\omega}_h)$ illetve $\mathbb{F}(\omega_h)$ az $\overline{\omega}_h$ illetve ω_h rácshálókon értelmezett rácsfüggvények vektorterét, $L_{0,h}^{(1)} : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\omega_h)$ pedig azt az operátort, amely valamely $v_h \in \mathbb{F}(\overline{\omega}_h)$ függvény esetén a következő módon hat:

$$\left(L_{0,h}^{(1)} v_h\right)(t) = -\frac{v_h(t+h) - 2v_h(t) + v_h(t-h)}{h^2} + p_i \frac{v_h(t+h) - v_h(t)}{h} + q_i v_h(t), \ t \in \omega_h. \tag{10.5.75}$$

Jelölje továbbá $B_{1,h}, B_{2,h}$ azokat az $\mathbb{F}(\overline{\omega}_h)$-beli rácsfüggvényeken értelmezett operátorokat, amelyekre

$$B_{1,h} v_h = v_h(t_0 = a), \quad B_{1,h} v_h = v_h(t_{N+1} = b). \tag{10.5.76}$$

Ekkor tehát $L_{0,h} : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\omega_h)$, $B_{1,h}, B_{2,h} : \mathbb{F}(\overline{\omega}_h) \to \mathbb{R}$ típusú lineáris operátorok. Jelölje $L_h^{(1)}$ azt az operátort, amely az $\mathbb{F}(\overline{\omega}_h)$ halmazon van értelmezve és

$$L_h^{(1)} v_h = \begin{bmatrix} L_{0,h}^{(1)} v_h \\ B_{1,h} v_h \\ B_{2,h} v_h \end{bmatrix}. \tag{10.5.77}$$

Tehát $L_h^{(1)}$ egy $\mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\omega_h) \times \mathbb{R} \times \mathbb{R} \equiv \mathbb{F}(\overline{\omega}_h)$ típusú lineáris operátor, amely a v_h ($\overline{\omega}_h$ rácsháló pontjaiban definiált) rácsfüggvényhez az

$$L_h^{(1)} v_h = \begin{bmatrix} -\frac{v_h(t+h) - 2v_h(t) + v_h(t-h)}{h^2} + p_i \frac{v_h(t+h) - v_h(t)}{h} + q_i v_h(t), \ t \in \omega_h \\ v_h(a) \\ v_h(b) \end{bmatrix} \in \mathbb{F}(\overline{\omega}_h) \tag{10.5.78}$$

(ugyancsak az $\overline{\omega}_h$ rácsháló pontjaiban definiált) rácsfüggvényt rendeli hozzá.
Jelölje r_h azt az $\mathbb{F}(\omega_h)$-beli rácsfüggvényt, amelyre $r_h(t) = r(t)$ mindegyik $t \in \omega_h$ pontban. Ekkor a (10.5.23) feladat felírható a következő módon. Az

$$f_h := \begin{bmatrix} -r_h \\ \alpha \\ \beta \end{bmatrix} \in \mathbb{F}(\overline{\omega}_h) \tag{10.5.79}$$

jelölés mellett keressük azon $v_h \in \mathbb{F}(\overline{\omega}_h)$ elemet (vagyis az $\overline{\omega}_h$ rácson értelmezett rácsfüggvényt), amelyre

$$L_h^{(1)} v_h = f_h. \tag{10.5.80}$$

Vegyük észre, hogy (10.5.80) – a jelöléseinket figyelembe véve – egy lineáris algebrai egyenletrendszert jelent, amelynek mérete megegyezik az $\overline{\omega}_h$ rácsháló pontjainak a számával, azaz $N+2$ ismeretlent tartalmaz. Tehát az ismeretlen v_h illetve az adott f_h rácsfüggvényeket $N+2$ dimenziós vektorokkal, míg az $L_h^{(1)}$ lineáris operátort egy $(N+2)\times(N+2)$ dimenziós mátrix segítségével adhatjuk meg. Ez azt jelenti, hogy a (10.5.80) feladat felírható

$$\overline{\mathbf{A}}_h^{(1)}\overline{\mathbf{y}}_h = \bar{\mathbf{f}}_h \tag{10.5.81}$$

alakban, ahol $\overline{\mathbf{A}}_h^{(1)} \in \mathbb{R}^{(N+2)\times(N+2)}$ és $\overline{\mathbf{y}}_h, \bar{\mathbf{f}}_h \in \mathbb{R}^{N+2}$. Mivel a vektorok i-edik koordinátája a $t_i \in \overline{\omega}_h$ rácsponthoz tartozó értékeket jelenti, ezért a (10.5.81) egyenletben

$$(\overline{\mathbf{y}}_h)_i = v_h(t_i), \quad (\bar{\mathbf{f}}_h)_i = f_h(t_i), \quad i = 0,1,\dots,N+1, \tag{10.5.82}$$

és (10.5.78) alapján az $\overline{\mathbf{A}}_h^{(1)}$ mátrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ a_1 & d_1 & c_1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & a_2 & d_2 & c_2 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & a_N & d_N & c_N \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{pmatrix} \tag{10.5.83}$$

alakú, ahol a_i, d_i, c_i értékei a (10.5.24) képlet szerintiek.

Mivel $(\overline{\mathbf{y}}_h)_0$ és $(\overline{\mathbf{y}}_h)_{N+1}$ értékei ismertek (a peremfeltételek miatt ezek értéke α és β), ezért a (10.5.81) egyenlet ekvivalens az

$$\mathbf{A}_h^{(1)}\mathbf{y}_h = \mathbf{f}_h \tag{10.5.84}$$

feladattal, ahol $\mathbf{A}_h^{(1)} \in \mathbb{R}^{N\times N}$, $\mathbf{y}_h, \mathbf{f}_h \in \mathbb{R}^N$ és

$$\begin{gathered}(\mathbf{f}_h)_1 = (\bar{\mathbf{f}}_h)_1 - a_1\alpha, \quad (\mathbf{f}_h)_i = (\bar{\mathbf{f}}_h)_i, \ i = 2,3,\dots,N-1, \quad (\mathbf{f}_h)_N = (\bar{\mathbf{f}}_h)_N - c_N\beta, \\ (\mathbf{y}_h)_i = (\overline{\mathbf{y}}_h)_i, \ i = 1,2,\dots,N, \\ \mathbf{A}_h^{(1)} = \begin{pmatrix} d_1 & c_1 & 0 & \dots & 0 & 0 & 0 \\ a_2 & d_2 & c_2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & a_N & d_N \end{pmatrix}. \end{gathered} \tag{10.5.85}$$

A (10.5.84)-(10.5.85) feladat megegyezik a (10.5.28) feladattal.

10.5.11. megjegyzés. A másik két diszkretizációra is az értelemszerűen definiált $\mathbf{A}_h^{(2)}$ és $\mathbf{A}_h^{(3)}$ mátrixokkal (10.5.84) érvényes. ◇

A továbbiakban arra vagyunk kíváncsiak, hogy a fenti sémák valamelyikével definiált numerikus megoldások használhatók-e az eredeti peremérték-feladat megoldásának közelítésére.

Ezen szakasz elején megmutattuk a következőket.

- A 10.5.2. tétel alapján az alkalmasan megválasztott h értékek mellett az $\overline{\mathbf{A}}_h^{(k)}$ mátrixok szigorúan diagonálisan dominánsak, ezért regulárisak. Tehát megfelelően kis h esetén az $L_h^{(k)}v_h = f_h$ $(k = 1,2,3)$ feladatok mindegyike egyértelműen megoldható.

- A 10.5.6. tétel alapján az e_h hibafüggvény nullához tart, azaz a közelítő megoldások sorozata $h \to 0$ esetén tart a pontos megoldáshoz. Ennek belátása két lépésben történt:
  1. Felírtuk a (10.5.33) hibaegyenletet.
  2. Megmutattuk, hogy ezen rendszer megoldásai ($h \to 0$ esetén) nullához tartanak.

10.5.12. megjegyzés. Ezzel kapcsolatosan két dologra hívjuk fel a figyelmet.

1. A (10.5.33) hibaegyenlet jobb oldalán szereplő kifejezés a lokális approximációs hiba. Tehát a hibaegyenlet jobb oldalán egy olyan vektor áll, amelynek i-edik koordinátája azt mutatja meg, hogy a $t_i \in \omega_h$ pontban a rácsháló pontjaiban értelmezett pontos megoldás $L_h^{(k)}$-képe milyen közel van a pontos megoldás L-képének értékéhez.
2. A hibavektort a (10.5.33) rendszer megoldásával állítjuk elő. Ezért a globális hiba viselkedése (nevezetesen nullához tartása) azon múlik, hogy a rendszer jobb oldalán szereplő vektorok hogyan viselkednek a rendszer együtthatómátrixának inverzén, azaz $h \to 0$ esetén az $\left(\overline{\mathbf{A}}_h^{(k)}\right)^{-1}$ mátrixok hogyan hatnak a lokális approximációs hibavektorra.

◇

A továbbiakban egy általános tárgyalásmódot adunk eredményeink leírására. Legyenek $\mathbb{F}_1[a,b] = C^2[a,b]$ és $\mathbb{F}_2[a,b] = C[a,b] \times \mathbb{R} \times \mathbb{R}$ adott függvényterek, $L : \mathbb{F}_1[a,b] \to \mathbb{F}_2[a,b]$ egy olyan lineáris operátor, amelyre minden $f \in \mathbb{F}_2[a,b]$ esetén az $Lu = f$ egyenlet korrekt kitűzésű. Célunk ezen feladat diszkretizációjának meghatározása és vizsgálata.

Legyen szokásosan $\mathbb{F}(\overline{\omega}_h)$ az $\overline{\omega}_h$ rácson értelmezett rácsfüggvények vektortere, $L_{0,h} : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\omega_h)$ egy adott lineáris operátor, azaz egy olyan leképezés, amely egy $\overline{\omega}_h$ rácson értelmezett rácsfüggvényhez valamilyen lineáris leképezési szabály szerint megfeleltet egy ω_h-n értelmezett másik rácsfüggvényt. Legyenek $B_{1,h}$ és $B_{2,h}$ $\mathbb{F}(\overline{\omega}_h) \to \mathbb{R}$ tipusú, szintén lineáris leképezések. Képezzük ezen operátorok segítségével a következő

$$L_h v_h = \begin{bmatrix} L_{0,h} v_h \\ B_{1,h} v_h \\ B_{2,h} v_h \end{bmatrix} \tag{10.5.86}$$

$\mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\overline{\omega}_h)$ típusú lineáris operátor. (Tehát L_h jelentheti az $L_h^{(k)}$ $(k = 1,2,3)$ operátorok valamelyikét, de jelenthet valamely más lineáris operátort is.) A továbbiakban az L_h operátornak megfeleltetett mátrixokat (korábbiakkal megegyező módon) $\overline{\mathbf{A}}_h$ illetve $\mathbf{A}_h$-val jelöljük. Legyen $P_h^{(2)}$ egy olyan leképezés, amely egy $\mathbb{F}_2[a,b]$ elemnek megfeleltet egy $\mathbb{F}(\overline{\omega}_h)$-beli rácsfüggvényt, és jelölje

$$f_h := P_h^{(2)} f \in \mathbb{F}(\overline{\omega}_h). \tag{10.5.87}$$

A diszkretizációs eljárások általános tárgyalása során a következő kérdések megválaszolása szükséges.

1. Létezik-e az
$$L_h v_h = f_h \tag{10.5.88}$$
feladatnak egyértelmű megoldása?
2. Hogyan értelmezhető egy $\mathbb{F}(\overline{\omega}_h)$-beli rácsfüggvény és egy $\mathbb{F}_i[a,b]$-beli függvény $(i = 1,2)$ távolsága?

3. Milyen közel van a $v_h \in \mathbb{F}(\overline{\omega}_h)$ függvény a (10.5.74) egyenlet $u \in \mathbb{F}_1[a,b]$ megoldásához?

4. A $h \to 0$ esetben milyen feltételek mellett tart a közelítő és a pontos megoldás eltérése nullához?

A továbbiakban ezekkel a kérdésekkel foglalkozunk.

1. Az általános tárgyalás során mindig feltesszük, hogy a (10.5.88) feladatnak létezik egyértelmű megoldása. Ugyanakkor minden konkrét L_h megválasztás esetén ezt be kell bizonyítani, azaz azt kell megmutatni, hogy az $\overline{\mathbf{A}}_h$ (avagy az $\mathbf{A}_h$) mátrix reguláris.

2. Ennél a kérdésnél az az alapvető probléma, hogy az $[a,b]$ intervallumon értelmezett $w \in F_1[a,b]$ és az $\overline{\omega}_h$ rácspontokban értelmezett $g_h \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvény különböző alaphalmazokon vannak értelmezve. Ezért legyen $P_h^{(1)}$ egy $\mathbb{F}_1[a,b] \to \mathbb{F}(\overline{\omega}_h)$ típusú lineáris leképezést. (Tehát $P_h^{(1)}$ egy olyan leképezés, amely az intervallumon értelmezett függvényt leképezi egy, a rácspontokban értelmezett függvényre.) Ekkor a két függvény távolságán a $\|P_h^{(1)}w - g_h\|_h$ számot értjük, ahol $\|\cdot\|_h$ egy $\mathbb{F}(\overline{\omega}_h)$ térbeli rögzített norma.

3. Legyen a $v_h \in \mathbb{F}(\overline{\omega}_h)$ függvény a (10.5.88) feladat megoldása. Jelölje u_h azt az $\mathbb{F}(\overline{\omega}_h)$-beli rácsfüggvényt, amelyet a (10.5.74) egyenlet $u \in \mathbb{F}_1[a,b]$ megoldásának $P_h^{(1)}$-képeként kapunk, azaz

$$u_h = P_h^{(1)}u, \tag{10.5.89}$$

továbbá $e_h \in \mathbb{F}(\overline{\omega}_h)$ azt a rácsfüggvényt, amelyre

$$e_h = v_h - u_h. \tag{10.5.90}$$

Ekkor az e_h rácsfüggvényt a v_h közelítő megoldáshoz tartozó *globális hibának*, és (a 2. pontnak megfelelően) a $\|e_h\|_h$ értéket a közelítő megoldás hibájának nevezzük. A vizsgált módszert akkor nevezzük *konvergensnek*, amikor $\lim_{h\to 0}\|e_h\|_h = 0$. Ha $\|e_h\|_h = \mathcal{O}(h^p)$, akkor a konvergenciát p-ed rendűnek nevezzük.

10.5.13. megjegyzés. A numerikus módszer alkalmazásával az a célunk, hogy a numerikus megoldások a (10.5.74) folytonos feladat megoldásának $u_h = P_h^{(1)}u$ képéhez kerüljenek közel, és ne egy esetleges más elem $P_h^{(1)}$ képéhez. Ezért a $\|\cdot\|_h$ normára kikötjük a következő ún. *kompatibilitási* feltételt. Legyen $L : \mathbb{F}_1[a,b] \to \mathbb{F}_2[a,b]$, és jelölje $\|\cdot\|$ az $\mathbb{F}_1[a,b]$ térbeli normát. Azt mondjuk, hogy a $\|\cdot\|_h$ norma *kompatibilis* a $\|\cdot\|$ normával, ha $\lim_{h\to 0}\|P_h^{(1)}w\|_h = \|w\|$ minden $w \in \mathbb{F}_1[a,b]$ esetén. Ekkor megmutatható a kívánt egyértelműség.[13] ◇

4. Az előző pontban definiált konvergencia biztosítására alkalmazzuk az $L_h = L_h^{(k)}$ megválasztás esetén alkalmazott módszerünket. (Lásd a 10.5.12. megjegyzést.) Legyen tetszőleges $w \in \mathbb{F}_1[a,b]$ esetén

$$l_h(w) = L_h(P_h^{(1)}w) - P_h^{(2)}(Lw) \in \mathbb{F}(\overline{\omega}_h). \tag{10.5.91}$$

10.5.14. definíció.

Azt mondjuk, hogy az L_h *operátor konzisztens az* L *operátorral*, ha valamely $p > 0$ mellett $l_h(w) = \mathcal{O}(h^p)$ minden $w \in \mathbb{F}_2[a,b]$ függvényre, és a p számot a *konzisztencia rendjének* nevezzük.

[13] Ugyanis, indirekt módon tegyük fel, hogy léteznek olyan $w_1, w_2 \in \mathbb{F}_1[a,b]$, $(w_1 \neq w_2)$ elemek, amelyekre $\lim_{h\to 0}\|w_h - P_h^{(1)}w_1\|_h = 0$ és $\lim_{h\to 0}\|w_h - P_h^{(1)}w_2\|_h = 0$. Ekkor $\|P_h^{(1)}w_1 - P_h^{(1)}w_2\|_h = \|P_h^{(1)}w_1 - w_h + w_h - P_h^{(1)}w_2\|_h \le \|P_h^{(1)}w_1 - w_h\|_h + \|w_h - P_h^{(1)}w_2\|_h \to 0$. Mivel a kompatibilitási feltételt felhasználva ekkor $\|w_1 - w_2\| = \lim_{h\to 0}\|P_h^{(1)}w_1 - P_h^{(1)}w_2\|_h = 0$, ezért $w_1 = w_2$, ami állításunkat igazolja. ■

A differenciaséma egy másik fontos tulajdonsága (a differenciálegyenletek elméletéhez hasonlóan) a *stabilitás*, amely azt fejezi ki, hogy a megoldás folytonosan függ a feladatot meghatározó adatoktól. (Más kifejezéssel: két közeli adatokhoz tartozó megoldás is közel van egymáshoz.) Ezt a fogalmat definiáljuk a következőben.

10.5.15. definíció.

Azt mondjuk, hogy a (10.5.88) feladat stabil kitűzésű (a továbbiakban: *az L_h operátor stabil*), ha minden $f_h \in \mathbb{F}(\overline{\omega}_h)$ esetén létezik a feladatnak egyértelmű megoldása, és a megoldásokra teljesül a

$$\|v_h\|_h \leq C\|f_h\|_h \tag{10.5.92}$$

egyenlőtlenség, ahol $C > 0$ egy, a h-tól független állandó.

A stabilitás (10.5.92) tulajdonságának ellenőrzése elég nehéz. Megmutatható[14], hogy ez a fogalom ekvivalens az L_h operátorsereg következő, gyakran könnyebben ellenőrizhető tulajdonságával: az invertálható L_h operátorokhoz létezik olyan $C \geq 0$ h-tól független állandó, amely mellett

$$\|L_h^{-1}\|_h \leq C. \tag{10.5.93}$$

10.5.16. tétel.

Tegyük fel, hogy

1. L_h p-ed rendben konzisztens az L operátorral,
2. az L_h operátor stabil.

Ekkor a numerikus módszer p-ed rendben konvergens, azaz

$$\|e_h\|_h = \mathcal{O}(h^p). \tag{10.5.94}$$

Bizonyítás. Felhasználva az e_h globális hiba (10.5.90) definícióját, illetve a (10.5.89) definíciót, a következő egyenlőség érvényes:

$$L_h e_h = L_h u_h - L_h v_h = L_h(P_h^{(1)}u) - L_h v_h = \underbrace{L_h(P_h^{(1)}u) - P_h^{(2)}(Lu)}_{=:l_h(u)} + \underbrace{P_h^{(2)}(Lu) - L_h v_h}_{=:l_h(f)}. \tag{10.5.95}$$

A (10.5.74) egyenlet mindkét oldalának $P_h^{(2)}$-képét véve és figyelembe véve a (10.5.87) összefüggést,

$$P_h^{(2)}(Lu) = P_h^{(2)} f = f_h. \tag{10.5.96}$$

Másrészt, a (10.5.88) egyenlet alapján $L_h v_h = f_h$. Így a (10.5.95) relációban $l_h(f) = 0$, tehát

$$L_h e_h = l_h(u). \tag{10.5.97}$$

Innen a stabilitás (10.5.92) tulajdonsága miatt érvényes az

$$\|e_h\|_h \leq C\|l_h(u)\|_h = \mathcal{O}(h^p) \tag{10.5.98}$$

[14] A (10.5.92) és a (10.5.93) tulajdonság ekvivalenciáját a következők módon láthatjuk be. Mivel $v_h = L_h^{-1} f_h$, a norma tulajdonságai miatt $\|v_h\|_h = \|L_h^{-1} f_h\|_h \leq \|L_h^{-1}\|_h \|f_h\|_h$, és így a (10.5.93) $\Rightarrow$ (10.5.92) implikáció nyilvánvaló. Másrészt, (10.5.92) esetén (ismételten a $v_h = L_h^{-1} f_h$ egyenlőség miatt) $\|L_h^{-1} f_h\|_h \leq C\|f_h\|_h$, azaz $\|L_h^{-1} f_h\|_h / \|f_h\|_h \leq C$, ami a norma definíciója miatt a (10.5.92) $\Rightarrow$ (10.5.93) implikációt bizonyítja.

egyenlőtlenség, amely a tétel állítását bizonyítja. ■

10.5.17. megjegyzés. A (10.5.93) feltételt a *numerikus módszer stabilitásának* is szokásos nevezni. Tehát leegyszerűsítve a 10.5.16. tétel így fogalmazható meg: "konzisztencia + stabilitás = konvergencia". ◇

10.5.5. A lineáris peremérték-feladatok M-mátrixokkal

A 10.5.4. szakaszban megfogalmaztuk a peremérték-feladatok véges differenciás approximációjának azon elméleti kérdéseit, amelyeket egy adott L_h-módszer esetén megválaszolni szükséges. Nevezetesen, be kell látnunk, hogy mindegyik h esetén

1. az L_h lineáris operátort reprezentáló $\overline{\mathbf{A}}_h$ mátrix reguláris;
2. érvényes a stabilitás, azaz teljesül a (10.5.93) stabilitási tulajdonság:

$$\|(\overline{\mathbf{A}}_h)^{-1}\|_h \leq C, \tag{10.5.99}$$

ahol C egy h-tól független állandó.

A továbbiakban megmutatjuk, hogy amikor mindegyik $\overline{\mathbf{A}}_h$ speciális tulajdonságú M-mátrix, akkor a maximumnormában a válasz viszonylag egyszerűen megfogalmazható. Emellett eredményeinket felhasználva egységesen megmutatjuk a 10.5.3. szakaszban megfogalmazott numerikus sémák konvergenciáját, illetve eredményeinket kiterjesztjük újabb feladatosztályra is.

Legyenek $\mathbf{M}_h \in \mathbb{R}^{k\times k}, h = 1/(k-1), k = 3, 4, \ldots$ adott ún. Z-mátrixok sorozata, azaz az $\mathbf{M}_h$ mátrix mindegyik főátlón kívüli eleme nempozitív. (Tehát $(\mathbf{M}_h)_{ij} \leq 0$ minden $i \neq j$ esetén.)

10.5.18. definíció.

Azt mondjuk, hogy az ilyen $\mathbf{M}_h$ mátrixsorozat *egyenletes M-mátrix*, ha léteznek olyan $\mathbf{g}_h \in \mathbb{R}^k$, $\mathbf{g}_h > \mathbf{0}$ vektorok és $0 < g_1, g_2 < \infty$ állandók, amelyekre

$$(\mathbf{M}_h\mathbf{g}_h)_i \geq g_1, \tag{10.5.100}$$

$$\|\mathbf{g}_h\|_\infty \leq g_2, \tag{10.5.101}$$

mindegyik h esetén.

Az egyenletes M-mátrix tulajdonság egy, nyilvánvalóan M-mátrixokokból álló halmazra, azaz egy mátrixseregre vonatkozik. Az egyszerűség kedvéért a definíciót az egyes elemek tulajdonságaiként fogalmaztuk meg.

Az 1.2.40. tételből a (10.5.100) és a (10.5.101) tulajdonságok felhasználásával közvetlenül nyerjük az alábbi állítást.

10.5.19. lemma.

Tegyük fel, hogy $\mathbf{M}_h$ egyenletes M-mátrix. Ekkor

$$\|\mathbf{M}_h^{-1}\|_\infty \leq \frac{g_2}{g_1}. \tag{10.5.102}$$

Eredményeinket az alábbiakban foglalhatjuk össze.

10.5.20. tétel.
Tegyük fel, hogy a (10.5.88) alakú diszkretizáció

a. a maximumnormában p-ed rendben konzisztens,

b. az L_h operátornak megfeleltetett $\overline{\mathbf{A}}_h$ egyenletes M-mátrix.

Ekkor a numerikus módszer a maximumnormában p-ed rendben konvergens.

Bizonyítás. Mivel a 10.5.18. definíció alapján mindegyik $\overline{\mathbf{A}}_h$ M-mátrix, ezért reguláris is. A 10.5.19. lemma alapján a feladat stabil kitűzésű is, azaz érvényes a (10.5.99) tulajdonság a $C = g_2/g_1$ állandóval. Ezért a 10.5.16. tételünk alapján az állításunk igaz. ■

10.5.21. megjegyzés. A (10.5.98) összefüggés alapján becslést is adhatunk a hibafüggvényre:

$$\|e_h\|_\infty \le \frac{g_2}{g_1}\|l_h(u)\|_\infty. \tag{10.5.103}$$

◇

Alkalmazzuk a 10.5.20. tételt a 10.5.3. fejezetben vizsgált $L_h^{(k)}$ $(k = 1, 2, 3)$ operátorokra! A (10.5.24), valamint a (10.5.26) és a (10.5.27) összefüggésekből látható, hogy mindhárom esetben megfelelően kis h esetén az $\overline{\mathbf{A}}_h^{(k)}$ (10.5.83) alakú mátrixok diagonálon kívüli elemei nempozitívak, a diagonális elemek pedig pozitívak lesznek. Így a (10.5.22) összefüggés miatt az $\overline{\mathbf{A}}_h^{(k)}$ mátrix szigorúan diagonálisan domináns M-mátrix. Legyen $\mathbf{g}_h = [q_{\min}, 1, 1, \dots, 1, q_{\min}]^T \in \mathbb{R}^{N+2}$. Ekkor nyilvánvalóan $\|\mathbf{g}_h\|_\infty \le g_2 := \max\{1, q_{\max}\} > 0$, és a (10.5.22) következtében

$$\overline{\mathbf{A}}_h^{(k)}\mathbf{g}_h = \mathbf{q}_h,$$

ahol $\mathbf{q}_h = [q_{\min}, q_1, q_2, \dots, q_N, q_{\min}]^T \in \mathbb{R}^{N+2}$. Legyen $g_1 = \min\{1, q_{\min}\}$. A fenti g_1 és g_2 megválasztással a (10.5.100) és a (10.5.101) összefüggések egyaránt érvényesek, ezért teljesül a (10.5.102) stabilitási feltétel, és így (10.5.103) érvényes. Mivel a konzisztencát $(P_h^{(1)}w)(t_i) = (P_h^{(2)}w)(t_i) = w(t_i)$ megválasztással mindhárom operátorra már korábban megmutattuk, a tétel állítása igaz. ■

10.5.22. megjegyzés. Vegyük észre, hogy a konvergencia belátásához nekünk a (10.5.33) hibaegyenletet kell vizsgálnunk, azaz azt kell megmutatni, hogy az L_h operátorok inverzei korlátosak maradnak a $[0, e_1, e_2, \dots, e_N, 0]$ típusú vektorokon. Ez azt jelenti, hogy a (10.5.99) feltétel helyett elegendő belátni az

$$\|\mathbf{A}_h^{-1}\|_h \le C \tag{10.5.104}$$

tulajdonságot. Ezért az $L_h = L_h^{(k)}$ esetén a $\mathbf{g}_h = [1, 1, \dots, 1] \in \mathbb{R}^N$ egy alkalmas megválasztás. Ebben az esetben $g_1 = q_{\min}$ és $g_2 = 1$, és ekkor a (10.5.103) becslés megegyezik a (10.5.39) becsléssel. ◇

Befejezésül térjünk át a (10.5.20) azon speciális esetére, amikor $p = q = 0$, azaz tekintsük az

$$\begin{aligned} u'' &= r(t), \quad t \in (0,1), \\ u(0) &= \alpha, \quad u(1) = \beta \end{aligned} \tag{10.5.105}$$

peremérték-feladatot. A feladat tárgyalását az indokolja, hogy

1. A fejezet elején ismertetett motivációs feladat ilyen típusú peremérték-feladathoz vezet. (Az egyszerűség kedvéért az intervallumot egységnyi hosszúságúnak tekintjük.) Az előzőekben a konvergenciát csak a speciális $\|\cdot\|_{2,h}$ normában mutattuk meg, a maximumnormában viszont nem.

2. Az eddigi maximumnormabeli konvergenciára vonatkozó eredményeink nem alkamazhatók, mivel korábban feltettük, hogy $p > 0$.

A szokásos diszkretizáció után a megoldandó lineáris algebrai egyenletrendszerünk felírható

$$\overline{\mathbf{A}}_h^{(4)} \overline{\mathbf{y}}_h = \bar{\mathbf{f}}_h \tag{10.5.106}$$

alakban, ahol az $\overline{\mathbf{A}}_h^{(4)} \in \mathbb{R}^{(N+2)\times(N+2)}$ mátrix alakja

$$\frac{1}{h^2}\begin{pmatrix} h^2 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & h^2 \end{pmatrix} \tag{10.5.107}$$

és $\bar{\mathbf{f}}_h \in \mathbb{R}^{N+2}$ a korábban definiált vektor. Könnyen látható, hogy $\overline{\mathbf{A}}_h^{(4)}$ nem marad szigorúan diagonálisan domináns. Ugyanakkor megmutatjuk, hogy egyenletes M-mátrix. Ehhez elegendő megmutatni, hogy található hozzá $\mathbf{g}_h$ vektor a megkövetelt tulajdonságokkal.
Legyen

$$g_i^h = 1 + ih(1 - ih), \;\; h = 1/(N+1), \;\; i = 0, 1, \dots, N+1. \tag{10.5.108}$$

A $\mathbf{g}_h = \left(g_i^h\right)_{i=0}^{N+1} \in \mathbb{R}^{N+2}$ vektorra teljesülnek az alábbi relációk.

1. Egyszerű számolással közvetlenül ellenőrizhető, hogy $\overline{\mathbf{A}}_h^{(4)} \mathbf{g}_h = [1, 2, 2, \dots, 2, 1]^T \in \mathbb{R}^{N+2}$.

2. A számtani-mértani közepek közötti egyenlőtlenségből következően $ih(1 - ih) \le 1/4$, azaz $g_i^h \le 5/4$.

Tehát a $g_1 = 1$ és a $g_2 = 5/4$ megválasztással a (10.5.100) és a (10.5.101) egyenlőtlenségek érvényesek. Ezért a globális hibára fennáll az

$$\|e_h\|_\infty \le \frac{5}{4} \|l_h(u)\|_\infty \tag{10.5.109}$$

egyenlőtlenség, ami a második derivált approximációjának hibabecslését figyelembevéve az

$$\|e_h\|_\infty \le \tilde{C} h^2, \quad \tilde{C} = \frac{5}{48} M_4 \tag{10.5.110}$$

másodrendű konvergenciát bizonyító becslést eredményezi.

10.5.6. A diszkrét maximumelv és következményei

Ebben a szakaszban megismerkedünk a diszkrét maximumelvvel és megmutatjuk alkalmazhatóságát a stabilitás (és azon keresztül a konvergencia) bizonyítására.[15]

[15] Ez a rész a 10.5.5. szakasz ismerete nélkül is olvasható.

Tekintsük az

$$\begin{aligned} A_i y_{i-1} + D_i y_i + C_i y_{i+1} = B_i, \quad i = 1, 2, \ldots, N \\ y_0 = B_0, \quad y_{N+1} = B_{N+1} \end{aligned} \tag{10.5.111}$$

alakú lineáris algebrai egyenletrendszert, ahol A_i, D_i, C_i és B_i adott számok. A továbbiakban feltesszük, hogy teljesülnek az

$$A_i, C_i < 0, \quad D_i > 0, \quad A_i + D_i + C_i \geq 0, \ i = 1, 2, \ldots, N \tag{10.5.112}$$

feltételek.

10.5.23. tétel.

Tegyük fel, hogy a (10.5.111) lineáris algebrai egyenletrendszerre a (10.5.112) tulajdonság mellett teljesül a $B_i \leq 0$ $(i = 1, 2, \ldots, N)$ feltétel is. Ekkor minden $i = 1, 2, \ldots, N$ indexre érvényes az

$$y_i \leq \max\{0, B_0, B_{N+1}\} := y^{\gamma}_{\max} \tag{10.5.113}$$

egyenlőtlenség.

Bizonyítás. Jelölje $y^\star := \max_{i=1,2,\ldots,N} y_i$. Indirekt módon tegyük fel, hogy az állítás nem igaz: $y^\star > y^{\gamma}_{\max}$. Ez azt jelenti, hogy létezik olyan $k \in \{1, 2, \ldots, N\}$ index, amelyre $y_k = y^\star$, és az ilyen tulajdonságú indexekre $y_k > y^{\gamma}_{\max}$. Ekkor

$$\begin{aligned} 0 \geq & B_k = A_k y_{k-1} + D_k y_k + C_k y_{k+1} = A_k y_{k-1} + D_k y^\star + C_k y_{k+1} \geq \\ \geq & A_k y^\star + D_k y^\star + C_k y^\star = \underbrace{(A_k + D_k + C_k)}_{\geq 0} \underbrace{y^\star}_{>0} \geq 0. \end{aligned} \tag{10.5.114}$$

Mivel ezen becslés mindkét oldalán nulla áll, ezért mindenütt $\geq$ helyett $=$ jel írható. Így érvényes az

$$A_k y_{k-1} + D_k y^\star + C_k y_{k+1} = A_k y^\star + D_k y^\star + C_k y^\star \tag{10.5.115}$$

egyenlőség. Ezt átrendezve:

$$-A_k(y^\star - y_{k-1}) - C_k(y^\star - y_{k+1}) = 0. \tag{10.5.116}$$

Feltételeink következtében $-A_k, -C_k > 0$ és $y^\star \geq y_{k-1}, y_{k+1}$. Ezért a (10.5.116) egyenlőtlenség csak az

$$y_{k-1} = y_{k+1} = y^\star \tag{10.5.117}$$

egyenlőség teljesülése esetén lehetséges. Tehát ha valamely $i = 1, 2, \ldots, N$ indexre a (10.5.111) megoldása $y^\star$, akkor a megfelelő koordináta jobb illetve bal oldali szomszédjának értéke is $y^\star$. Ezért a (10.5.117) egyenlőség miatt érvényes az

$$y_{k-2} = y_{k+2} = y^\star \tag{10.5.118}$$

egyenlőség is. Folytatva ezt a gondolatmenetet azt kapjuk, hogy

$$y_i = y^\star, \quad i = 0, 1, 2, \ldots, N, N+1. \tag{10.5.119}$$

Mivel $y_0 = B_0$ és $y_{N+1} = B_{N+1}$, ekkor (10.5.119) következtében $B_0 = B_{N+1} = y^\star$. Ez az indirekt feltétel figyelembevételével a

$$B_0 = B_{N+1} > \max\{0, B_0, B_{N+1}\} (= y^{\gamma}_{\min}) \tag{10.5.120}$$

relációt eredményezi, ami nyilvánvalóan nem lehetséges. ■

A diszkrét maximumelvnek számos következménye van.

10.5.24. következmény. Tegyük fel, hogy a (10.5.111) lineáris algebrai egyenletrendszerre a korábbi (10.5.112) tulajdonság mellett teljesül a $B_i \geq 0$ $(i = 1, 2, \ldots, N)$ feltétel is. Ekkor minden $i = 1, 2, \ldots, N$ indexre érvényes az

$$y_i \geq \min\{0, B_0, B_{N+1}\} := y^{\gamma}_{\min} \tag{10.5.121}$$

reláció. ◇

Bizonyítás. Az $y_i \sim -y_i$ transzformáció után az állítás közvetlenül következik a diszkrét maximumelvből. ■

Innen közvetlenül adódik az alábbi tulajdonság.[16]

10.5.25. következmény. Tegyük fel, hogy a (10.5.111) lineáris algebrai egyenletrendszerre a fenti (10.5.112) tulajdonság mellett teljesül a $B_i \geq 0$ $(i = 0, 1, 2, \ldots, N, N+1)$ feltétel is. Ekkor minden $i = 0, 1, 2, \ldots, N, N+1$ indexre érvényes az

$$y_i \geq 0 \tag{10.5.122}$$

egyenlőtlenség. ◇

A diszkrét maximumelv és következménye tehát azt mondja ki, hogy a (10.5.111) alakú és (10.5.112) tulajdonságú lineáris algebrai egyenletrendszer megoldása nempozitív jobb oldal esetén a nemnegatív maximumát (illetve nemnegatív jobb oldal esetén a nempozitív minimumát) felveszi azokban a koordinátákban, ahol a megoldások adottak.

10.5.26. következmény. Tegyük fel, hogy a (10.5.111)-(10.5.113) lineáris algebrai egyenletrendszerben minden $i = 1, 2, \ldots, N$ esetén $B_i = 0$. Ekkor

$$|y_i| \leq \max\{|B_0|, |B_{N+1}|\}, \quad i = 0, 1, 2, \ldots, N, N+1. \tag{10.5.123}$$

◇

10.5.27. következmény. A (10.5.111)-(10.5.113) lineáris algebrai egyenletrendszernek legfeljebb egy megoldása lehetséges.[17] ◇

Bizonyítás. Állításunkhoz azt szükséges belátnunk, hogy a homogén egyenletnek (azaz amikor $B_i = 0$ minden $i = 0, 1, \ldots, N+1$ indexre) csak a triviális megoldás ($y_i = 0$, $i = 0, 1, \ldots, N+1$) az egyetlen megoldása. Ez viszont a 10.5.25. következményből közvetlenül adódik: mivel feltevésünk miatt $B_0 = B_{N+1} = 0$, ezért (10.5.123) alapján minden $i = 0, 1, \ldots, N+1$ indexre $|y_i| = 0$, ami az állításunkat jelenti.

10.5.28. következmény. Tekintsük a (10.5.111) lineáris algebrai egyenletrendszer mellett az

$$\begin{gathered} A_i\widetilde{y_{i-1}} + D_i\widetilde{y_i} + C_i\widetilde{y_{i+1}} = \widetilde{B_i}, \quad i = 1, 2, \ldots, N, \\ \widetilde{y_0} = \widetilde{B_0}, \quad \widetilde{y_{N+1}} = \widetilde{B_{N+1}} \end{gathered} \tag{10.5.124}$$

[16] Ezt a tulajdonságot az irodalomban *nemnegativitás-megmaradási elvnek* szokásos nevezni.
[17] Ez a következmény a (10.5.111) lineáris algebrai egyenletrendszer megoldásának *unicitását* jelenti.

feladatot is. Ha a (10.5.112) tulajdonság mellett feltesszük, hogy

$$|B_i| \le \widetilde{B_i}, \quad i = 0, 1, \dots, N+1, \tag{10.5.125}$$

akkor a megoldásokra érvényes az

$$|y_i| \le \widetilde{y_i}, \quad i = 0, 1, \dots, N+1 \tag{10.5.126}$$

becslés.[18] ◇

Bizonyítás. Vezessük be a

$$v_i = \widetilde{y_i} - y_i, \quad w_i = \widetilde{y_i} + y_i \tag{10.5.127}$$

jelöléseket! Ekkor ezen új vektorok kielégítik az

$$\begin{gathered} A_i v_{i-1} + D_i v_i + C_i v_{i+1} = \widetilde{B_i} - B_i, \quad i = 1, 2, \dots, N \\ v_0 = \widetilde{B_0} - B_0, \quad v_{N+1} = \widetilde{B_{N+1}} - B_{N+1} \end{gathered} \tag{10.5.128}$$

illetve az

$$\begin{gathered} A_i w_{i-1} + D_i w_i + C_i w_{i+1} = \widetilde{B_i} + B_i, \quad i = 1, 2, \dots, N \\ w_0 = \widetilde{B_0} + B_0, \quad w_{N+1} = \widetilde{B_{N+1}} + B_{N+1} \end{gathered} \tag{10.5.129}$$

lineáris algebrai egyenletrendszereket. Mivel mindkét rendszer jobb oldala nemnegatív, ezért a 10.5.25. következmény szerint minden $i = 0, 1, \dots, N+1$ indexre $v_i \ge 0$ és $w_i \ge 0$. Ez viszont az $\widetilde{y_i} \ge y_i$ és az $\widetilde{y_i} \ge -y_i$ relációt, azaz állításunkat eredményezi. ■

A 10.5.28. következmény jól alkalmazható az olyan (10.5.112) tulajdonságú (10.5.111) alakú lineáris algebrai egyenletrendszerek megoldásának becsléséhez, amelyben $B_0 = B_{N+1} = 0$. Tehát tekintsük az

$$\begin{gathered} A_i y_{i-1} + D_i y_i + C_i y_{i+1} = B_i, \quad i = 1, 2, \dots, N \\ y_0 = 0, \quad y_{N+1} = 0 \end{gathered} \tag{10.5.130}$$

feladatot! Vezessük be az

$$E_i = D_i + A_i + C_i, \quad E_{\min} = \min_{i=1,2,\dots,N} E_i, \quad F_{\min} = \min_{i=1,2,\dots,N} (-A_i - C_i), \quad B_{\max} = \max |B_i| \tag{10.5.131}$$

jelöléseket! A (10.5.112) feltétel miatt $E_{\min} \ge 0$ és $F_{\min} > 0$.

[18] Ezt a következményt a (10.5.111) lineáris algebrai egyenletrendszer (jobb oldal szerinti) *monotonitásának* szokásos nevezni.

10.5.29. tétel.
Tegyük fel, hogy a (10.5.130) lineáris algebrai egyenletrendszerre

a. érvényesek a (10.5.112) tulajdonságok,

b. ha van olyan egyenlet, ahol az együtthatók összege nulla, akkor az A_i és C_i együtthatók egyenlők, azaz

$$A_i = C_i,\ i = 1, 2, \dots, N, \quad \text{ha} \quad E_{\min} = 0. \tag{10.5.132}$$

Ekkor az

$$\begin{cases} \widetilde{y_i} := \dfrac{B_{\max}}{E_{\min}},\ i = 0, 1, 2, \dots, N+1, & \text{ha}\ E_{\min} > 0; \\ \widetilde{y_i} := \dfrac{B_{\max}}{F_{\min}} i(N+1-i),\ i = 0, 1, 2, \dots, N+1, & \text{ha}\ E_{\min} = 0 \end{cases} \tag{10.5.133}$$

rácsfüggvény majorálja a (10.5.112) és (10.5.132) tulajdonságú (10.5.130) lineáris algebrai egyenletrendszer y_i megoldásának abszolút értékét, azaz

$$|y_i| \le \widetilde{y_i}, \quad i = 0, 1, \dots, N+1. \tag{10.5.134}$$

Bizonyítás. Először tegyük fel, hogy $E_{\min} > 0$. Ekkor behelyettesítve a (10.5.133) konstans $B_{\max}/E_{\min}$ értékű rácsfüggvényt a (10.5.130) egyenlet bal oldalába az

$$A_i \frac{B_{\max}}{E_{\min}} + D_i \frac{B_{\max}}{E_{\min}} + C_i \frac{B_{\max}}{E_{\min}} = (A_i + D_i + C_i) \frac{B_{\max}}{E_{\min}} = E_i \frac{B_{\max}}{E_{\min}} \ge B_{\max} \ge |B_i| \tag{10.5.135}$$

relációt nyerjük minden $i = 1, 2, \dots, N$ esetén. Mivel nyilvánvalóan $B_{\max}/E_{\min} \ge 0 = B_0 = B_{N+1}$, ezért a 10.5.28. következmény alapján erre az esetre a (10.5.134) állításunkat beláttuk.

Térjünk át az $E_{\min} = 0$ esetre! Helyettesítsük be most is a (10.5.133) szerinti rácsfüggvényt a (10.5.130) egyenlet bal oldalába. Könnyen ellenőrizhető, hogy ebben az esetben

$$\begin{aligned} &A_i \widetilde{y_{i-1}} + D_i \widetilde{y_i} + C_i \widetilde{y_{i+1}} = \\ &= \frac{B_{\max}}{F_{\min}} \left[A_i (i-1)(N+1-i+1) + D_i i (N+1-i) + C_i (i+1)(N+1-i-1) \right] = \\ &= \frac{B_{\max}}{F_{\min}} \left[i(N+1-i)(A_i + D_i + C_i) + (C_i - A_i)(N+1-i) + i(A_i - C_i) - (A_i + C_i) \right]. \end{aligned} \tag{10.5.136}$$

Mivel ebben az esetben $A_i = C_i$, ezért a (10.5.136) alapján az

$$\begin{aligned} &A_i \widetilde{y_{i-1}} + D_i \widetilde{y_i} + C_i \widetilde{y_{i+1}} = \frac{B_{\max}}{F_{\min}} \left[i(N+1-i) E_i - (A_i + C_i) \right] \ge \\ &\ge \frac{B_{\max}}{F_{\min}} (-(A_i + C_i)) \ge B_{\max} \ge |B_i| \end{aligned} \tag{10.5.137}$$

relációt nyerjük minden $i = 1, 2, \dots, N$ esetén. Mivel nyilvánvalóan $\widetilde{y_0} = \widetilde{y_{N+1}} = 0 = B_0 = B_{N+1}$, ezért a 10.5.28. következmény alapján erre az esetre is beláttuk a (10.5.134) állításunkat. ■

10.5.30. következmény. A számtani-mértani közepek közötti egyenlőtlenség alapján

$$\sqrt{i(N+1-i)} \le \frac{i + (N+1-i)}{2} = \frac{N+1}{2}.$$

Ezért

$$\frac{B_{\max}}{F_{\min}} i(N+1-i) \leq \frac{(N+1)^2}{4} \frac{B_{\max}}{F_{\min}}.$$

Így a (10.5.130) lineáris algebrai egyenletrendszer megoldására érvényes az

$$|y_i| \leq \begin{cases} \dfrac{B_{\max}}{E_{\min}}, \ i = 0,1,2,\ldots,N+1, & \text{ha } E_{\min} > 0; \\ \dfrac{B_{\max}}{F_{\min}} \dfrac{(N+1)^2}{4}, \ i = 0,1,2,\ldots,N+1, & \text{ha } E_{\min} = 0 \end{cases} \tag{10.5.138}$$

becslés. ◇

10.5.31. következmény. Írjuk fel a (10.5.111) lineáris algebrai egyenletrendszert a szokásos

$$\mathbf{A}_h \mathbf{y}_h = \mathbf{b}_h \tag{10.5.139}$$

mátrix-vektor alakban. A 10.5.30. következnény alapján ekkor érvényes a

$$\|\mathbf{y}_h\|_\infty \leq \begin{cases} \dfrac{1}{E_{\min}} \|\mathbf{b}_h\|_\infty, & \text{ha } E_{\min} > 0; \\ \dfrac{(N+1)^2}{4F_{\min}} \|\mathbf{b}_h\|_\infty, & \text{ha } E_{\min} = 0. \end{cases} \tag{10.5.140}$$

becslés. Mivel $\mathbf{y}_h = \mathbf{A}_h^{-1} \mathbf{b}_h$, ezért tehát

$$\|\mathbf{A}_h^{-1}\|_\infty \leq \begin{cases} \dfrac{1}{E_{\min}}, & \text{ha } E_{\min} > 0; \\ \dfrac{(N+1)^2}{4F_{\min}}, & \text{ha } E_{\min} = 0. \end{cases} \tag{10.5.141}$$

◇

A 10.5.16. tétel és a 10.5.31. következmény alapján kimondható az alábbi absztrakt állítást.

10.5.32. tétel.

Tegyük fel, hogy (10.5.88) egy olyan diszkretizációja a (10.5.74) operátoregyenletnek, amelyre teljesülnek az alábbi tulajdonságok.

1. Az L_h operátor p-ed rendben konzisztens az L operátorral.
2. Az L_h operátor leírható a (10.5.111) bal oldalán szereplő tridiagonális mátrix segítségével.
3. A mátrix együtthatóira teljesülnek a (10.5.112) feltételek.
4. A (10.5.131) jelölések mellett tegyük fel, hogy $E_{\min} = 0$ esetén (10.5.132) teljesül, és létezik olyan C N-től független pozitív állandó, amelyre

$$\frac{(N+1)^2}{4F_{\min}} \leq C \tag{10.5.142}$$

minden N esetén.

Ekkor az L_h numerikus módszerrel kapott közelítő megoldás p-ed rendben konvergens.

Befejezésül egy példán mutatjuk meg a 10.5.32. tétel alkalmazását.

10.5.33. példa. Mutassuk meg, hogy a (10.5.27) paraméterek megválasztása esetén a (10.5.21) séma másodrendben konvergál a (10.5.20) peremérték-feladat megoldásához!

A 10.5.32. tételt felhasználva elegendő megmutatnunk, hogy a tétel feltételei teljesülnek.

1. A másodrendű konzisztenciát a szakasz elején már megmutattuk. (Lásd pl. a (10.5.32)).
2. Ez a tulajdonság a feladatból nyilvánvaló.
3. Az együtthatók (10.5.27) alakjából illetve a 10.5.2. tételből láthatóan megfelelően kis h esetén ez a tulajdonság is teljesül.
4. Mivel $E_{\min} = \min q(t)$, ezért $q(t) \geq q_{\min} > 0$ esetén $E_{\min} > 0$, és így a (10.5.141) miatt a séma stabil a $C = 1/E_{\min}$ stabilitási állandóval. Amikor $p(t) = q(t) = 0$, vagyis a (10.5.105) feladatot tekintjük, akkor $E_{\min} = 0$. Ebben az esetben (10.5.27) alapján nyilvánvalóan $a_i = c_i$ és $F_i = 2/h^2$. Ezért $F_{\min} = 2/h^2$, és így a $h = 1/(N+1)$ összefüggés következtében

$$\frac{(N+1)^2}{4F_{\min}} = \frac{h^2(N+1)^2}{8} = \frac{1}{8} =: C. \tag{10.5.143}$$

Tehát a tétel feltételei teljesülnek, ezért a séma valóban másodrendben konvergál. ◇

10.5.34. megjegyzés. Az előzőekben megmutattuk, hogy a konvergencia másodrendű, azaz $h \to 0$ esetén a globális hiba $\tilde{C} \cdot h^2$ módon tart nullához. Ugyanakkor a gyakorlatban fontos a $\tilde{C}$ konstansnak (avagy egy éles becslésének) az ismerete. Ezt az $l(u)$ lokális approximációs hiba h^2-es tagjának együtthatójából és a stabilitási állandóból kaphatjuk. Ezért a példánkban $q_{\min} > 0$

esetén ez a konstans

$$\tilde{C} = \left(\frac{M_4}{12} + \frac{p_{\max} M_3}{6}\right) / q_{\min}, \quad M_j = \max_{[a,b]} |u^{(j)}|, \quad p_{\max} = \max_{[a,b]} |p|,$$

és így a (10.5.39) becsléssel megegyező eredményt kapunk. Ha $p = q = 0$, akkor $\tilde{C} = M_4/96$, tehát becslésünk élesebb, mint a (10.5.110) hibabecslés. ⋄

10.6. A peremérték-feladatok numerikus megoldása MATLAB segítségével

A peremérték-feladatok numerikus megoldására a MATLAB a BVP4C rutint javasolja. Ez a rutin a kétpontos peremértékfeladatokat elég általános esetben képes megoldani: egyrészt nem szeparált peremfeltételek is megadhatók (amikor nem csak a két végpontbeli függvényértékek illetve deriváltak adottak külön-külön a végpontokban, hanem azok kombinációi), másrészt a peremfeltételben paraméter is megadható. Módszerként az ún. kollokációs módszert alkalmazza, amelynek eredményeként egy nemlineáris algebrai egyenletrendszert nyerünk. Ennek megoldására a Newton módszert alkalmazzuk. Mivel ehhez a bemenő függvények parciális deriváltjai is szükségesek, ezeket is közelítőleg, nevezetesen véges differenciákkal határozzuk meg. A módszer részletei megtalálhatók a
http://200.13.98.241/~martin/irq/tareas1/bvp_paper.pdf
linken a "Solving Boundary Value Problems for Ordinary Differential Equations in MATLAB with bvp4c (Lawrence F. Shampine, Jacek Kierzenka, Mark W. Reichelt) leírásban. Nagyon hasznos továbbá a
http://www.mathworks.com/matlabcentral/fileexchange/3819
link, ahol a BVP4C rutin használatára található egy oktatóanyag és számos kidolgozott példa. Javasoljuk még a
http://www.mathworks.com/help/techdoc/ref/bvp4c.html
linket, ahol a BVP4C leírása mellett további területeken való alkalmazhatósága is szerepel.

Az anyagunkban szereplő két alapvető módszerünkre, a belövéses módszerre illetve a véges differenciák módszerére a megfelelő MATLAB programok önállóan is elkészíthetők. Ennek leírásával és egy modellfeladaton való alkalmazásával a továbbiakban foglalkozunk.

10.6.1. A modellfeladat: stacionárius hőeloszlás homogén vezetékben

Tegyük fel, hogy egy hosszú és vékony vezeték két, állandó hőmérsékletű fal között helyezkedik el. A vezeték vastagsága a hosszúságához képest elhanyagolható, így a sugárirányú hőmérsékletváltozás (radiális sugárzás) elhanyagolható, és ezért a hőmérséklet csak az x egydimenziós térbeli koordinátától függ. A hőáramlás egyrészt a vezetékben történő hosszirányú hőáramlás, másrészt a vezeték és a vezetéket körbevevő állandó hőmérsékletű gáz közötti konvekció hatására történik. Feladatunk a stacinárius hőeloszlás meghatározása.
Írjuk fel először a mérlegegyenletet egy Δx hosszúságú elemre! Jelölje $q(x)$ az x ponthoz tartozó hőfluxust [mértékegysége: $J/(m^2 s)$], A_c a keresztmetszet területét [m^2]), azaz $A_c = \pi r^2$, ahol r a vezeték sugara. Legyen D_{kon} a konvekciós hővezetési együttható [$J/(Km^2 s)$], ahol K a Kelvin hőfok, A_s az elem felülete [m^2], azaz $A_s = 2\pi r \Delta x$, és T_∞ a körbevevő gáz hőmérséklete [K]. Az x pontbeli ismeretlen hőmérsékletet $T(x)$ jelöli. Ekkor a mérlegegyenlet alapján

$$q(x)A_c = q(x + \Delta x)A_c - D_{kon} A_s (T_\infty - T(x)), \tag{10.6.1}$$

ahol a bal oldalon a bemenő hőmennyiség, a jobb oldalon pedig a kimenő hőmennyiség és a konvekció okozta veszteség szerepel. Leosztva a (10.6.1) egyenlőséget az elemi rész térfogatával ($\pi r^2 \Delta x$), átrendezés után a

$$-\frac{q(x+\Delta x)-q(x)}{\Delta x}+\frac{2D_{kon}}{r}(T_\infty - T(x))=0 \tag{10.6.2}$$

egyenlőséget kapjuk. Áttérve a $\Delta x \to 0$ határértékre, (10.6.2) a

$$-q'(x)+\frac{2D_{kon}}{r}(T_\infty - T(x))=0 \tag{10.6.3}$$

egyenletet eredményezi. Mivel Fourier törvénye alapján a fluxusra $q(x) = -D_{hov}T'(x)$, ahol D_{hov} $[J/(Kms)]$ a hővezetési együttható, ezért a (10.6.3) egyenlet felírható

$$T''(x)+D(T_\infty - T(x))=0 \tag{10.6.4}$$

alakban, ahol

$$D=\frac{2D_{kon}}{rD_{hov}}$$

a hőleadást jellemző állandó $[m^{-2}]$. Legyen a vezeték hosszúsága L, és jelölje

$$T(0)=T_0, \quad T(L)=T_1, \tag{10.6.5}$$

ahol T_0 és T_1 a bal illetve jobb oldali fal hőmérséklete. Ekkor tehát matematikai modellünk a (10.6.4)–(10.6.5) kétpontos peremérték-feladat.

A (10.6.4)–(10.6.5) feladat megoldása analitikusan előállítható. Ugyanis az egyenlet átírható

$$T''(x)-DT(x))=-DT_\infty \tag{10.6.6}$$

alakra, amelynek megoldásához a homogén egyenlet általános megoldása és egy partikuláris megoldás ismerete szükséges. Mivel a $\lambda = \pm\sqrt{D}$ jelölésekkel a homogén feladat általános megoldása $T_{hom}(x) = C_1 e^{\lambda x} + C_2 e^{-\lambda x}$, ahol C_1 és C_2 tetszőleges állandók, a $T_{part}(x) = T_\infty$ pedig egy partikuláris megoldás, ezért a (10.6.6) egyenlet általános megoldása

$$T(x)=C_1 e^{\lambda x}+C_2 e^{-\lambda x}+T_\infty. \tag{10.6.7}$$

A képletben szereplő állandókat a (10.6.5) peremfeltételekből határozhatjuk meg:

$$C_1=\frac{(T_0-T_\infty)e^{-\lambda L}-(T_1-T_\infty)}{e^{-\lambda L}-e^{\lambda L}}, \quad C_2=\frac{-(T_0-T_\infty)e^{\lambda L}+(T_1-T_\infty)}{e^{-\lambda L}-e^{\lambda L}}. \tag{10.6.8}$$

A továbbiakban az alábbi adatokkal számolunk: $L = 10$, $D_{kon} = 1$, $D_{hov} = 200$, $r = 0.2$, $T_\infty = 200$, $T_0 = 300$, $T_1 = 400$. (Adataink a korábban megadott mértékegységekben értendők.) Ekkor $D = 0.05$, és a feladatunk megoldása a

$$T(x)=20.4671e^{\sqrt{0.5}x}+79.5329e^{-\sqrt{0.5}x}+200 \tag{10.6.9}$$

függvény.[19]

[19] Amikor a pontos megoldást beprogramozzuk, nem a (10.6.9) képletben szereplő állandókat adjuk meg közvetlenül, hanem a MATLAB segítségével a (10.6.7) képletből számoljuk ki.

10.6.2. A tesztfeladat numerikus megoldása MATLAB segítségével

Alkalmazzuk a belövéses módszert a (10.6.4)–(10.6.5) feladatra. Első lépésben felírjuk a másodrendű közönséges differenciálegyenletet egy kétismeretlenes, elsőrendű rendszerként:

$$\begin{aligned} T'(x) &= z(x), \\ z'(x) &= -D(T_\infty - T(x)). \end{aligned} \tag{10.6.10}$$

Kezdeti feltételként a

$$T(0) = T_0, \quad z(0) = z_0 \tag{10.6.11}$$

feltételeket adjuk meg, ahol T_0 ismert, z_0 értékét a belövéses módszernek megfelelően pedig úgy választjuk meg, hogy a (10.6.10)–(10.6.11) kezdetiérték-feladat $T(x)$ megoldására a $T(L) = T_1$ feltétel teljesüljön.

A belövéses módszer algoritmusát követve MATLAB programunkat két rutin segítségével adjuk meg. A BVPS rutin adott z_0 érték mellett az explicit Euler-módszerrel kiszámolja a numerikus megoldást. Bemenő paraméterként a $[0, L]$ intervallum felosztásához szükséges osztásrészek száma (Nx) illetve az ismeretlen z komponens-függvény kezdeti értéke ($z0$) szerepel. Kimenő paraméterként a végponti ($x = L$) hőmérséklet (TL), a csomópontokban számolt hőmérsékletek ($Tvect$) és a csomópontok koordinátái ($xvect$) szerepelnek. A rutin a következő:

```
function [TL,Tvect,xvect] = bvpsee(Nx,z0)
%
% Belövéses módszer (shooting method) egy L hosszúságú rúd stacionárius
% hőeloszlásának kiszámítására. Ez a következő kétpontos peremérték-feladat
% megoldását igényli:
%
% d^2T
% --- +D(Tinf - T) = 0,    T(0) = T0,    T(L) = T1
% dx^2
% A kezdetiérték-feladat megoldására az egyszerű explicit Euler-módszert
% alkalmazzuk.
% Nx: a térbeli osztásrészek száma
% z0: a kezdeti meredekség
T0 = 300; % bal oldali végponteli peremfeltétel
T1 = 400; % jobb oldali végponteli peremfeltétel
Tinf = 200; % külső hőmérséklet
D = 0.05;  % állandó
L = 10; % a rúd hossza
xs = 0; % kezdőpont koordinátája
T = T0; % kezdeti feltétel
deltax=L/Nx; Tvect(1) = T; xvect(1) = xs; z=z0;
for i=2:Nx+1
dTdx = z;
dzdx = -c * (Tinf - T);
T = T + dTdx * deltax; % Euler módszer
z = z + dzdx * deltax;  xvect(i) = xs + (i-1)*deltax; Tvect(i) = T;
end;
TL=T;
```

Ezzel a rutinnal tehát egy adott kezdeti z értékkel tudjuk meghatározni a hőmérsékleteloszlást. Nyilvánvalóan ez önmagában nem elegendő, hiszen a kiszámolt TL különbözik $T1$-től. A belövéses módszer algoritmusát követve a BVPS rutinnal különböző kezdeti $z(0)$ értékekre kiszámoljuk a végpontbel TL hőmérsékletet, és az így kapott értékekkel kiszámoljuk a $z(0)$ értékhez tartozó $\varepsilon(z(0)) = TL(z(0)) - T1$ eltéréseket. Mivel olyan $z^\star(0)$ értéket keresünk, amelyre $\varepsilon(z^\star(0)) = 0$, ezért a $(z(0), \varepsilon(z(0)))$ pontokra interpolációs polinomot fektetünk, és a keresett $z^\star(0)$ ennek a zérushelye lesz. Ezt a két lépést a programban a (10.4.10) szerinti inverz interpolációval végezzük el.

A fenti lépéseket a SHOOTING rutin hajtja végre, amelynek bemenő paraméterei a következők:

- Nx: a $[0, L]$ intervallum felosztásához szükséges osztásrészek száma,
- $zstart$: a $z(0)$ kezdeti értéke (ami az ismeretlen T függvény gradiensét jelenti az $x = 0$ pontban, azaz a hőmérséklet gradiense a kezdet időpontban),
- $deltaz$: a különböző $z(0)$ értékek meghatározására szolgáló megváltozás,
- Nz: a kezdeti $zstart$ értékektől jobbra és balra további Nz számú $z(0)$ értékkel számolunk, nevezetesen a $z(0) = zstart \pm k \cdot deltaz$ $(k = 1, 2, \ldots, Nz)$ kezdeti értékekkel is meghatározzuk TL értékét. (Ezzel állítjuk elő az interpolációs alappontokat.)

A rutin opcionálisan kiírja a numerikus megoldást, és a pontos megoldás ismeretében kiírja a maximumnormabeli hibát és kirajzolja a pontos és közelítő megoldásokat.

```
function shooting(Nx,zstart,deltaz,Nz)
%
% Belövéses módszer (shooting method) egy L hosszúságú rúd stacionárius
% hőeloszlásának kiszámítására. Ez a következő kétpontos peremérték-feladat
% megoldását igényli:
%
% d^2T
% --- +D(Tinf - T) = 0,   T(0) = T0,   T(L) = T1
% dx^2
% A kezdetiérték-feladat megoldására a BVPS nevű rutint alkalamzzuk.
% Nx: a térbeli osztásrészek száma
% z0: a kezdeti meredekség
T0 = 300; % bal oldali végponteli peremfeltétel
T1 = 400; % jobb oldali végponteli peremfeltétel
Tinf = 200; % külső hőmérséklet
D = 0.05; % állandó
L = 10; % a rúd hossza
xs = 0; % kezdőpont koordinátája
T = T0; % kezdeti feltétel
deltax=L/Nx; zv(Nz+1)=zstart; z=zv(Nz+1);
[T,Tvect,xvect]=bvpsee(Nx,z);Tvegpont(Nz+1)=T;
for i=1:Nz
zv(i)=zstart-(Nz+1-i)*deltaz; z=zv(i);
[T,Tvect,xvect]=bvpsee(Nx,z);Tvegpont(i)=T;
zv(Nz+1+i)=zstart+i*deltaz; z=zv(Nz+1+i);
```

Δx	1	0.1	0.01	0.001	0.0001
e_h	$64.0278e+000$	$4.6133e-001$	$4.6786e-002$	$4.6852e-003$	$4.6859e-004$
rend		$1.1453e-001$	$1.0142e-001$	$1.0014e-001$	$1.0001e-001$

10.6.1. táblázat: A belövéses módszer explicit Euler-módszeres változatának hibája a maximumnormában a tesztfeladatra, és a konvergencia rendje.

```
[T,Tvect,xvect]=bvpsee(Nx,z);Tvegpont(Nz+1+i)=T;
end
for i=1:2*Nz+1
Tvegpont(i); zv(i);
end
% A gyök megkeresésére az inverz interpolációt alkalmazzuk.
Tint=Tvegpont-Tb; z=interp1(Tint,zv,0);
fprintf('A meredekség: %fn',z)
[Tfinal,Tvect,xvect]=bvpsee(Nx,z);
% A Meg nevű mátrixba összerakjuk a numerikus megoldást és utána
% szükség esetén kiiratjuk
Meg=ones(Nx+1,3);
for i=1:Nx+1
Meg(i,1)=i; Meg(i,2)=xvect(i); Meg(i,3)=Tvect(i);
end
% disp(' csomópont koordin. hőmérséklet')
% disp(Meg)
% plot(xvect,Tvekfinal);
% Ha ismerjük a pontos megoldást, akkor adjuk meg
% egy PONTOSSHOOTING nevű függvényben.
% Ekkor a hiba kiszámitható és opcionálisan kirajzolhatjuk a pontos
% és közelítő megoldásokat.
% [Tpontos,zpontos]=pontosshooting(Nx,xvect);
% hiba=norm(Tvect-Tpontos,inf);
% disp('lépésköz és hiba max. normában:'),deltax, hiba
% i=1:Nx+1;
% plot(i,Tvect,'r', i, Tpontos,'b') xlabel('rúd'),
% ylabel('belövéses módszer (piros),
% pontos megoldás (kék)')
```

Mivel a (10.6.9) képlet segítségével ismerjük a pontos megoldást, ezért a belövéses módszert különböző paraméterek mellett tesztelni tudjuk. A Δx lépésköz függvényében a hibavektor maximumnormáját a 10.6.1. táblázat mutatja. A pontos és numerikus megoldások az egyes rácshálókon a 10.6.1.-10.6.4. ábrákon láthatók.

10.6.1. megjegyzés. Futtatásainkat $Nz = 4$ és $zstart = -12$ értékekkel hajtottuk végre. Ha nagyobb Nz értéket választunk, akkor sem változik a futás eredménye. Ennek oka, hogy a tesztfeladatunk lineáris, azaz a 10.4.2. szakasz értelmében direkt módon is kiszámolható két adat

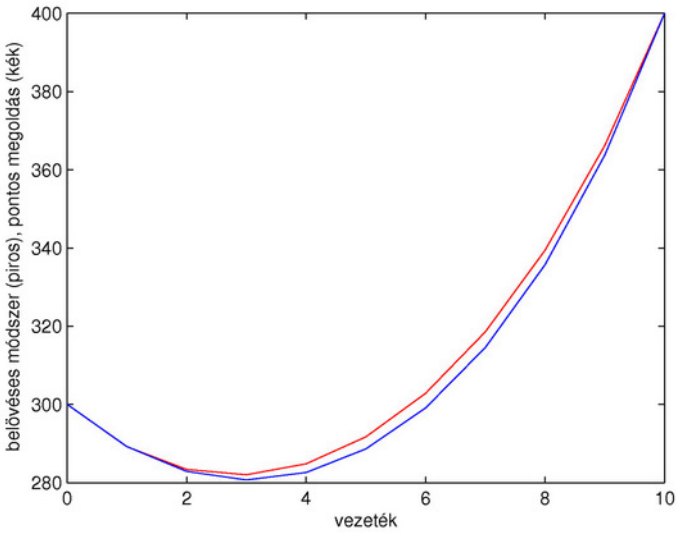
Δx	1	0.1	0.01	0.001	0.0001
e_h	$6.3073e-001$	$6.6929e-003$	$6.7386e-005$	$6.7433e-007$	$6.7446e-009$
rend		$1.0611e-002$	$1.0068e-002$	$1.0007e-002$	$1.0002e-002$

10.6.2. táblázat: A belövéses módszer javított explicit Euler-módszeres változatának hibája a maximumnormában a tesztfeladatra és a konvergencia rendje.

ismeretében a kezdeti meredekség. Ezért lényegében nem szükséges több pont megadása, sőt, valójában az $Nz = 2$ is elegendő. Ehhez elegendő összehasonlítani a 10.6.5. és a 10.6.1. ábrákat, amelyeket ugyanolyan térbeli felosztásra, de különböző Nz értékekre ($Nz = 2$ és $Nz = 10$) kaptunk. ⋄

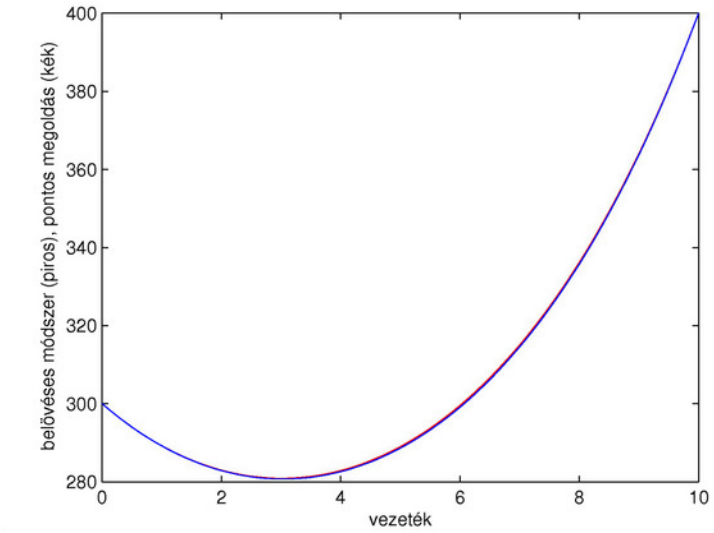
10.6.2. megjegyzés. Ha nagyon távoli *zstart* értéket adunk meg, akkor kevés interpolációs alappont esetén előfordulhat, hogy az inverz interpoláció nem működik. (A nulla érték kivűl esik az alappontokon.) Ilyenkor célszerű előzetesen néhány önálló futtatást végezni a BVPS rutinnal, és közelítőleg meghatározni a keresett z értéket. ⋄

10.6.3. megjegyzés. Ha az explicit Euler-módszer helyett a másodrendű javított Euler-módszerrel számolunk, akkor eredményeink hibái a 10.6.2. táblázatban láthatók. Mint az várható volt, a módszerünk másodrendben konvergens. ⋄

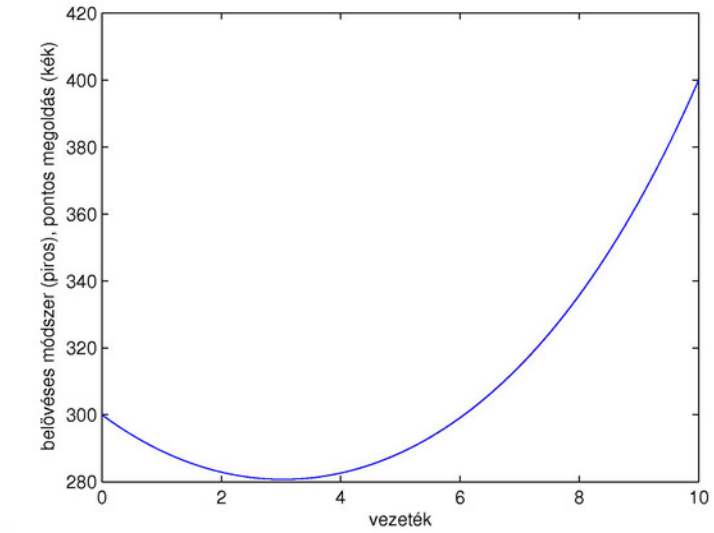


10.6.1. ábra: A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben 10 osztáspontú ekvidisztáns rácshálón.

Térjünk át a (10.6.4)–(10.6.5) feladat véges differenciás megoldására. Nem részletezve az al-

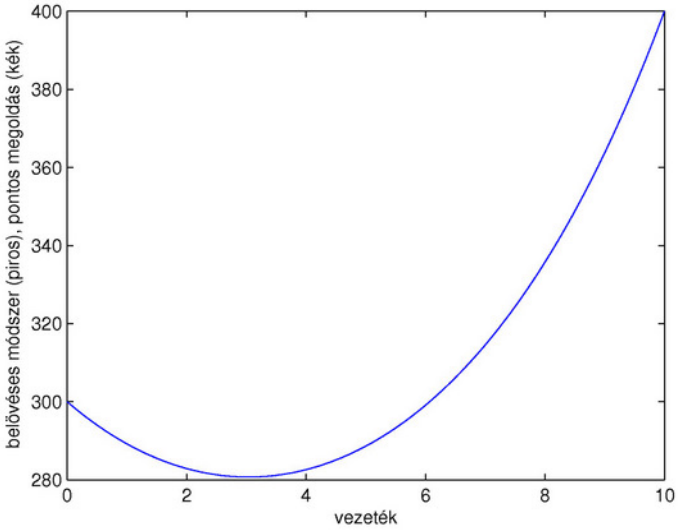


10.6.2. ábra: A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben 100 osztáspontú ekvidisztáns rácshálón.

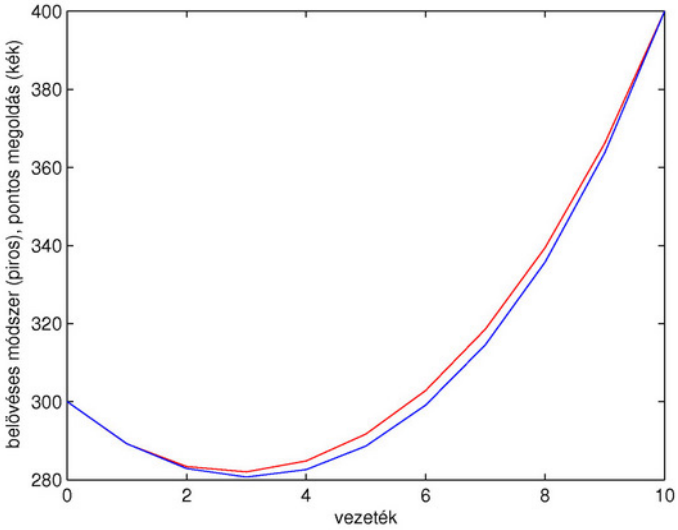


10.6.3. ábra: A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben 1000 osztáspontú ekvidisztáns rácshálón.

goritmust, megadjuk a VDM1 rutint, amely a véges differenciák módszerével megoldja a fenti feladatot. Itt bemenő paraméterként az `L` hosszúságú vezeték osztásrészeinek számát (`Nx`) kell megadni. Kimenő adatként a csomópontok koordinátáit (`xvect`) illetve a csomópontokhoz tar-



10.6.4. ábra: A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben 10000 osztáspontú ekvidisztáns rácshálón.



10.6.5. ábra: A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben 10 osztáspontú ekvidisztáns rácshálón két alappontra ($Nz = 2$) támaszkodó interpolációval.

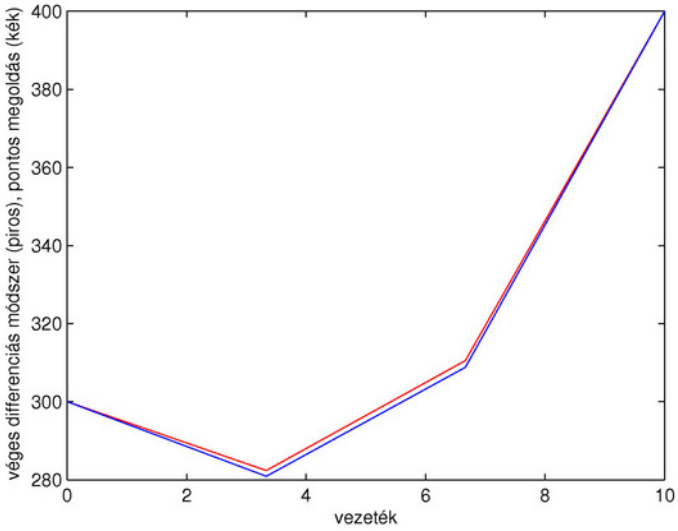
tozó közelítő megoldást tartalmazó vektort (`Tmego`) kapjuk.

osztásrészek száma	hiba max. normában	rend
3	$1.7002e+000$	
6	$4.5604e-001$	$2.6823e-001$
12	$1.1647e-001$	$2.5540e-001$
24	$2.9205e-002$	$2.5074e-001$
48	$7.3158e-003$	$2.5050e-001$
96	$1.8293e-003$	$2.5004e-001$
192	$4.5734e-004$	$2.5001e-001$
384	$1.1434e-004$	$2.5000e-001$
768	$2.8582e-005$	$2.4999e-001$
1536	$7.1573e-006$	$2.5041e-001$

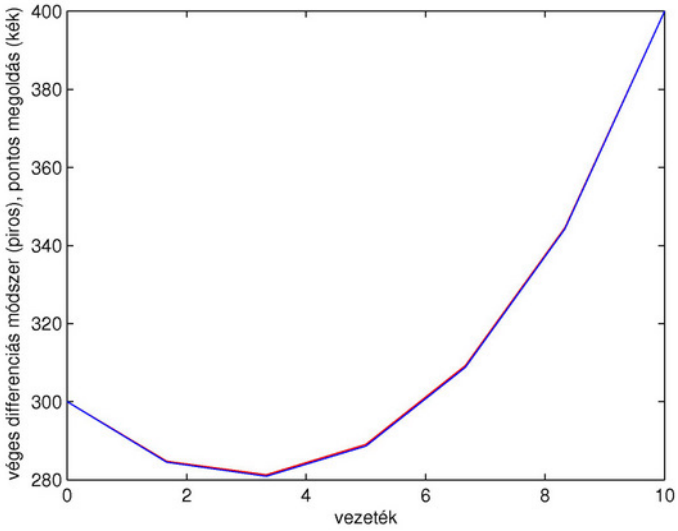
10.6.3. táblázat: A véges differenciák módszerének hibája feleződő lépéshosszal. A harmadik oszlop a rendet mutatja. (Az elméleti rendnem megfelelő hányados 0.25.)

```
function[xvect,Tmego]=vdm1(Nx)
Ta   = 300;      % bal oldali végponteli peremfeltétel
Tb   = 400;      % jobb oldali végponteli peremfeltétel
Tinf = 200;      % külső hőmérséklet
c    = 0.05;     % állandó
L    = 10;       % a vezeték hossza
ndivs  = Nx; nunknowns = ndivs - 1; deltax = L/ndivs;
A = -(2 + deltax^2*c); B = -deltax^2*c*Tinf;
for i=1:Nx+1,    % a diszkretizációs alappontok előállítása
    xvect(i)=(i-1)*deltax;
end;
matrix = zeros(nunknowns); % a lin. egyenletrendszer összeállítása kezdődik
matrix(1,1) = A;  % az első egyenlet összeállítása
matrix(1,2) = 1; rhs(1)= B - Ta;
for i = 2:nunknowns - 1    % a belső pontokhoz tartozó egyenletek
matrix(i,i-1) = 1;
matrix(i,i)   = A;
matrix(i,i+1) = 1;
rhs(i)= B;
end;
matrix(nunknowns, nunknowns-1) = 1;  % az utolsó egyenlet összeállítása
matrix(nunknowns, nunknowns)   = A; rhs(nunknowns)= B - Tb;
T = matrix\rhs';  % a lineáris egyenlet megoldása
Tmego(1)= Ta;     % a teljes megoldásvektor előállítása
Tmego(2:1 + nunknowns) = T(:); Tmego(nunknowns + 2)   = Tb;
```

Eredményeinket a tesztfeladatra a 10.6.3. táblázatban adjuk meg. Az első három esethez (azaz amikor $Nx = 3, 6, 9$) tartozó véges differenciás megoldást a pontos megoldással együtt a 10.6.6.-10.6.8. ábrákon ábrázoltuk.

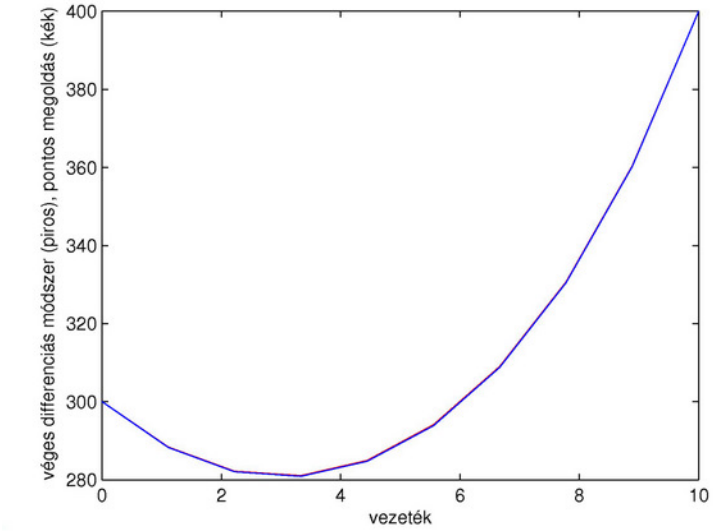


10.6.6. ábra: A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben véges differenciákkal a négy pontból álló ekvidisztáns rácshálón.



10.6.7. ábra: A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben véges differenciákkal a 7 pontból álló ekvidisztáns rácshálón.

Befejezésül hasonlítsuk össze módszereinket ugyanazon diszkrét rácshálón! Legyen $Nx = 8$, azaz 9 osztáspontú, $\Delta x = 1.25$ lépésközű rácshálón hasonlítjuk össze az explicit Euler- és javított explicit Euler-módszeres belövéses eredményeinket, valamint a véges differenciás módszert a pon-



10.6.8. ábra: A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben véges differenciákkal a 10 pontból álló ekvidisztáns rácshálón.

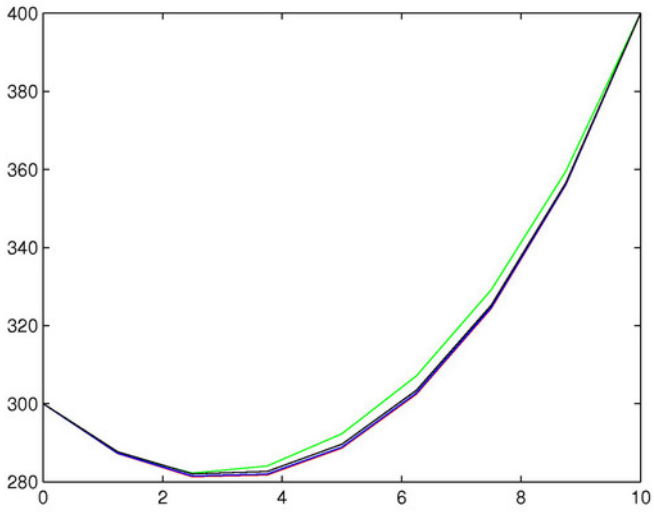
rácspont	explicit Euler	javított EE	véges differenciák	pontos megoldás
0	300.0000	300.0000	300.0000	300.0000
1.2500	287.2008	287.6551	287.3173	287.3173
2.5000	282.2141	282.0058	281.4562	281.4562
3.7500	284.0400	282.6293	281.9589	281.9589
5.0000	292.2889	289.5832	288.8646	288.8646
6.2500	307.1033	303.4096	302.7129	302.7129
7.5000	329.1279	325.1783	324.5856	324.5856
8.7500	359.5199	356.5687	356.1916	356.1916
10.0000	400.0000	400.0000	400.0000	400.0000

10.6.4. táblázat: A különböző módszerek eredményei a tesztfeladatra $Nx = 8$ felosztás esetén. A pontos megoldás piros, az explicit Euler-módszeres belövéses módszeré zöld, a javított explicit Euler-módszeres belövéses módszeré fekete, a véges differencás megoldásé pedig kék.

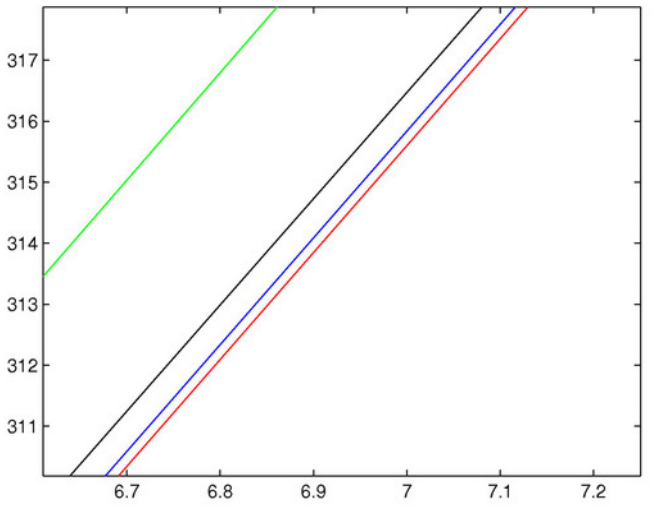
tos megoldással. Eredményeinket a 10.6.4. táblázat tartalmazza. A megoldásokat a 10.6.9. ábrán láthatjuk. Mivel a megoldások közel haladnak egymáshoz, az ábra egy része kinagyítva látható a 10.6.10. ábrán.

10.7. Feladatok

Közönséges differenciálegyenletek peremérték feladatai



10.6.9. ábra: A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben a különböző módszerekkel a $\Delta x = 1.25$ lépésközű ekvidisztáns rácshálón.



10.6.10. ábra: Kinagyítva az előző ábra. A pontos megoldás színe piros, az explicit Euler-módszeres belövéses módszeré zöld, a javított explicit Euler-módszeres belövéses módszeré fekete, a véges differenciás megoldásé pedig kék.

10.7.1. feladat. Mutassuk meg, hogy az

$$u'' = (5u + \sin(3u))\, e^t; \quad u(0) = 0,\ u(1) = 0$$

feladatnak létezik egyértelmű megoldása. (Útmutatás. Vizsgáljuk meg a jobb oldali függvény lipschitzességét!)

10.7.2. feladat. Mutassuk meg, hogy az

$$u'' = \sin(tu) + u^2; \quad u(1) = 3,\ u(4) = 7$$

és az

$$y'' = 16\left(\sin(t(4s+1)y) + y^2\right); \quad y(0) = 3,\ y(1) = 7$$

peremérték-feladatok ekvivalensek egymással. (Útmutatás. Alkalmazzuk a független változó transzformációját!)

10.7.3. feladat. Mutassuk meg, hogy az

$$u'' = 2\exp(t\cos u); \quad u(0) = 0,\ u(1) = 0$$

feladatnak létezik egyértelmű megoldása. (Útmutatás. Vizsgáljuk meg a jobb oldali függvény lipschitzességét!)

10.7.4. feladat. Határozzuk meg azon (α, β) párokat, amelyekre az

$$u'' = u; \quad u(0) = \alpha,\ u(1) = \beta$$

feladatnak létezik egyértelmű megoldása.

10.7.5. feladat. Határozzuk meg

$$u'' - 2' + u = 0; \quad u(0) = \alpha,\ u(1) = \beta$$

feladat megoldását! Van olyan (α, β) pár, amelyre a feladatnak nem létezik megoldása?

10.7.6. feladat. Állítsuk elő az

$$u'' = u^2 \quad u(0) = 2/3,\ u(1) = 3/8$$

feladat megoldását!

10.7.7. feladat. Vizsgáljuk meg a 10.7.6. feladatot az

$$u(0) = 0,\ u(1) = 1$$

peremfeltételekkel!

10.7.8. feladat. Tekintsük az

$$u'' = -4u, \quad u(0) = 1,\ u(\pi/2) = -1$$

peremérték-feladatot! Melyik állítás igaz az alábbiak közül?

- nincs megoldása;
- pontosan két megoldása van;
- pontosan egy megoldása van;
- az elemi függvények körében nincs megoldása;

- egynél több megoldása van.

10.7.9. feladat. Adjuk meg a 10.7.8. feladat kérdéseire a helyes válaszokat, ha a feladat peremfeletételei $u(0) = 1,\ u(\pi/2) = 2$ alakúak!

Belövéses módszer

10.7.10. feladat. Határozzuk meg a belövéses módszer során meghatározandó (10.4.6) szerinti $h(c)$ függvényt az

$$u'' = -u, \quad u(0) = 1,\ u(\pi/2) = 3$$

peremérték-feladatra! Oldjuk meg a $h(c) = 0$ egyenletet!

10.7.11. feladat. Határozzuk meg a $h(c)$ függvényt a

$$u'' = -u'u^{-1}, \quad u(1) = 3,\ u(2) = 5$$

peremérték-feladatra! Oldjuk meg a feladatot a h függvény felhasználásával!

10.7.12. feladat. Oldjuk meg az

$$u'' + y' + 2 + 2(t-2) = 0, \quad u(1) = 0,\ u(2) = 1$$

peremérték-feladatot a belövéses módszer segítségével. Írjunk MATLAB programot a módszer végrehajtására! Alkalmazzuk az EE-módszert a kezdetiérték-feladatok megoldására!

10.7.13. feladat. Oldjuk meg az

$$u'' = e^t + u\cos t - (t+1)u', \quad u(0) = 1, u(1) = 3$$

peremérték-feladatot a belövéses módszer segítségével! Alkalmazzuk az RK2 módszert a kezdetiérték-feladatok megoldására!

10.7.14. feladat. Oldjuk meg a a 10.7.13. peremérték-feladatot a belövéses módszer segítségével! Alkalmazzuk most az RK4 módszert $h = 0.01$ megválasztással a kezdetiérték-feladatok megoldására!

10.7.15. feladat. Írjuk át a BVPSEE.M és SHOOTING.M m-fájlokat arra az esetre, amikor a kezdetiérték-feladatok megoldására ez explicit Euler-módszer helyett a negyedrendű RK4-módszert alkalmazzuk!

10.7.16. feladat. Írjuk át a BVPSEE.M és SHOOTING.M m-fájlokat úgy, hogy az inverz interpoláció helyett felírjuk a h függvény Lagrange-interpolációját és annak a gyökét keressük meg!

Véges differenciák módszere

10.7.17. feladat. Határozzuk meg $h = 0.5$ lépésköz mellett véges differenciák módszerével az

$$u'' + 2u' + 10t = 0, \quad u(0) = 1,\ u(1) = 2$$

peremérték-feladat megoldását a $t = 0.5$ pontban!

10.7.18. feladat. Írjunk fel egy véges differenciák módszerén alapuló diszkretizációt az

$$u'' = -u'u^{-1}, \quad u(1) = 3,\ u(2) = 5$$

peremérték-feladat megoldására!

10.7.19. feladat. Írjuk fel az operátoregyenletes alakot az

$$-u'' + cu = f, \quad x \in (0, l), \quad u'(0) = \alpha, \quad u(l) = \beta$$

feladatra!

10.7.20. feladat. Írjuk fel a 10.7.19. feladat véges differenciás közelítését és annak operátoregyenletes alakját!

10.7.21. feladat. Mutassuk meg a 10.7.20. feladatban meghatározott közelítések konvergenciáját! (Használjuk fel, hogy a megfelelő $\mathbf{L}_h$ mátrix M-mátrix!)

10.7.22. feladat. Mutassuk meg a 10.5.6. tétel második állítását az ottani bizonyítás felhasználásával!

10.7.23. feladat. Jelölje $L^{(2)}_{0,h}, L^{(3)}_{0,h} : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\omega_h)$ azokat az operátorokat, amelyek egy $v_h \in \mathbb{F}(\overline{\omega}_h)$ függvényre a következő módon hatnak:

$$\left(L^{(2)}_{0,h} v_h\right)(t) = -\frac{v_h(t+h) - 2v_h(t) + v_h(t-h)}{h^2} + p_i \frac{v_h(t) - v_h(t-h)}{h} + q_i v_h(t),\ t \in \omega_h, \tag{10.7.1}$$

$$\left(L^{(3)}_{0,h} v_h\right)(t) = -\frac{v_h(t+h) - 2v_h(t) + v_h(t-h)}{h^2} + p_i \frac{v_h(t+h) - v_h(t-h)}{2h} + q_i v_h(t),\ t \in \omega_h. \tag{10.7.2}$$

Legyen

$$L^{(k)}_h v_h = \begin{pmatrix} L^{(k)}_{0,h} v_h \\ B_{1,h} v_h \\ B_{2,h} v_h \end{pmatrix},\ k = 2, 3. \tag{10.7.3}$$

Mutassuk meg, hogy ebben az esetben is $L^{(k)}_h$ felírható (10.5.83) alakú mátrix alakjában, ahol a_i, d_i, c_i értékei a (10.5.26) illetve a (10.5.27) képlet szerintiek.

10.7.24. feladat. Írjuk át a VDM1.M rutint arra az esetre, amikor a $t = 0$ pontban az $u'(0) = 0$ peremfeltétel adott!

10.7.25. feladat. Írjuk át a VDM1.M rutint úgy, hogy előállítsa a

$$u'' + a(t)u' + b(t)u = f(t), \quad u(0) = T_a,\ u(1) = T_b$$

peremérték-feladat megoldását!

10.7.26. feladat. Oldjuk meg a MATLAB program segítségével az $u''(t) + t\cos u(t) = 0,\ u(0) = 0,\ u(1) = 0$ feladatot véges differencia módszerrel!

10.7.27. feladat. Oldjuk meg az $u''(t) = -u,\ u(0) = 3,\ u(\pi/2) = 7$ feladatot véges differencia módszerrel. Készítsünk táblázatot a lépésköz és a hiba kapcsolatáról! Ábrázoljuk ezt a függvényt! (A pontos megoldás: $u(t) = 7\sin t + 3\cos t$.)

10.7.28. feladat. Oldjuk meg az $u''(t) = 2e^t - u,\ u(0) = 2,\ u(1) = e + \cos 1$ feladatot véges differencia módszerrel. Készítsünk táblázatot a lépésköz és a hiba kapcsolatáról! Ábrázoljuk ezt a függvényt! (A pontos megoldás: $u(t) = e^t + \cos t$.)

Ellenőrző kérdések

1. Mit nevezünk közönséges differenciálegyenlet peremérték-feladatának?
2. Mi a szerepe a peremfeltételeknek?
3. Mit nevezünk Lipschitz-féle tulajdonságnak és mi a kapcsolata a peremérték-feladatok megoldhatóságával?
4. Ismertesse a lineáris peremérték-feladat megoldhatóságának feltételét!
5. Mi a belövéses módszer, hogyan kapcsolódik a kezdetiérték-feladatok numerikus megoldásához?
6. Milyen módszereket alkalmazunk a belövéses módszer során a $h(c) = 0$ egyenlet gyökeinek meghatározására? Rendszerek esetén mely módszerek alkalmazhatók közülük?
7. Milyen módszerrel oldhatók meg a lineáris peremérték-feladatok?
8. Ismertesse a véges differenciás megoldási módszer lényegét!
9. Mikor nevezünk egy lineáris peremérték-feladatot megoldó numerikus módszert konvergensnek? Mi a konzisztencia és a stabilitás? Mi a kapcsolat közöttük?
10. Mutassa meg, hogy a lineáris peremérték-feladatok esetén a véges differenciás módszer konzisztens!
11. Mutassa meg, hogy a lineáris peremérték-feladatok esetén a véges differenciák módszere konvergens!
12. Mi a szerepe az M-mátrixoknak peremérték-feladatok numerikus megoldásában?
13. Mit nevezünk diszkrét maximumelvnek? Mi a szerepe peremérték-feladatok numerikus mgoldásában?
14. Milyen MATLAB programokat ismer a peremérték-feladatok megoldására?
15. Milyen alapon működnek a beépített MATLAB programok?

11. A parciális differenciálegyenletek numerikus módszerei

Ebben a fejezetben a parciális differenciálegyenletek feladatainak elméleti összefoglalása után azok numerikus megoldási módszereivel foglalkozunk. Ismertetjük a legtipikusabb véges differenciák módszerét a különböző feladatokra.

11.1. A parciális differenciálegyenletek alapfogalmai

A differenciálegyenletek közös vonása, hogy a függvény és deriváltjai közötti ismert kapcsolatból kell magát a függvényt meghatározni. Amikor a keresett függvény egyváltozós, akkor közönséges differenciálegyenletnek nevezzük a problémát. Az ezzel kapcsolatos kérdéseket, beleértve a numerikus megoldási módszereket, az előző két szakaszban már tárgyaltuk erre az esetre. Amikor az ismeretlen függvény többváltozós, és így az említett kapcsolat az ismeretlen függvény és annak *parciális deriváltjai* közötti kapcsolatot jelenti, *parciális differenciálegyenletről* beszélünk. Néhány tipikus kétváltozós parciális differenciálegyenlet és elnevezésük:

a. Laplace-egyenlet:
$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 0, \tag{11.1.1}$$

b. Poisson-egyenlet:
$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = f(x,y), \tag{11.1.2}$$

c. hővezetési egyenlet:
$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \tag{11.1.3}$$

d. hullámegyenlet:
$$\frac{\partial^2 u(x,t)}{\partial t^2} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \tag{11.1.4}$$

e. advekciós egyenlet:
$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial \left(k(x,t)u(x,t)\right)}{\partial x} = 0, \tag{11.1.5}$$

f. diffúziós egyenlet:
$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial}{\partial x}\left(D(u,x,t)\frac{\partial u(x,t)}{\partial x}\right) = 0, \tag{11.1.6}$$

g. reakció-diffúziós egyenlet:
$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = R(u), \tag{11.1.7}$$

ahol $R(u)$ egy adott függvény,

h. biharmonikus egyenlet:

$$\frac{\partial^4 u(x,y)}{\partial x^4} + 2\frac{\partial^4 u(x,y)}{\partial x^2 \partial y^2} + \frac{\partial^4 u(x,y)}{\partial y^4} = 0. \tag{11.1.8}$$

A fenti példákban x, y a *térbeli változókat*, t az *időbeli változót*, u pedig az *ismeretlen függvényt* jelöli. A parciális differenciálegyenletek elméletében használatos néhány elnevezés. *Rendnek* nevezzük az u függvény legmagasabb előforduló parciális deriválási rendjét. Ezért e. elsőrendű, h. negyedrendű, a többi pedig másodrendű parciális differenciálegyenlet. Megkülönböztetjük azokat az eseteket, amikor az ismeretlen függvény parciális deriváltjainak együtthatói *állandóak* avagy *függvények*. Példáinkban a., b., c., d., g. és h. állandó együtthatós parciális differenciálegyenlet, a többi pedig függvényegyütthatós. Fontos osztályozási szempont a linearitás: ha az egyenletekben az ismeretlen függvény és annak deriváltjai közötti kapcsolat lineáris, akkor *lineáris parciális differenciálegyenletnek*, ellenkező esetben *nemlineáris parciális differenciálegyenletnek* nevezzük a feladatunkat. Példáinkban f. és nemlineáris R függvény esetén g. nemlineáris, a többi pedig lineáris parciális differenciálegyenlet. Végezetül, lineáris feladatok esetén szokásos megkülönböztetni a *homogén* és az *inhomogén* parciális differenciálegyenleteket. Az utóbbi azt jelenti, hogy az egyenletben szerepel az ismeretlen u függvénytől illetve annak parciális deriváltjaitól nem függő tag. Példáinkban b. inhomogén, a többi pedig homogén.

Külön térjünk ki a független változók szerepére! Általában t az időt, x és y pedig a helyet jelöli. Ezek szerepe nem azonos: a folyamatokat általában $t \geq 0$ esetén vizsgáljuk, a térbeli változók viszont akármilyen előjelűek lehetnek. (Emellett a modellekben t mindig növekedő irányban változik, és a jelenségek tipikusan időben nem megfordíthatók, a térbeli változók viszont akármilyen irányban változhatnak.) Az olyan feladatot, amelyben az ismeretlen függvény időben nem változik (azaz u nem függ t-től), *stacionárius feladatnak* nevezzük, ellenkező esetben *időfüggő (instacionárius) feladatról* beszélünk.

Mi a továbbiakban a kétváltozós, másodrendű, lineáris parciális differenciálegyenletekkel foglalkozunk. Tekintsük tehát az $\Omega \subset \mathbb{R}^2$ tartományon az

$$\begin{aligned}(Lu)(x,y) = a(x,y)\frac{\partial^2 u(x,y)}{\partial x^2} + 2b(x,y)\frac{\partial^2 u(x,y)}{\partial x \partial y} + c(x,y)\frac{\partial^2 u(x,y)}{\partial y^2} + \\ d(x,y)\frac{\partial u(x,y)}{\partial x} + e(x,y)\frac{\partial u(x,y)}{\partial y} + g(x,y)u(x,y) = f(x,y)\end{aligned} \tag{11.1.9}$$

egyenletet, ahol az a, b, c, d, e, g *együtthatófüggvények* és az f *forrás* adottak. (Ezen függvények alkalmas megválasztásával, illetve az $y \sim t$ jelöléssel a korábbi lineáris másodrendű példák (a.-f.) mindegyike felírható.)

A (11.1.9) egyenletben szereplő L operátor

$$(L_0 u)(x,y) = a(x,y)\frac{\partial^2 u(x,y)}{\partial x^2} + 2b(x,y)\frac{\partial^2 u(x,y)}{\partial x \partial y} + c(x,y)\frac{\partial^2 u(x,y)}{\partial y^2} \tag{11.1.10}$$

részét az L operátor *főrészének* nevezzük. Az L_0 operátornak megfeleltethetjük a

$$B_{(xy)}(\alpha,\beta) = a(x,y)\alpha^2 + 2b(x,y)\alpha\beta + c(x,y)\beta^2 \tag{11.1.11}$$

kvadratikus alakot, és ez alapján az L operátor alábbi osztályozása lehetséges. Tekintsük valamely rögzített $(x_0, y_0) \in \Omega$ pontban a $B_{(x_0 y_0)}(\alpha,\beta) = const. > 0$ egyenlőséggel definiált másodrendű görbéket az (α,β) síkon. Az $a(x_0,y_0), b(x_0,y_0)$ és $c(x_0,y_0)$ értékektől függően (pontosabban, az $a(x_0,y_0)c(x_0,y_0) - b^2(x_0,y_0)$ előjelétől függően) ez egy ellipszist, parabolát vagy hiperbolát határoz meg. Ez motiválja a következő definíciót.

11.1.1. definíció.

Azt mondjuk, hogy az L operátor (másképpen, a (11.1.9) egyenlet)

- *elliptikus típusú* az $(x,y) \in \Omega$ pontban, ha $a(x,y)c(x,y) - b^2(x,y) > 0$;
- *parabolikus típusú* az $(x,y) \in \Omega$ pontban, ha $a(x,y)c(x,y) - b^2(x,y) = 0$;
- *hiperbolikus típusú* az $(x,y) \in \Omega$ pontban, ha $a(x,y)c(x,y) - b^2(x,y) < 0$.

Azt mondjuk, hogy elliptikus (parabolikus, hiperbolikus) típusú az $\Omega_1 \subset \Omega$ *tartományon*, ha elliptikus (parabolikus, hiperbolikus) típusú a Ω_1 tartomány mindegyik pontjában.

Ha (11.1.9) állandó együtthatós, akkor az Ω tartományon azonos típusú. Például a Laplace- és a Poisson-egyenletek elliptikus, a hővezetési egyenlet parabolikus, a hullámegyenlet pedig hiperbolikus típusú a teljes Ω tartományon. Ugyanakkor, az

$$y\frac{\partial^2 u(x,y)}{\partial x^2} + 2x\frac{\partial^2 u(x,y)}{\partial x \partial y} + y\frac{\partial^2 u(x,y)}{\partial y^2} = 0$$

függvényegyütthatós egyenlet az $\Omega_{ell} = \{(x,y) \in \mathbb{R}^2, |y| > |x|\}$ halmazon elliptikus, az $\Omega_{par} = \{(x,y) \in \mathbb{R}^2, |y| = |x|\}$ halmazon parabolikus, az $\Omega_{hip} = \{(x,y) \in \mathbb{R}^2, |y| < |x|\}$ halmazon pedig hiperbolikus típusú.

Az osztályozás után térjünk át a parciális differenciálegyenletkkel leírt modellek vizsgálatára. Célunk olyan matematikai modellek megadása, amelyek *korrekt kitűzésűek*, azaz rendelkeznek az alábbi tulajdonságokkal:

a. létezik megoldása (egzisztencia). Ez azt jelenti, hogy létezik olyan kellően sima függvény, amely kielégíti az egyenletet és a kiegészítő feltételeket.

b. Ez a megoldása egyértelmű (unicitás).

c. A megoldása folytonosan függ a feladatot meghatározó függvényektől (stabilitás).

Ezek a követelmények természetes módon következnek a matematikai modellezés jellegéből és céljából. Vegyük észre, hogy általános esetben a (11.1.9) alakú parciális differenciálegyenletnek, ha létezik is megoldása, akkor az nem egyértelmű. (Például a Laplace-egyenletnek megoldása az $u(x,y) = ax+by+c$ alakú lineáris függvény tetszőleges a, b és c állandók esetén.) Ezért tehát a parciális differenciálegyenletek önmagukban nem elegendőek a korrekt kitűzés biztosításához. Ehhez további feltételek megadása szükséges. Általában a megoldásfüggvényről a megoldási tartomány határán különböző információkkal rendelkezünk. Ezért a kiegészítő feltételeket ezen információk segítségével szokásos megadni.

Amikor az megoldási tartomány az (x,y) térváltozókból álló $\Omega \subset \mathbb{R}^2$ korlátos halmaz, akkor az Ω tartomány Γ peremén írunk elő *peremfeltételeket*. Amikor tér-idő változó egyaránt szerepel, azaz Ω az (x,t) típusú pontokból áll, és a térváltozóban korlátos a tartomány, akkor a $t = 0$ pontban *kezdeti feltételeket*, a térbeli tartomány határán pedig továbbra is peremfeltételeket adhatunk meg. A peremfeltételek megadásának három típusa van.

- *első (Dirichlet-) típusú peremfeltétel*, ami azt jelenti, hogy a Γ-beli perempontban rögzítjük a megoldásfüggvény értékét;
- *második (Neumann-) típusú peremfeltétel*, ami azt jelenti, hogy a Γ-beli perempontban ismerjük a normál irányú deriváltjának az értékét;

- *harmadik (Robin-) típusú peremfeltétel*, ami azt jelenti, hogy a Γ-beli perempontban előre megadjuk a megoldásfüggvény és annak külső normálvektor irányú deriváltjának valamely lineáris kombinációjának értékét.

A *kezdeti (Cauchy-) feltétel* megadása azt jelenti, hogy $t = 0$ időpontban megadjuk a megoldásfüggvény és/vagy annak t szerinti deriváltjának az értékét.

Célunk, hogy egy adott parciális differenciálegyenletet olyan kiegészítő feltételekkel lássunk el, amelyekkel a feladat korrekt kitűzésű lesz. Megmutatható ([9, 30]), hogy lineáris esetben

- az elliptikus feladatok Γ-n megadott első, második vagy harmadik peremfeltétellel,
- a parabolikus feladatok a $t = 0$-ban megadott $u(x, 0)$ kezdeti feltétellel és a térbeli határon megadott első, második vagy harmadik peremfeltételek egyikével,
- a hiperbolikus feladatok a $t = 0$ pontban megadott $u(x, 0)$ és $\dfrac{\partial u}{\partial t}(x, 0)$ kezdeti feltételekkel, valamint a térbeli határon megadott első, második vagy harmadik peremfeltételek valamelyikével

korrekt kitűzésűek.

11.2. Lineáris, másodrendű, elliptikus parciális differenciálegyenletek

Ebben a részben a lineáris, másodrendű, elliptikus parciális differenciálegyenletekkel és azok *véges differenciák módszerével* történő numerikus megoldásával foglalkozunk. Megmutatjuk, hogy az előző fejezetben tárgyalt kétpontos peremérték-feladatok és azok véges differencia módszeres numerikus megoldási technikája lényegében kiterjeszthető az ebben a szakaszban vizsgált feladatra.

11.2.1. A Laplace-egyenlet analitikus megoldása egységnégyzeten

Tekintsük a $\Omega = (0, 1) \times (0, 1)$ tartományon és annak Γ peremén az alábbi első peremérték-feladatot:

$$\frac{\partial^2 u(x, y)}{\partial x^2} + \frac{\partial^2 u(x, y)}{\partial y^2} = 0, \quad (x, y) \in \Omega, \tag{11.2.1}$$

$$u(x, y) = \mu(x, y), \quad (x, y) \in \Gamma. \tag{11.2.2}$$

Nyilvánvalóan feltehető, hogy a peremfeltételt leíró μ függvény nem azonosan nulla, hiszen ebben az esetben a feladat megoldása az $u(x, y) = 0$ függvény lenne, és ezen eset vizsgálata a számunkra érdektelen.)

Tegyük fel, hogy a peremfeltételt leíró, adott μ függvény csak az $y = 1$ oldal mentén nem nulla, vagyis a (11.2.2) peremfeltétel helyett az

$$\begin{aligned} u(0, y) = 0, \quad u(1, y) = 0, \quad y \in (0, 1), \\ u(x, 0) = 0, \quad u(x, 1) = \mu_4(x), \quad x \in (0, 1) \end{aligned} \tag{11.2.3}$$

peremfeltételekkel oldjuk meg a (11.2.1) egyenletet. Keressük a megoldást

$$u(x, y) = X(x) \cdot Y(y) \tag{11.2.4}$$

ún. *szétválasztható alakban*, ahol X és Y olyan, egyelőre ismeretlen, $C^2[0,1]$-beli függvények, amelyekre $X(x)$ és $Y(y)$ nem az azonosan nulla függvények a $(0,1)$ intervallumon. Behelyettesítve a (11.2.4) alakú u függvényt a (11.2.1) egyenletbe, az

$$X''(x)\cdot Y(y)+Y''(y)\cdot X(x)=0,\quad x,y\in(0,1) \tag{11.2.5}$$

egyenletet nyerjük. Ezért azokban a pontokban, ahol $X(x)\cdot Y(y)\neq 0$, (11.2.5) a

$$-\frac{X''(x)}{X(x)}=\frac{Y''(y)}{Y(y)} \tag{11.2.6}$$

azonosságot jelenti. Mivel a (11.2.6) bal oldala csak az x, a jobb oldala pedig csak az y változótól függ, ezért az egyenlőség csak akkor állhat fenn, ha mindkét oldal állandó: valamely $\lambda\in\mathbb{R}$ szám mellett (ez az ún. szeparációs állandó) minden $x\in(0,1)$ és $y\in(0,1)$ esetén érvényes a

$$-\frac{X''(x)}{X(x)}=\frac{Y''(y)}{Y(y)}=\lambda \tag{11.2.7}$$

egyenlőség. Innen a

$$-X''(x)=\lambda X(x),\quad x\in(0,1) \tag{11.2.8}$$

egyenletet kapjuk. Másrészt, behelyettesítve a (11.2.4) alakot a (11.2.3) peremfeltételek első két egyenletébe, az $X(0)Y(y)=X(1)Y(y)=0$ egyenlőséget kapjuk, amely az $Y(y)\neq 0$ miatt az

$$X(0)=X(1)=0 \tag{11.2.9}$$

feltételt jelenti. Célunk tehát olyan $\lambda\in\mathbb{R}$ szám meghatározása, amely mellett a (11.2.8)-(11.2.9) feladatnak létezik a triviális $X(x)=0$ függvénytől különböző $C^2[0,1]$-beli megoldása. Mivel (11.2.8) egy állandó együtthatós, másodrendű közönséges differenciálegyenlet, ezért általános megoldását az $s^2+\lambda=0$ karakterisztikus egyenletének gyökeivel határozhatjuk meg. Ez λ előjelének függvényében az alábbi esetekhez vezet.

- Tegyük fel, hogy $\lambda<0$. Ekkor a karakterisztikus egyenlet gyökei valósak, és a (11.2.8) egyenlet általános megoldása

  $$X(x)=C_1e^{\sqrt{-\lambda}x}+C_2e^{-\sqrt{-\lambda}x}.$$

  Mivel ekkor az $X(0)=0$ feltétel a $C_1+C_2=0$ egyenlőséget jelenti, ezért $X(x)=C_1\left(e^{\sqrt{-\lambda}x}-e^{-\sqrt{-\lambda}x}\right)$ alakú. A másik, $X(1)=0$ peremfeltételt ide behelyettesítve az $X(1)=C_1\left(e^{\sqrt{-\lambda}}-e^{-\sqrt{-\lambda}}\right)=0$ feltételt kapjuk. Mivel $e^{\sqrt{-\lambda}}>e^{-\sqrt{-\lambda}}$, ezért a fenti egyenlőség csak $C_1=0$ esetén lehetséges, ami az $X(x)=0$ megoldást eredményezi. Mivel ez nem megengedett, ezért ez az eset nem lehetséges.

- Tegyük fel, hogy $\lambda=0$. Ekkor a (11.2.8) egyenlet $X''(x)=0$ alakú, így az általános megoldása $X(x)=C_1x+C_2$. Erre a függvényre, amelynek képe egy egyenes, a (11.2.9) feltétel csak $C_1=C_2=0$ esetén lehetséges. Ez viszont szintén a nem megengedett $X(x)=0$ triviális megoldáshoz vezet.

- Tegyük fel, hogy $\lambda>0$. Ekkor a karakterisztikus egyenlet gyökei $\pm\mathtt{i}\sqrt{\lambda}$. Így a (11.2.8) egyenlet általános megoldása

  $$X(x)=C_1\sin(\sqrt{\lambda}x)+C_2\cos(\sqrt{\lambda}x).$$

Az $X(0) = 0$ feltétel miatt $C_2 = 0$. Tehát az $X(1) = 0$ feltétel a $C_1 \sin\sqrt{\lambda} = 0$ feltételt jelenti. Mivel $C_1 \neq 0$, ezért ez a $\sqrt{\lambda} = k\pi$ $(k = 1, 2, \ldots)$ feltételt adja, azaz a lehetséges λ értékekre a

$$\lambda_k = k^2\pi^2, \quad k = 1, 2, \ldots \tag{11.2.10}$$

értékeket kapjuk. Tehát az

$$X_k(x) = C_k \sin(k\pi x), \quad k = 1, 2, \ldots \tag{11.2.11}$$

függvények tetszőleges C_k állandók mellett megoldásai a (11.2.8)-(11.2.9) feladatnak.

Térjünk át az $Y(y)$ függvény meghatározására! A (11.2.7) egyenlőség felhasználásával az

$$Y''(y) = \lambda Y(y), \quad y \in (0, 1) \tag{11.2.12}$$

egyenletet kapjuk, ahol (11.2.10) alapján $\lambda = \lambda_k = k^2\pi^2$. Másrészt, a (11.2.3) összefüggésben szereplő $u(x, 0) = 0$ peremfeltétel következtében

$$Y(0) = 0. \tag{11.2.13}$$

Tehát olyan $Y_k(y)$ $(k = 1, 2, \ldots)$ nem nulla függvényeket keresünk, amelyekre

$$Y_k''(y) = k^2\pi^2 Y(y), \quad y \in (0, 1), \quad Y_k(0) = 0. \tag{11.2.14}$$

Az egyenlet általános megoldása $Y_k(y) = C_1 e^{k\pi y} + C_2 e^{-k\pi y}$, és így $Y(0) = C_1 + C_2 = 0$. Tehát $C_2 = -C_1$, vagyis a (11.2.14) tulajdonságú függvények felírhatók

$$Y_k(y) = C_k^1 \left(e^{k\pi y} - e^{-k\pi y}\right) = 2C_k^1 \left(\frac{e^{k\pi y} - e^{-k\pi y}}{2}\right) = \tilde{C}_k^1 \sinh(k\pi y) \tag{11.2.15}$$

alakban[1], ahol C_k^1 illetve $\tilde{C}_k^1$ tetszőleges állandók.

Összesítve a keresett megoldás (11.2.4) alakját a (11.2.11) és a (11.2.15) képletekkel, azt kapjuk, hogy minden $k = 1, 2, \ldots$ esetén bármely tetszőleges C_k állandó mellett az

$$u_k(x, y) = X_k(x)Y_k(y) = C_k \sin(k\pi x) \sinh(k\pi y) \tag{11.2.16}$$

függvények olyan függvények, amelyek megoldásai a (11.2.1) egyenletnek, és kielégítik a (11.2.3) első három (homogén) peremfeltételét. Így az

$$u(x, y) = \sum_{k=1}^{\infty} u_k(x, y) \tag{11.2.17}$$

függvény is rendelkezik ezekkel a tulajdonságokkal. Válasszuk meg a tetszőleges C_k állandókat úgy, hogy a negyedik, $y = 1$ mentén adott inhomogén peremfeltétel is teljesüljön erre az u függvényre! Nyilvánvalóan (11.2.17) és (11.2.16) alapján

$$u(x, 1) = \sum_{k=1}^{\infty} C_k \sin(k\pi x) \sinh(k\pi). \tag{11.2.18}$$

Másrészt a $\mu_4(x)$ függvény Fourier-sora

$$\mu_4(x) = \sum_{k=1}^{\infty} \mu_4^k \sin(k\pi x) \tag{11.2.19}$$

[1] A sinh a jól ismert szinusz hiperbolikus függvény, amelyet gyakran az sh szimbolummal is jelölnek.

alakú, ahol

$$\mu_4^k = 2\int_0^1 \mu_4(s)\sin(k\pi s)ds. \tag{11.2.20}$$

A (11.2.18) és a (11.2.19) képletek összevetéséből

$$C_k = \frac{\mu_4^k}{\sinh(k\pi)}. \tag{11.2.21}$$

Összegezve: az

$$u(x,y) = \sum_{k=1}^{\infty} \frac{\mu_4^k}{\sinh(k\pi)} \sin(k\pi x)\sinh(k\pi y) \tag{11.2.22}$$

függvénysorral definiált $u(x,y)$ függvény a függvénysor egyenletes konvergenciája esetén megoldása lesz a (11.2.1)-(11.2.3) feladatnak.

11.2.1. megjegyzés. Ezek után az $\Omega = (0,1)\times(0,1)$ egységnégyzeten illetve annak Γ peremén kitűzött (11.2.1)-(11.2.2) első peremérték-feladat megoldása *tetszőleges* $\mu(x,y)$ *peremfeltétel esetén* előállítható a következő módon. Vezessük be a $\mu(0,y)=\mu_1(y)$, $\mu(1,y)=\mu_2(y)$, $\mu(x,0)=\mu_3(x)$, $\mu(x,1)=\mu_4(x)$ új függvényeket, és tekintsük a következő négy peremfeltétel-rendszert:

- 1. eset:
$$\begin{aligned} u(0,y)=\mu_1(y), \quad u(1,y)=0, \quad y\in(0,1),\\ u(x,0)=0, \quad u(x,1)=0, \quad x\in(0,1). \end{aligned} \tag{11.2.23}$$
- 2. eset:
$$\begin{aligned} u(0,y)=0, \quad u(1,y)=\mu_2(y), \quad y\in(0,1),\\ u(x,0)=0, \quad u(x,1)=0, \quad x\in(0,1). \end{aligned} \tag{11.2.24}$$
- 3. eset:
$$\begin{aligned} u(0,y)=0, \quad u(1,y)=0, \quad y\in(0,1),\\ u(x,0)=\mu_3(x), \quad u(x,1)=0, \quad x\in(0,1). \end{aligned} \tag{11.2.25}$$
- 4. eset:
$$\begin{aligned} u(0,y)=0, \quad u(1,y)=0, \quad y\in(0,1),\\ u(x,0)=0, \quad u(x,1)=\mu_4(x), \quad x\in(0,1). \end{aligned} \tag{11.2.26}$$

Jelölje rendre $u_1(x,y)$, $u_2(x,y)$, $u_3(x,y)$ és $u_4(x,y)$ a (11.2.1) egyenlet megoldását a (11.2.23)-(11.2.26) peremfeltételekkel. (Az előzőekben az $u_4(x,y)$ függvényt állítottuk elő. Ennek analógiájaként hasonlóan meghatározhatjuk a többi függvényt is.) Ezután az $u(x,y) = u_1(x,y) + u_2(x,y) + u_3(x,y) + u_4(x,y)$ függvény lesz a megoldása az eredeti (11.2.1)-(11.2.2) feladatnak.[2] További részletek és más típusú feladatok ilyen jellegű megoldása megtalálhatók pl. Stephenson könyvében [32]. ⋄

11.2.2. megjegyzés. Felmerülhet a kérdés: nem létezik-e a (11.2.1)-(11.2.2) feladatnak a (11.2.22) függvénytől eltérő más megoldása is? A válasz nemleges. Ugyanis ha egy $w(x,y)$ függvény az Ω

[2] Vegyük észre, hogy az u összegfüggvényben szereplő u_1, u_2, u_3, u_4 függvényekre megadott peremfeltételek nem a teljes Γ peremen adottak: a négy sarokpontban nem adjuk meg az értékeket. Ezért a'priori csak azt tudjuk, hogy ez az u megoldás csak a sarokpontokon kívül egyenlő μ-vel a Γ-n. Ugyanakkor a Fourier-sor konvergenciájának tulajdonsága miatt a megfelelő függvények határértékét veszi fel a megoldás, azaz μ folytonossága miatt ezekben a sarokpontokban is μ értékeit veszi fel.

tartományon kielégíti a (11.2.1) egyenletet, és folytonos az $\overline{\Omega}$ halmazon, akkor teljesül rá az ún. *elliptikus maximum-minimum elv*: a függvény a legnagyobb és legkisebb értékét felveszi a Γ peremen. Ebből közvetlenül könnyen megmutatható, hogy a feladatnak csak egyetlen megoldása lehet. Emellett, a feladat stabil kitűzése is megmutatható: a megoldás a bemenő függvényektől folytonosan függ a maximumnormában [9, 30]. Tehát a (11.3.1)-(11.3.2) feladat korrekt kitűzésű. ⋄

Mint látható, a (11.2.1)-(11.2.2) feladat formálisan ugyan megoldható, de a megoldást valójában négy végtelen függvénysor összegének alakjában tudjuk csak felírni. Ez azt jelenti, hogy a gyakorlatban már az ilyen egyszerű feladat analitikus megoldása sem kivitelezhető. Tehát numerikus megoldás alkalmazása szükséges. A továbbiakban a *véges differenciák módszerével* fogjuk a közelítő megoldást előállítani, és megvizsgáljuk, hogy az alkalmasan meghatározott közelítő megoldások közel lesznek-e a pontos megoldáshoz.

11.2.2. Elliptikus egyenletek közelítő megoldása véges differenciák módszerével

Ebben a részben a Laplace-egyenletnél általánosabb alakú elliptikus típusú parciális differenciálegyenlet peremérték-feladatának véges differenciás megoldásával foglalkozunk. Tekintsük a

$$-\left(\frac{\partial^2 u(x,y)}{\partial x^2}+\frac{\partial^2 u(x,y)}{\partial y^2}\right)+c(x,y)u(x,y)=f(x,y),\quad (x,y)\in\Omega, \tag{11.2.27}$$

$$u(x,y)=\mu(x,y),\quad (x,y)\in\Gamma \tag{11.2.28}$$

másodrendű, elliptikus típusú parciális differenciálegyenletet az első (Dirichlet-féle) peremfeltétellel, ahol c, f és μ adott függvények. Jelölje $L: C^2(\overline{\Omega})\to C(\Omega)\cap C(\Gamma)$ a következő operátort:

$$Lw(x,y)=\begin{cases}\left[-\left(\dfrac{\partial^2 w}{\partial x^2}+\dfrac{\partial^2 w}{\partial y^2}\right)+cw\right](x,y), & \text{ha } (x,y)\in\Omega;\\ w(x,y), & \text{ha } (x,y)\in\Gamma,\end{cases} \tag{11.2.29}$$

továbbá $\tilde{f}:\overline{\Omega}\to\mathbb{R}$ a következő függvényt:

$$\tilde{f}(x,y)=\begin{cases}f(x,y), & \text{ha } (x,y)\in\Omega;\\ \mu(x,y), & \text{ha } (x,y)\in\Gamma.\end{cases} \tag{11.2.30}$$

Ekkor feladatunk az

$$Lu=\tilde{f} \tag{11.2.31}$$

operátoregyenlet megoldása, ahol $u\in C^2(\overline{\Omega})$ az ismeretlen függvény.

A továbbiakban feltesszük a következőket.

- $c\in C(\Omega)$ és $\tilde{f}\in C(\Omega)\cap C(\Gamma)$,
- $c\geq 0$,
- $\Omega=(0,l)\times(0,l)$.

Mivel a (11.2.31) feladat analitikus megoldását általános esetben nem tudjuk előállítani, ezért numerikus eljárást alkalmazunk. Ennek lényege a következő.

1. Definiálunk az $\overline{\Omega}$ halmazon rácsháló́kat az alábbi módon:

$$\omega_h = \{(x_i, y_j), \quad x_i = ih, \ y_j = jh, \ i,j = 1,2,\dots,N-1, \ h = l/N\}$$

$$\overline{\omega}_h = \{(x_i, y_j), \quad x_i = ih, \ y_j = jh, \ i,j = 0,1,\dots,N, \ h = l/N\}.$$

   Jelölje $\gamma_h = \overline{\omega}_h \backslash \omega_h \subset \Gamma$ az $\overline{\omega}_h$ rácsháló Γ peremre eső pontjait.

2. Jelölje $\mathbb{F}(\overline{\omega}_h)$ és $\mathbb{F}(\omega_h)$ az $\overline{\omega}_h$ és az ω_h rácsokon értelmezett, $\mathbb{R}$-be képező függvények vektorterét.

3. Célunk olyan $y_h \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvény meghatározása, amely $\overline{\omega}_h$ pontjaiban közel van a (11.2.31) feladat u megoldásához, és a rácsháló finomításával (azaz $h \to 0$ esetén) az eltérésük nullához tart[3].

Adjunk meg olyan $L_h : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\overline{\omega}_h)$ lineáris operátort és $b_h \in \mathbb{F}(\overline{\omega}_h)$ elemet, amelyekre az

$$L_h y_h = b_h \tag{11.2.32}$$

operátoregyenlet $y_h \in \mathbb{F}(\overline{\omega}_h)$ megoldása rendelkezik a fentiekben leírt tulajdonsággal.

Az L_h operátor megválasztásánál ötletként a véges differenciás approximáció szolgál. Egy tetszőleges $w_h \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvény esetén jelölje $w_h(x_i, y_j) = w_{i,j}$, valamint alkalmazzuk a $c(x_i, y_j) = c_{i,j}$ egyszerűsítő jelölést. Definiáljuk az L_h operátort a következő módon. Rendelje hozzá a w_h rácsfüggvényhez azt az $L_h w_h$-val jelölt $\mathbb{F}(\overline{\omega}_h)$-beli rácsfüggvényt, amely az $(x_i, y_j) \in \overline{\omega}_h$ rácspontokban az alábbi értékeket veszi fel:

$$\begin{cases} -\dfrac{w_{i+1,j} - 2w_{i,j} + w_{i-1,j}}{h^2} - \dfrac{w_{i,j+1} - 2w_{i,j} + w_{i,j-1}}{h^2} + c_{i,j} w_{i,j}, & \text{ha } (x_i, y_j) \in \omega_h; \\ w_{i,j}, & \text{ha } (x_i, y_j) \in \gamma_h. \end{cases} \tag{11.2.33}$$

Definiáljuk a $b_h \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvényt a következő módon:

$$b_h(x_i, y_j) = \begin{cases} f(x_i, y_j), & \text{ha } (x_i, y_j) \in \omega_h; \\ \mu(x_i, y_j), & \text{ha } (x_i, y_j) \in \gamma_h. \end{cases} \tag{11.2.34}$$

Ekkor b_h értéke $\overline{\omega}_h$ mindegyik rácspontjában meghatározható, és a (11.2.32) egyenlet azt jelenti, hogy keressük azon $y_h \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvényt, amelyet a (11.2.33) szerinti L_h operátor ebbe a b_h vektorba képez le.

11.2.3. megjegyzés. Mint az ismeretes, egy függvényt akkor tekintünk ismertnek, ha tudjuk, hogy az értelmezési tartományának pontjaiban milyen értékeket vesz fel. Ezért tehát y_h meghatározásához az $y_h(x_i, y_j)$ értékek ismerete szükséges. A (11.2.32) egyenlet az y_h függvény rácspontbeli értékeire nézve egy lineáris algebrai egyenletrendszert jelent, ahol az ismeretlenek száma egyenlő az egyenletek számával, és mindkettő $(N+1)^2$. Ezért a (11.2.32) feladat egy $(N+1)^2$ ismeretlenes lineáris algebrai egyenletrendszert jelent, amely felírható

$$\mathbf{L}_h \mathbf{y}_h = \mathbf{b}_h \tag{11.2.35}$$

[3] Ezt a fogalmat a későbbiekben pontosítjuk, hiszen a pontos megoldás a teljes $\overline{\Omega}$ halmazon, míg a numerikus megoldás annak csak bizonyos pontjaiban (az $\overline{\omega}_h$ rácsháló pontjaiban) van értelmezve. Ezért a két függvény különböző tereken van értelmezve, és így "eltérésük" nem értelmezhető a szokásos ("különbségük távolsága") módon.

alakban, ahol $\mathbf{y}_h \in \mathbb{R}^{(N+1)^2}$ az y_h rácsfüggvény rácspontbeli értékeiből álló ismeretlen vektor, $\mathbf{b}_h \in \mathbb{R}^{(N+1)^2}$ a b_h rácsfüggvény rácspontbeli értékeiből álló ismert vektor, és $\mathbf{L}_h \in \mathbb{R}^{(N+1)^2 \times (N+1)^2}$ adott mátrix, amelynek alakját később adjuk meg. $\diamond$

A továbbiakban megmutatjuk, hogy a finomodó rácshálók sorozatán a (11.2.32) feladatok megoldásával előállított $\mathbf{y}_h$ a h paraméter megfelelően kis megválasztása mellett jól közelíti a (11.2.27) feladat u megoldását. Ez a következőt jelenti. Minden $(x^\star, y^\star) \in \overline{\Omega}$ ponthoz tudunk olyan $(x_h, y_h) \in \overline{\omega}_h$ rácspontsorozatot definiálni, amelyre $h \to 0$ esetén $(x_h, y_h) \to (x^\star, y^\star)$. Ekkor a numerikus módszer konvergenciája a

$$\lim_{(x_h,y_h)\to(x^\star,y^\star)} (y_h(x_h, y_h) - u(x^\star, y^\star)) = 0 \tag{11.2.36}$$

relációt jelenti.

11.2.3. Általános kitűzés és az alaptétel

Ebben a részben a 11.2.2. részben megfogalmazott feladatokat általános formában leírjuk, majd az alaptételben megadunk egy olyan feltételt, amely mellett a (11.2.36) tulajdonság biztosítható.

Jelölje $P_h : C(\overline{\Omega}) \to \mathbb{F}(\overline{\omega}_h)$ a

$$(P_h v)(x_i, y_j) = v(x_i, y_j), \quad (x_i, y_j) \in \overline{\omega}_h \tag{11.2.37}$$

leképezést, azaz a $P_h v \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvény a v függvény értékeit veszi fel az $\overline{\omega}_h$ rácsháló pontjaiban[4].

Legyen $L : C^2(\overline{\Omega}) \to C(\Omega) \cap C(\Gamma)$ egy olyan lineáris operátor, amelyről feltesszük, hogy az

$$Lu = \tilde{f} \tag{11.2.38}$$

egyenlet minden $\tilde{f} \in C(\Omega) \cap C(\Gamma)$ esetén korrekt kitűzésű. Legyen továbbá $L_h : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\overline{\omega}_h)$ olyan lineáris operátor, és $b_h \in \mathbb{F}(\overline{\omega}_h)$ olyan rögzített elem, amely mellett az

$$L_h y_h = b_h \tag{11.2.39}$$

operátoregyenlet is korrekt kitűzésű. Így léteznek (és egyértelműek) a (11.2.38) és a (11.2.39) feladatok u illetve y_h megoldásai.

Az

$$e_h = y_h - P_h u \in \mathbb{F}(\overline{\omega}_h) \tag{11.2.40}$$

hibafüggvény jelölés bevezetésével a (11.2.36) tulajdonság azt jelenti, hogy

$$\lim_{h\to 0} \|e_h\|_{\mathbb{F}(\overline{\omega}_h)} = 0. \tag{11.2.41}$$

Itt $\|\cdot\|_{\mathbb{F}(\overline{\omega}_h)}$ az $\mathbb{F}(\overline{\omega}_h)$-beli rácsfüggvényeken értelmezett valamelyik normát jelenti.

11.2.4. definíció.
Az y_h numerikus megoldást előállító numerikus módszert az $\|\cdot\|_{\mathbb{F}(\overline{\omega}_h)}$ normában konvergensnek nevezzük, ha (11.2.41) teljesül. Ha $\|e_h\|_{\mathbb{F}(\overline{\omega}_h)} = \mathcal{O}(h^p)$, akkor a módszert p-ed rendben konvergensnek nevezzük.

[4] A P_h operátort *projekciós operátornak* (más szóval: $\overline{\omega}_h$*-ra képező projekciónak*) nevezzük.

A továbbiakban a valamely norma melletti konvergenciával és annak rendjével foglalkozunk.

A (11.2.40) összefüggésből $y_h = e_h + P_h u$. Ezt behelyettesítve a (11.2.39) numerikus módszert leíró egyenletbe az

$$L_h e_h = b_h - L_h P_h u \tag{11.2.42}$$

összefüggést kapjuk. Bevezetve a (11.2.42) jobb oldalán szereplő rácsfüggvényre a $\Psi_h \in \mathbb{F}(\overline{\omega}_h)$ jelölést, azaz a

$$\Psi_h = b_h - L_h P_h u \tag{11.2.43}$$

jelöléssel az

$$L_h e_h = \Psi_h \tag{11.2.44}$$

ún. *hibaegyenletet* kapjuk. A (11.2.38) alapján $P_h L u = P_h \tilde{f}$. Ezért a (11.2.43) alapján Ψ_h felírható a

$$\Psi_h = (b_h - P_h \tilde{f}) + (P_h L u - L_h P_h u) \tag{11.2.45}$$

alakban. Ebben az alakban a jobb oldali első kifejezés azt mutatja, hogy a (11.2.38) és a (11.2.39) feladatokban a jobb oldalak milyen közel vannak egymáshoz, azaz hogy b_h milyen pontosan approximálja az $\tilde{f}$ függvényt az $\overline{\omega}_h$ rácshálón. A (11.2.45) összefüggésben a második tag azt mutatja, hogy az L_h operátor milyen pontosan approximálja az $\overline{\omega}_h$ rácshálón az L operátort az u megoldásfüggvényre. Természetes elvárás, hogy $h \to 0$ esetén Ψ_h nullához tartson.

11.2.5. definíció.

Azt mondjuk, hogy a numerikus módszer a $\|\cdot\|_{\mathbb{F}(\overline{\omega}_h)}$ normában konzisztens, ha $\lim_{h\to 0} \|\Psi_h\|_{\mathbb{F}(\overline{\omega}_h)} = 0$. Ha $\lim_{h\to 0} \|\Psi_h\|_{\mathbb{F}(\overline{\omega}_h)} = \mathcal{O}(h^p)$, akkor a módszert (az adott normában) p-ed rendben konzisztensnek nevezzük.

Mint azt az elöző fejezetben láttuk, a konzisztenciából még nem mutatható meg közvetlenül a konvergencia: ehhez az L_h operátorok egy további tulajdonsága is kellett. Eddig azt tettük fel, hogy a (11.2.38) és a (11.2.39) feladatok korrekt kitűzésűek. Ez azt jelenti, hogy mindkét feladatnak létezik egyértelmű megoldása (egzisztencia és unicitás), és ezek folytonosan függnek a feladatot leíró paraméterektől (stabilitás). Ez az L illetve L_h operátorokra nézve azt a követelményt eredményezi, hogy az operátorok invertálhatóak, és az inverz operátorok korlátosak. A numerikus módszert leíró L_h operátorról tehát megköveteljük, hogy létezzen L_h^{-1} és az korlátos legyen, azaz létezzen olyan $K(h) \geq 0$, amely mellett $\|L_h^{-1}\|_{\mathbb{F}(\overline{\omega}_h)} \leq K(h)$. Ugyanakkor a numerikus módszer viselkedését (pontosabban a konvergenciáját) $\lim_{h\to 0}$ esetén vizsgáljuk. Tehát, ha h csökkenésével a fenti $K(h)$ állandó kinő a végtelenbe, akkor ezt a korlátossági (azaz stabilitási) tulajdonságát a numerikus módszer elveszti. Definiálujnk egy olyan tulajdonságot az $\{\mathbf{L}_h\}$ operátorokra, amelyek mellett ez nem fordulhat elő.

11.2.6. definíció.

Azt mondjuk, hogy a (11.2.39) numerikus módszer stabil a $\|\cdot\|_{\mathbb{F}(\overline{\omega}_h)}$ normában, ha az (L_h^{-1}) operátorsereg *egyenletesen korlátos*, azaz létezik olyan $K > 0$, h-tól független állandó, amely mellett

$$\|L_h^{-1}\|_{\mathbb{F}(\overline{\omega}_h)} \leq K \tag{11.2.46}$$

minden megengedett h érték esetén.

11.2.7. megjegyzés. Ha a stabilitási tulajdonság minden h értékre igaz, akkor a módszert *feltétel nélkül stabilnak*, ha csak valamely $h_0 > 0$ szám melletti $h < h_0$ értékekre, akkor *feltételesen stabilnak* nevezzük. ◇

Most már bebizonyíthatjuk az alaptételt.

11.2.8. tétel.
Legyen a (11.2.38) feladat korrekt kitűzésű. Tegyük fel továbbá, hogy a (11.2.39) diszkretizált feladatok

- korrekt kitűzésűek,
- konzisztensek a $\|\cdot\|_{\mathbb{F}(\overline{\omega}_h)}$ normában,
- stabilak a $\|\cdot\|_{\mathbb{F}(\overline{\omega}_h)}$ normában.

Ekkor a numerikus módszer konvergens a $\|\cdot\|_{\mathbb{F}(\overline{\omega}_h)}$ normában, és konvergenciájának rendje megegyezik a konzisztenciájának rendjével.

Bizonyítás. A (11.2.44) egyenlőség és a tétel feltételeinek következtében $e_h = L_h^{-1}\Psi_h$, azaz

$$\|e_h\|_{\mathbb{F}(\overline{\omega}_h)} \leq \|L_h^{-1}\|_{\mathbb{F}(\overline{\omega}_h)}\|\Psi_h\|_{\mathbb{F}(\overline{\omega}_h)} \leq K\|\Psi_h\|_{\mathbb{F}(\overline{\omega}_h)}. \tag{11.2.47}$$

Ezért, ha a konzisztencia p-ed rendű, azaz $\Psi_h = \mathcal{O}(h^p)$, akkor

$$\|e_h\|_{\mathbb{F}(\overline{\omega}_h)} \leq K \cdot \mathcal{O}(h^p) = \mathcal{O}(h^p), \tag{11.2.48}$$

ami a tétel állítását igazolja. ■

Fontos megjegyeznünk, hogy a 11.2.8. tétel azt mutatja, hogy a közvetlenül nem (vagy csak nagyon ritkán) bizonyítható konvergencia két, lényegesen könnyebben ellenőrizhető tulajdonsággal biztosítható, nevezetesen a konzisztenciával és a stabilitással. Vegyük észre ugyanis, hogy a konvergencia közvetlen belátásához a (11.2.38) folytonos feladat u megoldásának ismerete szükséges. (Ezt viszont tipikusan nem ismerjük, hiszen a numerikus módszerek alkalmazásának éppen az az oka, hogy a feladatot nem tudjuk analitikusan megoldani.) A konzisztencia és stabilitás belátásához viszont nem szükséges u ismerete. A konzisztenciát (és annak rendjét) egy megfelelően sima függvényosztályon mutatjuk meg, és az ismeretlen megoldás azon tulajdonságát használjuk fel csupán, hogy kielégíti az egyenletet. A stabilitás a numerikus séma egy belső tulajdonsága, a (11.2.38) folytonos feladat u megoldásának nincs szerepe benne.

11.2.4. Az elliptikus feladatok numerikus közelítésének konvergenciája

Ebben részben a (11.2.32) feladat numerikus megoldásának maximumnormabeli konvergenciájával foglalkozunk, ahol az egyenletben az operátor és a jobb oldal megválasztása (11.2.33) és (11.2.34) alakú. Ennek belátásához a 11.2.8. tételt alkalmazzuk. Tehát további feladatunk a maximumnormabeli konzisztencia és stabilitás vizsgálata.

11.2.9. tétel.
A (11.2.32)-(11.2.34) véges differenciás approximáció a maximumnormában másodrendben konzisztens a (11.2.27)-(11.2.28) feladattal.

Bizonyítás. Azt kell megmutatnunk, hogy minden $(x_i, y_j) \in \overline{\omega}_h$ pontban $\Psi_h(x_i, y_j) = \mathcal{O}(h^2)$. Mivel $(x_i, y_j) \in \gamma_h$ esetén $\Psi_h(x_i, y_j) = 0$, ezért elegendő az állítást csak az $(x_i, y_j) \in \omega_h$ belső

rácspontokra megmutatnunk. Ezekben a pontokban

$$\begin{aligned}&\Psi_h(x_i,y_j) = b_h(x_i,y_j) - (L_hP_hu)(x_i,y_j) = \\ &f(x_i,y_j) + \frac{1}{h^2}\left(u(x_i+h,y_j)+u(x_i-h,y_j)-2u(x_i,y_j)\right)+ \\ &\frac{1}{h^2}\left(u(x_i,y_j+h)+u(x_i,y_j-h)-2u(x_i,y_j)\right) - c_{i,j}u(x_i,y_j).\end{aligned} \tag{11.2.49}$$

Mivel

$$f(x_i,y_j) = (Lu)(x_i,y_j) = -\frac{\partial^2 u}{\partial x^2}(x_i,y_j) - \frac{\partial^2 u}{\partial y^2}(x_i,y_j) + c(x_i,y_j)u(x_i,y_j), \tag{11.2.50}$$

ezért a (11.2.49) kifejezés átírható a következő alakra:

$$\begin{aligned}&\Psi_h(x_i,y_j) = \\ &\left(\frac{u(x_i+h,y_j)+u(x_i-h,y_j)-2u(x_i,y_j)}{h^2} - \frac{\partial^2 u}{\partial x^2}(x_i,y_j)\right)+ \\ &\left(\frac{u(x_i,y_j+h)+u(x_i,y_j-h)-2u(x_i,y_j)}{h^2} - \frac{\partial^2 u}{\partial y^2}(x_i,y_j)\right)+ \\ &\left(c(x_i,y_j)u(x_i,y_j) - c_{i,j}u(x_i,y_j)\right).\end{aligned} \tag{11.2.51}$$

Ekkor $u(x_i \pm h, y_j)$ és $u(x_i, y_j \pm h)$ kifejezések szokásos Taylor-sorba fejtésével, valamint a $c_{i,j} = c(x_i,y_j)$ egyenlőség figyelembevételével

$$\Psi_h(x_i,y_j) = \frac{1}{12}\left(\frac{\partial^4 u}{\partial x^4}(x_i,y_j) + \frac{\partial^4 u}{\partial y^4}(x_i,y_j)\right)h^2 + \mathcal{O}(h^4). \tag{11.2.52}$$

Ezért tehát $\Psi_h(x_i,y_j) = \mathcal{O}(h^2)$, ahol a h^2-es vezető tag együtthatója $M_4/6$. Ezzel beláttuk a tételt jelentő

$$\|\Psi_h\|_\infty = \mathcal{O}(h^2) \tag{11.2.53}$$

állítást. ■

Térjünk át a stabilitás vizsgálatára! Első lépésben megvizsgáljuk a numerikus módszert leíró (11.2.35) feladat egy fontos tulajdonságát.

11.2.10. tétel.
Tekintsük a (11.2.33) szerint definiált $L_h : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\overline{\omega}_h)$ rácsoperátort, és legyen $\mathbf{L}_h \in \mathbb{R}^{(N+1)^2\times(N+1)^2}$ az ezen operátornak megfeleltetett mátrix. Ekkor $\mathbf{L}_h$ M-mátrix.

Bizonyítás. A (11.2.33) definícióból nyilvánvalóan az $\mathbf{L}_h$ mátrix diagonálon kívüli elemei nempozitívak. Ezért elegendő megmutatni, hogy létezik olyan $\mathbb{R}^{(N+1)^2}$-beli $\mathbf{g}_h > 0$ vektor, amelyre $\mathbf{L}_h\mathbf{g}_h > 0$. Legyen $g_h \in \mathbb{F}(\overline{\omega}_h)$ a következő rácsfüggvény:

$$g_h(x_i,y_j) = [1+ih(l-ih)] + [1+jh(l-jh)]. \tag{11.2.54}$$

Jelölje $\mathbf{g}_h$ azt a vektort, amelynek k-adik koordinátája $\mathbf{g}_{h,k} = g_h(x_i,y_j)$, ahol $k = j(N+1)+i$ és $i,j = 0,1,\ldots,N$.[5] Ekkor nyilvánvalóan $\mathbf{g}_{h,k} \geq 2$ és $(\mathbf{L}_h\mathbf{g}_h)_k = 2$ azokban a k-adik koordinátákban, amelyekre $i \in \{0,N\}$ és $j \in \{0,N\}$. (Vegyük észre, hogy ezen pontok a γ_h határpontokhoz tartoznak.) Egyszerű behelyettesítéssel ellenőrizhető, hogy $(x_i,y_j) \in \omega_h$ esetén

[5] Az egyszerűbb jelölés kedvéért a koordináták indexelését nullától indítjuk.

$(L_h g_h)(x_i, y_j) \geq 4$, vagyis az $i = 1, 2, \ldots, N-1$ és $j = 1, 2, \ldots, N-1$ értékekhez tartó k-adik koordinátákra $(\mathbf{L}_h \mathbf{g}_h)_k \geq 5/2$.[6] Ezért tehát a (11.2.54) megválasztású $\mathbf{g}_h$ vektor mellett $\mathbf{L}_h \mathbf{g}_h \geq 2$, ami az állításunkat igazolja. ■

A fenti tétel alapján az $\mathbf{L}_h$ mátrixok invertálhatók. A továbbiakban a megmutajuk, hogy az inverzeire a maximumnormában jó becslés adható.

11.2.11. tétel.
A (11.2.33) szerinti $L_h : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\overline{\omega}_h)$ rácsoperátornak megfeleltetett $\mathbf{L}_h \in \mathbb{R}^{(N+1)^2 \times (N+1)^2}$ reguláris mátrix inverzeire érvényes az

$$\|\mathbf{L}_h^{-1}\|_\infty \leq \frac{l^2 + 4}{4} \tag{11.2.55}$$

becslés.

Bizonyítás. Mivel $\mathbf{L}_h$ M-mátrix, ezért az inverzének maximumnormája felülről becsülhető az alábbi módon:

$$\|\mathbf{L}_h^{-1}\|_\infty \leq \frac{\|\mathbf{g}_h\|_\infty}{\min_i (\mathbf{L}_h \mathbf{g}_h)_i}, \tag{11.2.56}$$

ahol $\mathbf{g}_h$ a (11.2.54) szerinti vektor. Az előzőekben megmutattuk, hogy $\min_i (\mathbf{L}_h \mathbf{g}_h)_i = 2$. Másrészt, a számtani-mértani közepek közötti összefüggés alapján $ih(l-ih) \leq l^2/4$, azaz $\|g_h\|_\infty \leq (l^2+4)/2$. Innen a (11.2.55) állításunk közvetlenül adódik. ■

Így a 11.2.9. és a 11.2.11. tételek alapján a 11.2.8. alaptétel alkalmazható, és érvényes az alábbi, konvergenciára vonatkozó állítás.

11.2.12. tétel.
Tegyük fel, hogy a (11.2.27)-(11.2.28) feladatnak létezik $u \in C^4(\overline{\Omega})$ megoldása. Ekkor a (11.2.32)-(11.2.34) numerikus séma megoldása a maximumnormában másodrendben konvergens.

11.2.5. A numerikus módszer realizálásának algoritmusa

A (11.2.32)-(11.2.34) módszer az $y_h \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvény meghatározását jelenti. Ez nyilvánvalóan a (11.2.35) lineáris algebrai egyenletrendszer megoldásával történik, ami a gyakorlatban a következőt jelenti. Ha $y_{i,j}$ jelöli az y_h rácsfüggvény $(x_i, y_j) \in \bar{\omega}_h$ rácspontbeli értékét, akkor

$$\begin{cases} -\dfrac{y_{i+1,j} - 2y_{i,j} + y_{i-1,j}}{h^2} - \dfrac{y_{i,j+1} - 2y_{i,j} + y_{i,j-1}}{h^2} + c_{i,j} y_{i,j} = f_{i,j}, \;\; i,j = 1,2,\ldots,N-1; \\ \\ y_{i,j} = \mu_{i,j}, \qquad i \in \{0, N\} \quad \text{vagy} \quad j \in \{0, N\}. \end{cases} \tag{11.2.57}$$

(Itt a $c_{i,j} = c(x_i, y_j)$, $\mu_{i,j} = \mu(x_i, y_j)$ és az $f_{i,j} = f(x_i, y_j)$ egyszerűsítő jelöléseket használtuk.) Írjuk fel a (11.2.57) feladatot a (11.2.35) lineáris algebrai egyenletrendszer alakjában, azaz határozzuk meg a feladatban szereplő $\mathbf{L}_h$ mátrix és a $\mathbf{b}_h$ jobb oldali vektor alakját.

A (11.2.57) alakjából könnyen látható, hogy azon k indexekre, amelyekre $i, j \in \{0, N\}$ és $k = j(N+1) + i$, az $\mathbf{y}_h$ vektor k-adik koordinátájának értékét a peremfeltételből ismerjük. (Ezek

[6] Lásd az előző fejezet a 10.5.5. szakaszát.

valójában a γ_h pontjaiban felírt egyenleteknek felelnek meg.) Ezért $\mathbf{y}_h$ meghatározásához elegendő csak az ω_h pontjaihoz tartozó koordinátákat meghatározni. Ez azt jelenti, hogy a (11.2.35) feladat helyett egy kisebb dimenziójú lineáris algebrai egyenletrendszer megoldása szükséges: az $\mathbf{y}_h \in \mathbb{R}^{(N+1)^2}$ vektorból a fenti koordináták elhagyása után nyert $\mathbb{R}^{(N-1)^2}$-beli vektor meghatározása szükséges. Ezt a vektort a séma ω_h pontjaiban felírt feltételekből határozhatjuk meg, amely egy $(N-1)^2$ ismeretlenes lineáris algebrai egyenletrendszert jelent. Határozzuk meg ezt a feladatot a (11.2.57) alapján!

A $k=(i-1)(N-1)+j$ (ahol $i,j=1,2,\ldots,N-1$) átsorszámozással az ω_h rácsháló pontjaihoz tartozó $\tilde{\mathbf{y}}_h$ ismeretlen vektorra átírt feladat ekkor az

$$\tilde{\mathbf{L}}_h\tilde{\mathbf{y}}_h=\tilde{\mathbf{b}}_h \tag{11.2.58}$$

lineáris algebrai egyenletrendszer alakját ölti, ahol $\tilde{\mathbf{y}}_h \in \mathbb{R}^{(N-1)^2}$ az ismeretlen vektor, $\tilde{\mathbf{b}}_h \in \mathbb{R}^{(N-1)^2}$ ismert vektor, és $\tilde{\mathbf{L}}_h \in \mathbb{R}^{(N-1)^2\times(N-1)^2}$ adott mátrix. Határozzuk meg ezek alakját a (11.2.57) összefüggések alapján!

Vezessük be az $\mathbf{y}_i \in \mathbb{R}^{(N-1)}$, $\mathbf{b}_i \in \mathbb{R}^{(N-1)}$ (ahol $i=1,2,\ldots,N-1$) vektorokat úgy, hogy mindegyik rögzített i indexű vektor j-edik koordinátája ($j=1,2,\ldots,N-1$) az $\tilde{\mathbf{y}}_h$ és a $\tilde{\mathbf{b}}_h$ vektorok fenti átsorszámozási szabálynak megfelelő k-adik koordinátája legyen. Ekkor

$$\tilde{\mathbf{y}}_h=[\mathbf{y}_1^\top,\mathbf{y}_2^\top,\ldots,\mathbf{y}_{N-1}^\top]^\top,\quad \tilde{\mathbf{b}}_h=[\mathbf{b}_1^\top,\mathbf{b}_2^\top,\ldots,\mathbf{b}_{N-1}^\top]^\top.$$

Az egyszerűség kedvéért tegyük fel, hogy $c=$ állandó. Alkalmazva a $\mathbf{B}_h \in \mathbb{R}^{(N-1)\times(N-1)}$ jelölést a $\mathbf{B}_h=\mathrm{tridiag}[-1,4+c,-1]$ tridiagonális mátrixra, illetve a $\mathbf{0},\mathbf{E}_h \in \mathbb{R}^{(N-1)\times(N-1)}$ jelöléseket a zéro- illetve az egységmátrixra, a (11.2.58) egyenletben a mátrix

$$\tilde{\mathbf{L}}_h=\frac{1}{h^2}\begin{pmatrix}\mathbf{B}_h & -\mathbf{E}_h & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ -\mathbf{E}_h & \mathbf{B}_h & -\mathbf{E}_h & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & -\mathbf{E}_h & \mathbf{B}_h & -\mathbf{E}_h\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & -\mathbf{E}_h & \mathbf{B}_h\end{pmatrix} \tag{11.2.59}$$

alakú hipermátrix. A jobb oldali vektorban szereplő $\mathbf{b}_i$ vektorok elemeire a következő kifejezéseket kapjuk.

A $\mathbf{b}_1$ vektor elemei:

$$\mathbf{b}_{1,1}=\frac{1}{h^2}\left(\mu_{0,1}+\mu_{1,0}\right)+f_{1,1},\ \mathbf{b}_{1,j}=\frac{1}{h^2}\mu_{j,0}+f_{j,1},\ \mathbf{b}_{1,N-1}=\frac{1}{h^2}\left(\mu_{N,1}+\mu_{N-1,0}\right)+f_{N-1,1},$$

ahol $j=2,3,\ldots,N-2$. A $\mathbf{b}_i$ vektor elemei $i=2,3,\ldots,N-2$ esetén:

$$\mathbf{b}_{i,1}=\frac{1}{h^2}\mu_{0,i}+f_{1,i},\ \mathbf{b}_{i,j}=f_{j,i},\quad \mathbf{b}_{i,N-1}=\frac{1}{h^2}\mu_{N,i}+f_{N-1,i},$$

ahol $j=2,3,\ldots,N-2$. A $\mathbf{b}_{N-1}$ vektor elemei:

$$\mathbf{b}_{N-1,1}=\frac{1}{h^2}\left(\mu_{0,N-1}+\mu_{1,N}\right)+f_{1,N-1},\ \mathbf{b}_{N-1,j}=\frac{1}{h^2}\mu_{j,N}+f_{N,j},$$
$$\mathbf{b}_{N-1,N-1}=\frac{1}{h^2}\left(\mu_{N,N-1}+\mu_{N-1,N}\right)+f_{N-1,N-1},$$

ahol $j=2,3,\ldots,N-2$.

Tehát a numerikus megoldás előállításához a fenti módon összeállított (11.2.58) lineáris algebrai egyenletrendszert kell megoldanunk.

Az algoritmus MATLAB programjával és a számítógépes realizálási kérdésével a fejezet végén foglalkozunk.

11.3. Lineáris, másodrendű, parabolikus parciális differenciálegyenletek

Ebben a szakaszban a lineáris, másodrendű, parabolikus parciális differenciálegyenletekkel és azok véges differenciák módszerével történő numerikus megoldásával foglalkozunk. Felírjuk a forrásmentes hővezetési egyenlet homogén, első peremfeltételű feladatának analitikus megoldását, majd meghatározzuk a véges differencia módszeres numerikus megoldást. A módszer konvergenciájának vizsgálatánál megmutatjuk, hogy az elliptikus típusú feladatokra alkalmazott technika lényegében kiterjeszthető erre a feladatra is.

11.3.1. Az egydimenziós hővezetési egyenlet analitikus megoldása

Tetszőleges $t^\star > 0$ adott szám esetén jelölje $\Omega_{t^\star} = (0,1) \times (0,t^\star] \subset \mathbb{R}^2$ halmazt, $\Gamma_{t^\star}$ pedig a $t = 0$, $x = 0$ és $x = 1$ egyenesekkel határolt ponthalmazt, azaz $\Gamma_{t^\star} = \overline{\Omega}_{t^\star} \backslash \Omega_{t^\star}$. A $\Gamma_{t^\star}$ halmazt *parabolikus peremnek* nevezzük. Tekintsük a továbbiakban $\Omega_{t^\star}$ pontjaiban a

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \quad (x,t) \in \Omega_{t^\star} \tag{11.3.1}$$

egyenletet a $\Gamma_{t^\star}$ parabolikus peremen megadott alábbi kiegészítő ("kezdeti+perem") feltételekkel:

$$u(x,0) = \mu_0(x), \quad x \in (0,1); \quad u(0,t) = u(1,t) = 0, \quad t \in [0,t^\star]. \tag{11.3.2}$$

Ha $\mu_0 = 0$, akkor a feladat megoldása az $u = 0$ függvény. Ezért a továbbiakban feltesszük, hogy $\mu_0 \neq 0$, így a keresett megoldás $u \neq 0$.

Keressük a megoldást ismét az

$$u(x,t) = X(x) \cdot T(t) \tag{11.3.3}$$

szétválasztható alakban, ahol X és T a nem azonosan nulla, egyelőre ismeretlen és megfelelően sima függvények. Behelyettesítve a (11.3.3) alakú u függvényt a (11.3.1) egyenletbe, a

$$T'(t) \cdot X(x) - X''(x) \cdot T(t) = 0, \quad (x,t) \in \Omega_{t^\star} \tag{11.3.4}$$

egyenletet nyerjük. Ekkor a (11.3.4) összefüggés alapján az

$$\frac{X''(x)}{X(x)} = \frac{T'(t)}{T(t)} \tag{11.3.5}$$

azonosságot nyerjük, ami az elliptikus esethez hasonlóan azt jelenti, hogy mindkét oldal állandó: valamely $\lambda \in \mathbb{R}$ szám mellett érvényes az

$$\frac{X''(x)}{X(x)} = \frac{T'(t)}{T(t)} = \lambda \tag{11.3.6}$$

egyenlőség. Innen az

$$X''(x) = \lambda X(x), \quad x \in (0,1) \tag{11.3.7}$$

egyenletet kapjuk. Másrészt, behelyettesítve a (11.3.3) alakot a (11.3.2) két peremfeltételébe, az

$$X(0) = X(1) = 0 \tag{11.3.8}$$

feltételt nyerjük. Célunk tehát olyan $\lambda \in \mathbb{R}$ szám meghatározása, amely mellett a (11.3.7)-(11.3.8) feladatnak létezik a triviális $X(x) = 0$ függvénytől különböző megoldása. Követve a (11.2.8)-(11.2.9) feladat megoldását, innen a lehetséges λ értékekre a

$$\lambda_k = -k^2\pi^2, \quad k = 1, 2, \dots \tag{11.3.9}$$

értékeket kapjuk. Tehát az

$$X_k(x) = C_1^k \sin(k\pi x), \quad k = 1, 2, \dots \tag{11.3.10}$$

függvények tetszőleges C_1^k állandók mellett megoldásai a $\lambda = \lambda_k$ megválasztású (11.3.7)-(11.3.8) feladatnak.

Térjünk át a $T(t)$ függvény meghatározására! A (11.3.6) egyenlőség felhasználásával a

$$T'(t) = \lambda T(t), \quad t \in (0, t^\star] \tag{11.3.11}$$

egyenletet kapjuk, ahol (11.3.9) alapján $\lambda = \lambda_k = -k^2\pi^2$. Az egyenlet általános megoldása

$$T_k(t) = C_2^k e^{\lambda_k t}, \tag{11.3.12}$$

ahol C_2^k tetszőleges állandó. Tehát a (11.3.10) és a (11.3.12) képletekkel azt kapjuk, hogy minden $k = 1, 2, \dots$ esetén tetszőleges C_k állandó mellett az

$$u_k(x,t) = X_k(x)T_k(t) = C_k e^{-k^2\pi^2 t} \sin(k\pi x) \tag{11.3.13}$$

függvények olyan függvények, amelyek megoldásai a (11.3.1) egyenletnek, és kielégítik a (11.3.2) mindkét (homogén) peremfeltételt. Így az

$$u(x,t) = \sum_{k=1}^{\infty} u_k(x,t) \tag{11.3.14}$$

függvény is rendelkezik ezekkel a tulajdonságokkal. Válasszuk meg a tetszőleges C_k állandókat úgy, hogy a (11.3.2) kezdeti feltétele is teljesüljön erre az u függvényre! A (11.3.14) és a (11.3.13) képletek alapján

$$u(x,0) = \sum_{k=1}^{\infty} C_k \sin(k\pi x). \tag{11.3.15}$$

Ugyanakkor a $\mu_0(x)$ függvény Fourier-sora

$$\mu_0(x) = \sum_{k=1}^{\infty} \mu_0^k \sin(k\pi x) \tag{11.3.16}$$

alakú, ahol

$$\mu_0^k = 2\int_0^1 \mu_0(s) \sin(k\pi s)\, \mathrm{d}s. \tag{11.3.17}$$

A (11.3.15) és a (11.3.16) képletek összevetéséből tehát

$$C_k = \mu_0^k. \tag{11.3.18}$$

Összegezve: az

$$u(x,t) = \sum_{k=1}^{\infty} \mu_0^k e^{-k^2\pi^2 t} \sin(k\pi x) \tag{11.3.19}$$

függvénysorral definiált $u(x,t)$ függvény a függvénysor egyenletes konvergenciája esetén megoldása a (11.3.1)-(11.3.2) feladatnak.

11.3.1. megjegyzés. A fenti megoldás segítségével a

$$\frac{\partial U(x,t)}{\partial t} - \frac{\partial^2 U(x,t)}{\partial x^2} = 0, \quad (x,t) \in \Omega_{t^\star} \tag{11.3.20}$$

$$U(x,0) = \mathrm{M}_0(x), \quad x \in (0,1); \quad U(0,t) = \alpha, \quad u(1,t) = \beta, \quad t \in [0,t^\star] \tag{11.3.21}$$

(ahol α és β adott állandók) inhomogén peremfeltételű feladat megoldása is könnyen előállítható hasonló végtelen függvénysor alakjában. Ehhez vezessük be

$$\hat{u}(x,t) = \alpha(1-x) + \beta x, \quad x \in [0,1] \tag{11.3.22}$$

jelölést. (Nyilvánvalóan $\hat{u}(x,t)$ egy ismert függvény.) Ekkor az $u = U - \hat{u}$ függvény megoldása lesz a (11.3.1)-(11.3.2) feladatnak, ahol $\mu_0(x) = \mathrm{M}_0(x) - (\alpha(1-x) + \beta x)$. Ezért ennek a függvénynek a (11.3.17) képlet szerinti Fourier-együtthatóival előállított (11.3.19) függvénysora meghatározza az u függvényt, és ennek ismeretében az ismeretlen függvényt kiszámolhatjuk az $U = u + \hat{u}$ összefüggésből. ◇

11.3.2. megjegyzés. Felmerülhet a kérdés: nem létezik-e a (11.3.1)-(11.3.2) feladatnak a (11.3.19) függvénytől eltérő megoldása? A válasz nemleges. Ugyanis ha egy $w(x,t)$ függvény az $\Omega_{t^\star}$ halmazon kielégíti a (11.3.1) egyenletet, és folytonos az $\overline{\Omega}_{t^\star}$ halmazon, akkor teljesül rá az ún. *parabolikus maximum-minimum elv*: a függvény a legnagyobb és legkisebb értékét felveszi a $\Gamma_{t^\star}$ parabolikus peremen. Ebből közvetlenül könnyen megmutatható, hogy a feladatnak csak egyetlen megoldása lehet. Emellett, a feladat stabil kitűzése is megmutatható: a megoldás a bemenő függvényektől folytonsan függ a maximumnormában [9, 30]. Tehát a (11.3.1)-(11.3.2) feladat korrekt kitűzésű. ◇

Mint látható, a (11.3.1)-(11.3.2) feladat, a (11.2.1)-(11.2.2) elliptikus feladathoz hasonlóan, formálisan ugyan megoldható, de a megoldást valójában csak egy végtelen függvénysor összegének alakjában tudjuk felírni. Ez azt jelenti, hogy a gyakorlatban már az ilyen egyszerű feladat analitikus megoldása sem kivitelezhető. Tehát ismételten numerikus megoldás alkalmazása szükséges. A továbbiakban a *véges differenciák módszerével* fogjuk a közelítő megoldást előállítani egy, a (11.3.1)-(11.3.2) feladatnál általánosabb feladatra, és megvizsgáljuk, hogy a közelítő megoldások közel lesznek-e a pontos megoldáshoz.

11.3.2. A hővezetési feladat numerikus megoldása véges differenciák módszerével

Ebben a részben a (11.3.1)-(11.3.2) feladatnál általánosabb alakú parabolikus típusú parciális differenciálegyenlet véges differenciás megoldásával foglalkozunk. Tekintsük a

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = f(x,t), \quad (x,t) \in (0,l) \times (0,t^\star] \tag{11.3.23}$$

egyenletet az

$$u(x,0) = \mu_0(x), \quad x \in (0,l) \tag{11.3.24}$$

kezdeti és az

$$u(0,t) = \mu_1(t), \quad u(l,t) = \mu_2(t), \quad t \in (0,t^\star] \tag{11.3.25}$$

peremfeltétellel, ahol f és μ_0, μ_1, μ_2 adott függvények. Legyen $\Omega_{t^\star} = (0,l)\times(0,t^\star] \subset \mathbb{R}^2$ azon halmaz, amelynek pontjaiban a (11.3.23) egyenletet felírjuk. Jelölje $\Gamma_{t^\star}$ az $\overline{\Omega}_{t^\star}\backslash\Omega_{t^\star}$ parabolikus peremet, μ pedig azt a $\Gamma_{t^\star}$ halmazon értelmezett függvényt, amely az egyes szakaszokon megegyezik a μ_0, μ_1 és μ_2 függvényekkel. Tegyük fel, hogy $f \in C(\Omega_{t^\star})$ és $\mu \in C(\Gamma_{t^\star})$.

Vezessük be az $L : C^{2,1}(\overline{\Omega}_{t^\star}) \to C(\Omega_{t^\star}) \cap C(\Gamma_{t^\star})$ lineáris operátort[7] a következő módon:

$$Lw(x,t) = \begin{cases} \left(\dfrac{\partial w}{\partial t} - \dfrac{\partial^2 w}{\partial x^2}\right)(x,t), & \text{ha } (x,t) \in \Omega_{t^\star}; \\ w(x,t), & \text{ha } (x,t) \in \Gamma_{t^\star}, \end{cases} \tag{11.3.26}$$

továbbá $\tilde{f} : \overline{\Omega}_{t^\star} \to \mathbb{R}$ a következő függvényt:

$$\tilde{f}(x,t) = \begin{cases} f(x,t), & \text{ha } (x,t) \in \Omega_{t^\star}; \\ \mu(x,t), & \text{ha } (x,t) \in \Gamma_{t^\star}. \end{cases} \tag{11.3.27}$$

Ekkor a (11.3.23)-(11.3.24) feladatunk az

$$Lu = \tilde{f} \tag{11.3.28}$$

operátoregyenlet megoldását jelenti, ahol $u \in C^{2,1}(\overline{\Omega}_{t^\star})$ az ismeretlen függvény.

Mivel a (11.3.28) feladat analitikus megoldását általános esetben nem tudjuk előállítani, ezért ismételten numerikus eljárást alkalmazunk. Ennek lényege a következő.

1. Definiálunk az $\overline{\Omega}_{t^\star}$ halmazon rácshálókat a következő módon:

$$\omega_{h,\tau} = \{(x_i,t_n),\ \ x_i = ih,\ t_n = n\tau,\ i = 1,2,\dots,N_x-1,\ n = 1,2,\dots,N_t\}$$

$$\overline{\omega}_{h,\tau} = \{(x_i,t_n),\ \ x_i = ih,\ t_n = n\tau,\ i = 0,1,\dots,N_x,\ n = 0,1,\dots,N_t\}.$$

   Itt N_x és N_t jelölik az x és t irányú osztásrészek számát, $h = l/N_x$ és $\tau = t^\star/N_t$ pedig a diszkretizációs lépésközöket. Jelölje $\gamma_{h,\tau} = \overline{\omega}_{h,\tau}\backslash\omega_{h,\tau} \subset \Gamma_{t^\star}$ az $\overline{\omega}_{h,\tau}$ rácsháló $\Gamma_{t^\star}$ parabolikus peremre eső pontjait.

2. Jelölje $\mathbb{F}(\overline{\omega}_{h,\tau})$ és $\mathbb{F}(\omega_{h,\tau})$ az $\overline{\omega}_{h,\tau}$ és az $\omega_{h,\tau}$ rácsokon értelmezett, $\mathbb{R}$-be képező függvények vektorterét.

3. Célunk olyan $y_{h,\tau} \in \mathbb{F}(\overline{\omega}_{h,\tau})$ rácsfüggvény meghatározása, amely $\overline{\omega}_{h,\tau}$ pontjaiban közel van a (11.3.1)-(11.3.2) feladat u megoldásához, és a rácsháló finomításával (azaz $h,\tau \to 0$ esetén) az eltérésük nullához tart.

Adjunk meg tehát olyan $L_{h,\tau} : \mathbb{F}(\overline{\omega}_{h,\tau}) \to \mathbb{F}(\overline{\omega}_{h,\tau})$ lineáris operátort és $b_{h,\tau} \in \mathbb{F}(\overline{\omega}_h)$ elemet, amelyekre az

$$L_{h,\tau}y_{h,\tau} = b_{h,\tau} \tag{11.3.29}$$

operátoregyenlet $y_{h,\tau} \in \mathbb{F}(\overline{\omega}_{h,\tau})$ megoldása rendelkezik a fentiekben leírt tulajdonsággal.

Az $L_{h,\tau}$ operátor megválasztásánál ötletként a következő lemma szolgál.

[7] $C^{2,1}(\overline{\Omega}_{t^\star})$ jelöli az első változóban kétszer, a második változóban egyszer folytonosan differenciálható $u(x,t)$ (ahol $(x,t) \in \overline{\Omega}_{t^\star}$) függvények halmazát.

11.3.3. lemma.
Legyen $w \in C^{4,2}(\overline{\Omega}_{t^\star})$. Ekkor tetszőleges $\vartheta \in [0,1]$ szám esetén

$$\begin{aligned}\frac{\partial w}{\partial t}(x_i, t_{n-1}+\vartheta\tau) &= \frac{w(x_i,t_{n-1}+\tau)-w(x_i,t_{n-1})}{\tau} + \mathcal{O}(\tau),\\ \frac{\partial^2 w}{\partial x^2}(x_i, t_{n-1}+\vartheta\tau) &= \frac{w(x_{i+1},t_{n-1})-2w(x_i,t_{n-1})+w(x_{i-1},t_{n-1})}{h^2} + \mathcal{O}(\tau+h^2).\end{aligned} \tag{11.3.30}$$

Bizonyítás. Írjuk fel az első állítás bal oldalán szereplő függvény elsőrendű Taylor-sorbafejtését!

$$\frac{\partial w}{\partial t}(x_i, t_{n-1}+\vartheta\tau) = \frac{\partial w}{\partial t}(x_i,t_{n-1}) + \vartheta\tau\frac{\partial^2 w}{\partial t^2}(x_i,t_*) = \frac{\partial w}{\partial t}(x_i,t_{n-1}) + \mathcal{O}(\tau). \tag{11.3.31}$$

Mivel

$$w(x_i,t_{n-1}+\tau) = w(x_i,t_{n-1}) + \tau\frac{\partial w}{\partial t}(x_i,t_{n-1}) + \frac{\tau^2}{2}\frac{\partial^2 w}{\partial t^2}(x_i,t_{**}), \tag{11.3.32}$$

ezért

$$\frac{w(x_i,t_{n-1}+\tau)-w(x_i,t_{n-1})}{\tau} = \frac{\partial w}{\partial t}(x_i,t_{n-1}) + \mathcal{O}(\tau). \tag{11.3.33}$$

A (11.3.31) és a (11.3.33) összefüggésekből az első állítás közvetlenül leolvasható.

A második állítás az első állítás bizonyításával könnyen belátható, és ezért ezt az Olvasóra bízzuk. ■

Célunk a (11.3.26) szerinti $(Lw)(x_i,t_n)$ érték közelítése $\mathcal{O}(\tau+h^2)$ pontossággal a w függvény $\overline{\omega}_{h,\tau}$ rácspontbeli értékeivel. Tekintsük a 11.3.3. lemma állítását a $\vartheta = 1$ megválasztással! Ez motiválja a következőket.

Jelölje a $w_{h,\tau} \in \mathbb{F}(\overline{\omega}_{h,\tau})$ adott rácsfüggvény esetén $w_h(x_i,t_n) = w_i^n$. Definiáljuk az $L_{h,\tau} : \mathbb{F}(\overline{\omega}_{h,\tau}) \to \mathbb{F}(\overline{\omega}_{h,\tau})$ rácsoperátort az alábbi módon:

$$\left(L_{h,\tau}w_{h,\tau}\right)(x_i,t_n) = \begin{cases} \dfrac{w_i^n - w_i^{n-1}}{\tau} - \dfrac{w_{i+1}^{n-1} - 2w_i^{n-1} + w_{i-1}^{n-1}}{h^2}, & \text{ha } (x_i,t_n) \in \omega_{h,\tau};\\ w_i^n, & \text{ha } (x_i,t_n) \in \gamma_{h,\tau}. \end{cases} \tag{11.3.34}$$

Alkalmazva az $f(x_i,t_{n-1}+\vartheta\tau) =: f_i^{n,\vartheta}$ és a $\mu(x_i,t_n) = \mu_i^n$ jelöléseket, valamely rögzített $\vartheta \in [0,1]$ érték mellett definiáljuk a $b_{h,\tau}^\vartheta \in \mathbb{F}(\overline{\omega}_{h,\tau})$ rácsfüggvényt a következő módon:

$$b_{h,\tau}^\vartheta(x_i,t_n) = \begin{cases} f_i^{n,\vartheta}, & \text{ha } (x_i,t_n) \in \omega_{h,\tau};\\ \mu_i^n & \text{ha } (x_i,t_n) \in \gamma_{h,\tau}. \end{cases} \tag{11.3.35}$$

Ekkor $b_{h,\tau}^\vartheta$ értéke $\overline{\omega}_{h,\tau}$ minden rácspontjában meghatározható, [8] és a (11.3.29) egyenlet azt jelenti, hogy keressük azon $y_{h,\tau}^\vartheta \in \mathbb{F}(\overline{\omega}_{h,\tau})$ rácsfüggvényt, amelyet a (11.3.34) szerinti $L_{h,\tau}$ operátor ebbe a $b_{h,\tau}^\vartheta$ rácsfüggvénybe képez le.

A továbbiakban, ahol nem okoz félreértést, a jelöléseinken elhagyjuk a ϑ felső indexet. (Mint látni fogjuk a 11.3.7.. tételben, a ϑ paraméter konkrét megválasztása nincs kihatással a módszer konvergenciájára.)

[8] Ha $\vartheta = 0$, akkor az $n = 1$ értékre a képletben $f(0,x_i)$ szerepel. Bár az f függvényről eddig csak azt tettük fel, hogy az $\Omega_{t^\star}$ tartományon van értelmezve, és ott folytonos, a továbbiakban feltesszük, hogy folytonosan kiterjeszthető a $t = 0$ egyenesre. Ekkor az $f(0,x_i)$ értéken ezen kiterjesztett függvény értékét értjük, azaz a képlet értelmes az $n = 1$ esetén is.

11.3.4. megjegyzés. A 11.2.3. megjegyzéshez hasonlóan a (11.3.29) lineáris operátoregyenlet felírható

$$\mathbf{L}_{h,\tau}\mathbf{y}_{h,\tau} = \mathbf{b}_{h,\tau} \tag{11.3.36}$$

alakú lineáris algebrai egyenletrendszerként, amelynek mérete megegyezik az $\overline{\omega}_{h,\tau}$ rácsháló pontjainak számával, azaz $(N_x+1)(N_t+1)$. Később látni fogjuk, hogy esetünkben $\mathbf{y}_{h,\tau}$ meghatározásához valójában nincs szükség az $\mathbf{L}_{h,\tau}$ mátrix invertálására. ⋄

11.3.3. A véges differenciás közelítés konvergenciája

A továbbiakban a 11.3.2. részben megfogalmazott feladat megoldhatóságával és konvergenciájával foglalkozunk.

11.3.5. tétel.
Tegyük fel, hogy a rácsháló lépésközeire teljesül a

$$q := \tau/h^2 \leq 0.5 \tag{11.3.37}$$

feltétel. Ekkor a (11.3.29) feladatnak létezik egyértelmű megoldása.

Bizonyítás. A 11.3.4. megjegyzésnek megfelelően a (11.3.29) feladat egy (11.3.36) alakú lineáris algebrai egyenletrendszert jelent. Írjuk ki az egyes rácspontokhoz tartozó egyenleteket!

- az $(x_i,t_n)\in\omega_{h,\tau}$ belső rácspontokban (11.3.36) a következő egyenletet jelenti:

$$\frac{1}{\tau}y_i^n - \frac{1}{h^2}y_{i-1}^{n-1} - \frac{1}{h^2}y_{i+1}^{n-1} + \left(\frac{2}{h^2}-\frac{1}{\tau}\right)y_i^{n-1} = f_i^{n,\vartheta}. \tag{11.3.38}$$

- az $(x_i,t_n)\in\gamma_{h,\tau}$ perempontokban (11.3.36) pedig a következőt jelenti:

$$y_i^n = \mu_i^n. \tag{11.3.39}$$

Ezért az $\mathbf{L}_{h,\tau}$ mátrix főátlójában pozitív elemek $(1/\tau)$, míg a (11.3.37) feltétel következtében azon kívül nempozitív elemek állnak.

Tekinsük a $\mathbf{g}_i^n = 1+ih(l-ih)$ $(i=0,1,\dots,N_x; n=0,1,\dots,N_t)$ vektort. Nyilvánvalóan $\mathbf{g}_i^n \geq 1$, és $i=0$ illetve $i=N_x$ esetén $\mathbf{g}_i^n = 1$. Határozzuk meg az $\mathbf{L}_{h,\tau}\mathbf{g}$ vektort! Mivel a $\mathbf{g}$ vektor koordinátái az n indextől függetlenek, ezért a belső pontokhoz tartozó koordinátákra:

$$\begin{aligned}(\mathbf{L}_{h,\tau}\mathbf{g})_i^n = \frac{1}{\tau}\mathbf{g}_i^n - \frac{1}{h^2}\mathbf{g}_{i-1}^{n-1} - \frac{1}{h^2}\mathbf{g}_{i+1}^{n-1} + \left(\frac{2}{h^2}-\frac{1}{\tau}\right)\mathbf{g}_i^{n-1} = \\ \frac{1}{\tau}\underbrace{\left(\mathbf{g}_i^n - \mathbf{g}_i^{n-1}\right)}_{=0} + \frac{1}{h^2}\left(-\mathbf{g}_{i-1}^{n-1} + 2\mathbf{g}_i^{n-1} - \mathbf{g}_{i+1}^{n-1}\right) = 2,\end{aligned} \tag{11.3.40}$$

mivel egyszerű behelyettesítéssel ellenőrizhetően $-\mathbf{g}_{i-1}^n + 2\mathbf{g}_i^n - \mathbf{g}_{i+1}^n = 2h^2$. (V.ö. a 10.. fejezet 10.2.2. szakaszával.)
Tehát $\mathbf{L}_{h,\tau}\mathbf{g} > 0$. Ezért $\mathbf{L}_{h,\tau}$ M-mátrix, ami az állításunkat bizonyítja. ■

11.3.6. következmény. A szokásos módon ismét becslés adható az $\mathbf{L}_{h,\tau}$ mátrix inverzének maximumnormájára:

$$\|\mathbf{L}_{h,\tau}^{-1}\|_\infty \leq \frac{\|\mathbf{g}\|_\infty}{\min_{i,n}(\mathbf{L}_h\mathbf{g})_i^n}$$

Az előzőekben megmutattuk, hogy $\min_{i,n}(\mathbf{L}_h\mathbf{g})_i^n = 1$. Másrészt, a számtani-mértani közepek közötti összefüggés alapján $ih(l-ih) \leq l^2/4$, azaz $\|g_h\|_\infty \leq (l^2+4)/4$. Ezért tehát az $\mathbf{L}_{h,\tau}$ reguláris mátrix inverzére érvényes az

$$\|\mathbf{L}_{h,\tau}^{-1}\|_\infty \leq \frac{l^2+4}{4} \tag{11.3.41}$$

becslés. $\diamond$

Térjünk át a konvergencia vizsgálatára! Legyen $P_{h,\tau}$ a $C(\overline{\Omega}_{t^*}) \to \mathbb{F}(\overline{\omega}_{h,\tau})$ projekció, azaz

$$(P_{h,\tau}w)(x_i,t_n) = w(x_i,t_n), \quad (x_i,t_n) \in \overline{\omega}_{h,\tau}, \tag{11.3.42}$$

továbbá $w_{h,\tau} = P_{h,\tau}w \in \mathbb{F}(\overline{\omega}_{h,\tau})$ a w függvény projekcióját. Ekkor a konvergencia valamely $\|\cdot\|_{\mathbb{F}(\overline{\omega}_{h,\tau})}$ nomában azt jelenti, hogy az $e_{h,\tau} = y_{h,\tau} - P_{h,\tau}u = y_{h,\tau} - u_{h,\tau} \in \mathbb{F}(\overline{\omega}_h)$ rácsfüggvény h és τ nullához tartása esetén ebben a normában nullához tart. A következő tétel ezt a tulajdoságot mutatja meg.

11.3.7. tétel.
A (11.3.37) feltétel teljesülése esetén a (11.3.34) és (11.3.35) képletekkel definiált (11.3.29) egyenlet $y_{h,\tau} \in \mathbb{F}(\overline{\omega}_h)$ megoldásaiból álló rácsfüggvénysorozat a maximumnormában tetszőleges $\vartheta \in [0,1]$ érték esetén $\Psi_{h,\tau}(x_i,t_n) = \mathcal{O}(\tau + h^2)$ rendben konvergál a (11.3.28) feladat elegendően sima megoldásához.

Bizonyítás. Feltehető, hogy a (11.3.28) korrekt kitűzésű feladat (lásd a 11.3.1. szakaszt), ezért elegendő megmutatni, hogy a módszer a maximumnormában konzisztens és stabil.

Vizsgáljuk meg a konzisztenciát! Ehhez a lokális approximációs hiba, vagyis a

$$\Psi_{h,\tau}(x_i,t_n) = b_{h,\tau}(x_i,t_n) - (L_{h,\tau}P_{h,\tau}u)(x_i,t_n) = (b_{h,\tau} - (L_{h,\tau}u_{h,\tau}))(x_i,t_n)$$

kifejezés abszolút értékének maximumát kell becsülnünk. Mivel az $(x_i,t_n) \in \gamma_{h,\tau}$ perempontokban

$$\Psi_{h,\tau}(x_i,t_n) = \mu_i^n - (L_{h,\tau}u_{h,\tau})(x_i,t_n) = \mu_i^n - u_{h,\tau}(x_i,t_n) = \mu_i^n - \mu_i^n = 0,$$

ezért elegendő a lokális approximációs hibát csak az $(x_i,t_n) \in \omega_{h,\tau}$ belső rácspontokban vizsgálnunk. Ezekben a pontokban

$$\Psi_{h,\tau}(x_i,t_n) = f_i^{n,\vartheta} - (L_{h,\tau}u_{h,\tau})(x_i,t_n) = f(x_i,t_{n-1}+\vartheta\tau) - (L_{h,\tau}u_{h,\tau})(x_i,t_n). \tag{11.3.43}$$

A (11.3.23) egyenlet alapján

$$f(x_i,t_{n-1}+\vartheta\tau) = \frac{\partial u(x_i,x_i,t_{n-1}+\vartheta\tau)}{\partial t} - \frac{\partial^2 u(x_i,t_{n-1}+\vartheta\tau)}{\partial x^2}, \tag{11.3.44}$$

és az $u_{h,\tau}(x_i,t_n) = u_i^n$ jelöléssel a (11.3.34) összefüggés alapján

$$(L_{h,\tau}u_{h,\tau})(x_i,t_n) = \frac{u_i^n - u_i^{n-1}}{\tau} - \frac{u_{i+1}^{n-1} - 2u_i^{n-1} + u_{i-1}^{n-1}}{h^2}. \tag{11.3.45}$$

A (11.3.44) és a (11.3.45) összefüggések alapján tehát a (11.3.43) kifejezés átírható a következő alakban:

$$\Psi_{h,\tau}(x_i,t_n)=\left(\frac{\partial u(x_i,t_{n-1}+\vartheta\tau)}{\partial t}-\frac{u_i^{n+1}-u_i^n}{\tau}\right)+\left(\frac{u_{i+1}^n-2u_i^n+u_{i-1}^n}{h^2}-\frac{\partial^2 u(x_i,t_{n-1}+\vartheta\tau)}{\partial x^2}\right). \tag{11.3.46}$$

Ezért a 11.3.3. lemma állítása alapján tetszőleges $\vartheta\in[0,1]$ érték esetén

$$\Psi_{h,\tau}(x_i,t_n)=\mathcal{O}(\tau+h^2), \tag{11.3.47}$$

ami a maximumnormabeli konzisztenciát, és annak rendjét mutatja.

Mivel a 11.3.6.. következményből (konkrétan a (11.3.41) becslésből) a stabilitás közvetlenül következik, ezért a 11.2.8. tétel felhasználásával ezzel a tétel állítását beláttuk. ■

11.3.8. megjegyzés. A 11.3.7. tétel így is megfogalmazható: A (11.3.37) feltétel teljesülése esetén a (11.3.38)-(11.3.39) lineáris algebrai egyenletrendszer megoldása a maximumnormában konvergál a (11.3.23)-(11.3.25) hővezetési feladat elegendően sima megoldásához. ◇

11.3.4. A numerikus módszer realizálásának algoritmusa

A konvergencia megmutatása után térjünk át a módszer realizálásának kérdésére.

A 11.3.4. megjegyzésnek megfelelően a módszer realizálása a (11.3.36) alakú lineáris algebrai egyenletrendszer megoldásával ekvivalens. Mivel a megoldás a parabolikus perem mentén ismert, ezért valójában csak az $(x_i,t_n)\in\omega_{h,\tau}$ belső rácspontokhoz tartozó numerikus megoldások értékét kell meghatároznunk, azaz valójában egy

$$\tilde{\mathbf{L}}_{h,\tau}\tilde{\mathbf{y}}_{h,\tau}=\tilde{\mathbf{b}}_{h,\tau} \tag{11.3.48}$$

alakú lineáris egyenletrendszer megoldása szükséges, ahol $\tilde{\mathbf{L}}_{h,\tau}$ egy $(N_x-1)N_t\times(N_x-1)N_t$ méretű adott mátrix, $\tilde{\mathbf{b}}_{h,\tau}$ adott, $\tilde{\mathbf{y}}_{h,\tau}$ pedig az ismeretlen (11.3.36)-beli $\mathbf{y}_{h,\tau}$ vektor $\omega_{h,\tau}$ rácspontjaihoz tartozó komponenseiből álló, $(N_x-1)N_t$ méretű vektorok.

Határozzuk meg a (11.3.48) egyenletben szereplő $\tilde{\mathbf{L}}_{h,\tau}$ mátrix és $\tilde{\mathbf{b}}_{h,\tau}$ vektor alakját! Ehhez tehát a (11.3.38) egyenleteket kell felírnunk az ismert kezdeti és peremfeltételek (azaz μ függvény) felhasználásával. Nyilvánvalóan (11.3.38) felírható

$$y_i^n-qy_{i-1}^{n-1}-qy_{i+1}^{n-1}+(2q-1)\,y_i^{n-1}=\tau f_i^{n,\vartheta} \tag{11.3.49}$$

($i=1,2,\dots,N_x-1$ és $n=0,1,\dots,N_t-1$) alakban, ahol $q=\tau/h^2$. A (11.3.24) és a (11.3.25) feltételekből (vagyis a (11.3.39) egyenletekből) a (11.3.49) egyenletekben $i\in\{0,N_x\}$ illetve $n=0$ esetén ismerjük az y_i^n értékeit: ezekre az indexekre $y_i^n=\mu(x_i,t_n)$. Ez azt jelenti, hogy

$$y_i^0=\mu_0(x_i),\quad y_0^n=\mu_1(t_n);\quad y_{N_x}^n=\mu_2(t_n),\ i=0,1,\dots,N_x,\ n=1,2,\dots,N_t. \tag{11.3.50}$$

Vezessük be az $\mathbf{y}^n\in\mathbb{R}^{(N_x-1)}$, $\mathbf{b}^n\in\mathbb{R}^{(N_x-1)}$ (ahol $n=1,2,\dots,N_t$) szokásosan oszlopvektorokat az alábbi módon:

$$\mathbf{y}^n=[y_1^n,y_2^n,\dots,y_{N_x-1}^n]^\top,\quad \mathbf{b}^n=[b_1^n,b_2^n,\dots,b_{N_x-1}^n]^\top,$$

Ekkor

$$\tilde{\mathbf{y}}_{h,\tau}=[(\mathbf{y}^1)^\top,(\mathbf{y}^2)^\top,\dots,(\mathbf{y}^{N_t})^\top]^\top,\quad \tilde{\mathbf{b}}_{h,\tau}=[(\mathbf{b}^1)^\top,(\mathbf{b}^2)^\top,\dots,(\mathbf{b}^{N_t})^\top]^\top.$$

Az egyszerűség kedvéért tegyük fel, hogy $c =$ állandó. Alkalmazva a $\mathbf{B}_h \in \mathbb{R}^{(N_x-1)\times(N_x-1)}$ jelölést a $\mathbf{B}_h = \text{tridiag}[-q, 2q-1, -q]$ tridiagonális mátrixra, illetve a $\mathbf{0}, \mathbf{E}_h \in \mathbb{R}^{(N_x-1)\times(N_x-1)}$ jelöléseket a zéró- illetve az egységmátrixra, a (11.3.48) egyenletben a mátrix

$$\tilde{\mathbf{L}}_{h,\tau} = \begin{pmatrix} \mathbf{E}_h & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_h & \mathbf{E}_h & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{B}_h & \mathbf{E}_h & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{B}_h & \mathbf{E}_h \end{pmatrix} \tag{11.3.51}$$

alakban írható fel, ami egy N_t sorból álló alakú hipermátrix.

11.3.9. megjegyzés. Könnyen ellenőrizhető, hogy az $\tilde{\mathbf{L}}_{h,\tau}$ mátrix a (11.3.37) szerinti $q \le 0.5$ feltétel esetén M-mátrix. ⋄

A (11.3.49) egyenletekből közvetlenül felírhatók a $\mathbf{b}^n$ $N_x - 1$ dimenziós vektorok. Ezek alakja a következő:

$$\mathbf{b}^1 = \begin{pmatrix} \tau f_1^{0,\vartheta} + q y_0^0 - (2q-1) y_1^0 + q y_2^0 \\ \tau f_2^{0,\vartheta} + q y_1^0 - (2q-1) y_2^0 + q y_3^0 \\ \vdots \\ \tau f_{N_x-1}^{0,\vartheta} + q y_{N_x-2}^0 - (2q-1) y_{N_x-1}^0 + q y_{N_x}^0 \end{pmatrix}, \tag{11.3.52}$$

és az $n = 2, 3, \dots, N_t$ értékekre

$$\mathbf{b}^n = \begin{pmatrix} \tau f_1^{n-1,\vartheta} + q y_0^{n-1} \\ \tau f_2^{n-1,\vartheta} \\ \vdots \\ \tau f_{N_x-2,\vartheta}^{n-1} \\ \tau f_{N_x-1,\vartheta}^{n-1} + q y_{N_x}^{n-1} \end{pmatrix}. \tag{11.3.53}$$

11.3.10. megjegyzés. Mint beláttuk, a konvergencia rendjét nem befolyásolja a ϑ paraméter értéke. A gyakorlatban általában a $\vartheta = 0.5$ megválasztás a szokásos. Ennek oka, hogy olyan pontot kell választanunk, hogy az azon ponthoz tartozó függvényérték jól jellemezze az $f(x_i, t)$ függvény viselkedését a $[t_{n-1}, t_n]$ intervallumon. Mint azt a numerikus integrálásnál is láttuk, erre a felezőpont a legalkalmasabb. ⋄

A (11.3.50) összefüggések ismeretében tehát

$$\mathbf{b}^1 = \begin{pmatrix} \tau f_1^{0,\vartheta} + q\mu_0(x_0) - (2q-1)\mu_0(x_1) + q\mu_0(x_2) \\ \tau f_2^{0,\vartheta} + q\mu_0(x_1) - (2q-1)\mu_0(x_2) + q\mu_0(x_3) \\ \vdots \\ \tau f_{N_x-1}^{0,\vartheta} + q\mu_0(x_{N_x-2}) - (2q-1)\mu_0(x_{N_x-1}) + q\mu_0(x_{N_x}) \end{pmatrix}, \tag{11.3.54}$$

és az $n = 2, 3, \dots, N_t$ értékekre

$$\mathbf{b}^n = \begin{pmatrix} \tau f_1^{n-1,\vartheta} + q\mu_1(t_{n-1}) \\ \tau f_2^{n-1,\vartheta} \\ \vdots \\ \tau f_{N_x-2}^{n-1,\vartheta} \\ \tau f_{N_x-1}^{n-1,\vartheta} + q\mu_2(t_{n-1}) \end{pmatrix}. \tag{11.3.55}$$

Mindezek alapján soronként kiírva a (11.3.48) egyenletet a következőt kapjuk

$$\begin{aligned} \mathbf{E}_h\mathbf{y}^1 &= \mathbf{b}^1, \\ \mathbf{B}_h\mathbf{y}^{n-1} + \mathbf{E}_h\mathbf{y}^n &= \mathbf{b}^n, \quad n = 2, 3, \ldots, N_t, \end{aligned}$$

amelynek átrendezésével az alábbi megoldó algoritmust kapjuk:

$$\begin{aligned} \mathbf{y}^1 &= \mathbf{b}^1, \\ \mathbf{y}^n &= \mathbf{b}^n - \mathbf{B}_h\mathbf{y}^{n-1}, \quad n = 2, 3, \ldots, N_t. \end{aligned} \tag{11.3.56}$$

A (11.3.56) ún. *explicit* módszert jelent: az ismert $\mathbf{b}^n$ vektorok ismeretében közvetlenül nyerjük időrétegenként a közelítő megoldást. Emellett a módszer *egylépéses* (vagy más terminiológiában: *kétlépcsős*), hiszen az $\mathbf{y}^{n-1}$ vektorból határozzuk meg a $\mathbf{y}^n$ vektort, azaz a megoldás új időrétegen való kiszámolása a megelőző időrétegen már kiszámolt közelítés segítségével történik. Ezért a módszer realizálása meglehetősen egyszerű, és nem igényel sok műveletet: az egy időrétegen való megoldásvektor kiszámolásához egy $(N_x - 1)$ méretű tridiagonális mátrix és vektor összeszorzása, valamint egy vektorösszeadás szükséges.

11.3.5. Egy másik véges differenciás séma és vizsgálata

Az eddigiekben tárgyalt, a (11.3.34) és (11.3.35) képletekkel definiált (11.3.29) *explicit séma* konvergenciájának feltétele volt a $\tau/h^2 \leq 0.5$ (11.3.37) feltétel. (Lásd a 11.3.7. tételt.) Ez a feltétel eléggé megszorító, ha a rácshálót finomítjuk. Ugyanis, amikor az $\overline{\Omega}_{t^\star}$ megoldási halmazon értelmezett $\overline{\omega}_{h,\tau}$ rácsháló térbeli lépésközét csökkentjük, akkor ennek mértékének négyzetével kell az időbeli lépésközt is csökkenteni. (Például, h felezésével τ értékét negyedére kell csökkentenünk.) Ez viszont a rácsháló pontjainak számát (és így a számítási igényt) is jelentősen megnöveli. (Az előző példánk esetén N_t a négyszeresére növekszik, azaz amíg az eredeti rácsháló belső pontjainak – és ezért az ismeretleneknek is – a száma $(N_x - 1)N_t$, addig a finomított $(2N_x - 1)4N_t$ lesz, azaz kb. nyolcszorosára növekszik a pontok száma.)

Felmerülhet a kérdés: megadható-e olyan közelítés, amelynek konvergenciájához nem szükséges ilyen korlátozó feltétel?

A (11.3.29) diszkretizációban a (11.3.23)-(11.3.25) feladatban szereplő deriváltakat a (11.3.30) szerint közelítettük. (Ez igazából az (x_1, t_{n-1}) pontban felírt közelítést jelentette, ahol az idő szerinti deriváltat egy haladó véges differenciával közelítettük.) Írjunk fel egy másik, bizonyos értelemben természetesebb közelítést: a közelítést az (x_i, t_n) pontban (tehát a rácsháló pontjában) írjuk fel, és az idő szerinti deriváltat egy retrográd véges differenciával közelítjük. Tehát

$$\begin{aligned} \frac{\partial w}{\partial t}(x_i, t_n) &\simeq \frac{w(x_i, t_n) - w(x_i, t_n - \tau)}{\tau}, \\ \frac{\partial^2 w}{\partial x^2}(x_i, t_n) &\simeq \frac{w(x_{i+1}, t_n) - 2w(x_i, t_n) + w(x_{i-1}, t_n)}{h^2}, \end{aligned} \tag{11.3.57}$$

azaz a térbeli változó szerinti második deriváltat egy adott időrétegen a *rácshálóval megegyező időrétegen* vett középponti véges differenciákkal közelítjük. Hasonlóan, a jobb oldali rácsfüggvényre

$$b_{h,\tau}(x_i,t_n)=\begin{cases} f_i^n, & \text{ha } (x_i,t_n)\in\omega_{h,\tau};\\ \mu_i^n, & \text{ha } (x_i,t_n)\in\gamma_{h,\tau}.\end{cases} \tag{11.3.58}$$

Ekkor az $L_{h,\tau}$ operátor definíciója a (11.3.34) szerinti képlet helyett a következőre módosul:

$$\left(L_{h,\tau}w_{h,\tau}\right)(x_i,t_n)=\begin{cases} \dfrac{w_i^n-w_i^{n-1}}{\tau}-\dfrac{w_{i+1}^n-2w_i^n+w_{i-1}^n}{h^2}, & \text{ha } (x_i,t_n)\in\omega_{h,\tau};\\ w_i^n, & \text{ha } (x_i,t_n)\in\gamma_{h,\tau}.\end{cases} \tag{11.3.59}$$

A 11.3.5. tételhez hasonlóan a (11.3.59) operátor invertálhatósága (és így a megfelelő numerikus séma realizálhatósága) megmutatható.

11.3.11. tétel.
A (11.3.59) operátorú (11.3.29) feladatnak létezik egyértelmű megoldása.

Bizonyítás. A 11.3.4. megjegyzésnek megfelelően a (11.3.29) feladat egy (11.3.36) alakú lineáris algebrai egyenletrendszert jelent. Írjuk ki az egyes rácspontokhoz tartozó egyenleteket!

- Felhasználva az operátor (11.3.59) alakját, az $(x_i,t_n)\in\omega_{h,\tau}$ belső rácspontokban (11.3.36) a következő egyenletet jelenti:

$$\left(\frac{1}{\tau}+\frac{2}{h^2}\right)y_i^n-\frac{1}{h^2}y_{i-1}^n-\frac{1}{h^2}y_{i+1}^n-\frac{1}{\tau}y_i^{n-1}=f_i^n. \tag{11.3.60}$$

- Az $(x_i,t_n)\in\gamma_{h,\tau}$ perempontokban (11.3.36) a következőt jelenti:

$$y_i^n=\mu_i^n. \tag{11.3.61}$$

Ezért a (11.3.59) alakú $\mathbf{L}_{h,\tau}$ mátrix főátlójában pozitív elemek, míg azon kívül nempozitív elemek állnak.

Tekinsük a $\mathbf{g}_i^n=1+ih(l-ih)$ $(i=0,1,\dots,N_x;n=0,1,\dots,N_t)$ vektort! Nyilvánvalóan $\mathbf{g}_i^n\geq 1$, és $i=0$ illetve $i=N_x$ esetén $\mathbf{g}_i^n=1$. Határozzuk meg az $\mathbf{L}_{h,\tau}\mathbf{g}$ vektort! A perempontokhoz tartozó koordinátái nyilván $\mathbf{g}_i^n$. Mivel a $\mathbf{g}$ vektor koordinátái az n indextől függetlenek, ezért a belső pontokhoz tartozó koordinátákra:

$$\left(\mathbf{L}_{h,\tau}\mathbf{g}\right)_i^n=\frac{1}{\tau}\underbrace{\left(\mathbf{g}_i^n-\mathbf{g}_i^{n-1}\right)}_{=0}+\frac{1}{h^2}\left(-\mathbf{g}_{i-1}^n+2\mathbf{g}_i^n-\mathbf{g}_{i+1}^n\right)=2, \tag{11.3.62}$$

mivel, mint ismeretes $-\mathbf{g}_{i-1}^{n-1}+2\mathbf{g}_i^{n-1}-\mathbf{g}_{i+1}^{n-1}=2h^2$.

Tehát $\mathbf{L}_{h,\tau}\mathbf{g}>0$. Ezért $\mathbf{L}_{h,\tau}$ M-mátrix, ami az állításunkat bizonyítja. ■

11.3.12. következmény. A szokásos módon ismét becslés adható az $\mathbf{L}_{h,\tau}$ mátrix inverzének maximumnormájára:

$$\|\mathbf{L}_{h,\tau}^{-1}\|_\infty\leq\frac{\|\mathbf{g}\|_\infty}{\min_{i,n}(\mathbf{L}_h\mathbf{g})_i^n}.$$

Az előzőekben megmutattuk, hogy $\min_{i,n}(\mathbf{L}_h\mathbf{g})_i^n = 1$. Másrészt, a számtani-mértani közepek közötti összefüggés alapján $ih(l-ih) \leq l^2/4$, azaz $\|g_h\|_\infty \leq (l^2+4)/4$. Ezért tehát az $\mathbf{L}_{h,\tau}$ reguláris mátrix inverzére érvényes az

$$\|\mathbf{L}_{h,\tau}^{-1}\|_\infty \leq \frac{l^2+4}{4} \tag{11.3.63}$$

becslés. ◇

11.3.13. megjegyzés. A 11.3.5. és a 11.3.11. tételek közötti alapvető különbség, hogy míg a 11.3.5. tételben feltételt kötünk ki a τ/h^2 hányadosra (lásd (11.3.37) feltétel), addig a 11.3.11. tételben ilyen korlát nem szerepel. ◇

A következő állítás a módszer konvergenciájáról szól.

11.3.14. tétel.
A (11.3.59) és (11.3.58) képletekkel definiált (11.3.29) egyenlet $y_{h,\tau} \in \mathbb{F}(\overline{\omega}_h)$ megoldásaiból álló rácsfüggvénysorozat a maximumnormában $\mathcal{O}(\tau+h^2)$ rendben konvergál a (11.3.28) feladat elegendően sima megoldásához.

Bizonyítás. A tétel bizonyításához alkalmazzuk a 11.3.7. tétel bizonyítását.
A konzisztencia vizsgálata a $\Psi_{h,\tau}(x_i,t_n)$ lokális approximációs hiba rendjének vizsgálatát igényli. Nyilvánvalóan esetünkben

$$\Psi_{h,\tau}(x_i,t_n) = \left(\frac{\partial u(x_i,t_n)}{\partial t} - \frac{u_i^n - u_i^{n-1}}{\tau}\right) + \left(\frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} - \frac{\partial^2 u(x_i,t_n)}{\partial x^2}\right). \tag{11.3.64}$$

A fenti egyenlőség jobb oldalának első kifejezése

$$\frac{\partial u(x_i,t_n)}{\partial t} - \frac{u_i^n - u_i^{n-1}}{\tau} = \mathcal{O}(\tau).$$

A második kifejezésre a második derivált véges differenciás közelítésének pontossága miatt

$$\frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} - \frac{\partial^2 u(x_i,t_n)}{\partial x^2} = \mathcal{O}(h^2).$$

Ezért tehát

$$\Psi_{h,\tau}(x_i,t_n) = \mathcal{O}(\tau+h^2), \tag{11.3.65}$$

ami a maximumnorma-beli konzisztenciát mutatja.

A módszer stabilitása a 11.3.12. következményből (konkrétan a (11.3.63) becslésből) közvetlenül következik.

A 11.2.8. tétel állítását alkalmazva ezzel állításunkat beláttuk. ■

11.3.15. megjegyzés. A fenti tétel fontos következménye, hogy a módszerünk a τ és h tetszőleges módon történő nullához tartása esetén konvergens a maximumnormában, azaz a diszkretizációs lépésközöket csak az approximáció pontosságának kívánalmai szerint választhatjuk meg, nincs szükség az egyéb korlátozások figyelembevételére. ◇

Végezetül írjuk fel a módszer realizálását jelentő algoritmust! Ehhez határozzuk meg a (11.3.59) és (11.3.58) képletekkel definiált (11.3.29) operátoregyenletnek megfeleltetett (11.3.48) lineáris algebrai egyenletrendszer $\tilde{\mathbf{L}}_{h,\tau}$ mátrixát és $\tilde{\mathbf{b}}_{h,\tau}$ jobb oldali vektorát. Ezeket a (11.3.59) egyenletek átírásával a

$$-qy_{i-1}^n + (2+q)y_i^n - qy_{i+1}^n - y_i^{n-1} = \tau f_i^n \tag{11.3.66}$$

($i = 1, 2, \ldots, N_x - 1$ és $n = 1, 2, \ldots, N_t$) egyenletekből határozhatjuk meg. Emlékeztetünk, hogy a (11.3.24) és a (11.3.25) feltételekből (vagyis a (11.3.39) egyenletekből) a (11.3.66) egyenletekben $i \in \{0, N_x\}$ illetve $n = 0$ esetén ismerjük az y_i^n értékeit: ezekre az indexekre $y_i^n = \mu(x_i, t_n)$. Ezért

$$y_i^0 = \mu_0(x_i), \quad y_0^n = \mu_1(t_n); \quad y_{N_x}^n = \mu_2(t_n), \ i = 0, 1, \ldots, N_x, \ n = 1, 2, \ldots, N_t. \tag{11.3.67}$$

Megtartva a 11.3.4. szakasz $\mathbf{y}^n \in \mathbb{R}^{N_x-1}$, $\mathbf{b}^n \in \mathbb{R}^{N_x-1}$, valamint a $\mathbf{0}, \mathbf{E}_h \in \mathbb{R}^{(N_x-1)\times(N_x-1)}$ jelöléseit, illetve bevezetve a $\mathbf{C}_h = \text{tridiag}[-q, 1+2q, -q] \in \mathbb{R}^{(N_x-1)\times(N_x-1)}$ tridiagonális mátrix jelölését, a (11.3.48) egyenletben az

$$\tilde{\mathbf{L}}_{h,\tau} = \begin{pmatrix} \mathbf{C}_h & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{E}_h & \mathbf{C}_h & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & -\mathbf{E}_h & \mathbf{C}_h & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & -\mathbf{E}_h & \mathbf{C}_h \end{pmatrix} \tag{11.3.68}$$

N_t számú sorból álló hipermátrix. A $\mathbf{b}^n$ ($n = 1, 2, \ldots, N_t$) $(N_x - 1)$ dimenziós vektorok alakja a következő. Az $n = 1$ értékre

$$\mathbf{b}^1 = \begin{pmatrix} \tau f_1^1 + \mu_0(x_1) + q\mu_1(t_1) \\ \tau f_2^1 + \mu_0(x_2) \\ \vdots \\ \tau f_{N_x-2}^1 + \mu_0(x_{N_x-2}) \\ \tau f_{N_x-1}^1 + \mu_0(x_{N_x-1}) + q\mu_2(t_1) \end{pmatrix}, \tag{11.3.69}$$

az $n = 2, 3, \ldots, N_t$ értékekre pedig

$$\mathbf{b}^n = \begin{pmatrix} \tau f_1^n + q\mu_1(t_n) \\ \tau f_2^n \\ \vdots \\ \tau f_{N_x-2}^n \\ \tau f_{N_x-1}^n + q\mu_2(t_n) \end{pmatrix}. \tag{11.3.70}$$

Soronként kiírva a (11.3.48) egyenletet a következőt kapjuk

$$\begin{aligned} \mathbf{C}_h\mathbf{y}^1 &= \mathbf{b}^1, \\ -\mathbf{E}_h\mathbf{y}^{n-1} + \mathbf{C}_h\mathbf{y}^n &= \mathbf{b}^n, \quad n = 2, 3, \ldots, N_t. \end{aligned}$$

Ezt átrendezve az alábbi megoldó algoritmust nyerjük:

$$\begin{aligned} \mathbf{C}_h\mathbf{y}^1 &= \mathbf{b}^1, \\ \mathbf{C}_h\mathbf{y}^n &= \mathbf{b}^n + \mathbf{y}^{n-1}, \quad n = 2, 3, \ldots, N_t. \end{aligned} \tag{11.3.71}$$

A (11.3.71) ún. *implicit* módszert jelent: az ismert jobb oldali vektorok ismeretében egy $\mathbf{C}_h$ mátrixú lineáris algebrai egyenletrendszer megoldása szükséges mindegyik időrétegen. Ez a módszer is egylépéses, hiszen az $\mathbf{y}^{n-1}$ vektorból határozzuk meg az $\mathbf{y}^n$ vektort, azaz a megoldás új

időrétegen való kiszámolása a megelőző időrétegen már kiszámolt közelítés segítségével történik. (Megjegyezzük, hogy a $\mathbf{C}_h$ mátrix speciális struktúrájának következtében az időrétegenkénti lineáris algebrai egyenletrendszer megoldására alkalmazhatjuk a korábban már ismertetett speciális Gauss-eliminációt.)

11.3.6. Általánosítás és magasabb rendű módszerek

Foglaljuk össze a (11.3.23)-(11.3.25) hővezetési feladat numerikus megoldására vonatkozó eddigi eredményeinket! A numerikus módszerek felépítésénél az volt az alapkérdés, hogyan adjunk meg olyan $L_{h,\tau} : \mathbb{F}(\overline{\omega}_{h,\tau}) \to \mathbb{F}(\overline{\omega}_{h,\tau})$ lineáris operátort és $b_{h,\tau} \in \mathbb{F}(\overline{\omega}_{h,\tau})$ rácsfüggvényt, amelyekre a (11.3.29) operátoregyenlet $y_{h,\tau} \in \mathbb{F}(\overline{\omega}_{h,\tau})$ megoldása rendelkezik a szükséges tulajdonsággal. Ehhez ötletként a deriváltak véges differenciás approximációját alkalmaztuk. Az így nyert sémákra ezután megmutattuk a megoldhatóságot, illetve beláttuk a konvergenciát: mindkét séma a maximumnormában $\mathcal{O}(\tau + h^2)$ rendben tart a (11.3.23)-(11.3.25) feladat megfelelő simaságú megoldásához.

A fentiek két további természetes kérdést vetnek fel.

- Kaphatunk-e más megközelítéssel is újabb, a megfelelő tulajdonsággal rendelkező $L_{h,\tau}$ operátort és $b_{h,\tau}$ jobb oldali rácsfüggvényt?
- Megadható-e magasabb rendben pontos (azaz gyorsabban konvergáló) közelítés?

Vegyük észre, hogy az $\overline{\omega}_{h,\tau}$ rácsháló előállítható a 10. fejezetben alkalmazott rácshálónak megfelelő $\overline{\omega}_h = \{x_i = ih,\ i = 0, 1, \ldots, N_x,\ h = l/N_x\}$ térbeli és a 9. fejezetben alkalmazott rácshálónak megfelelő $\overline{\omega}_\tau = \{t_i = i\tau;\ i = 0, 1, \ldots, N_t;\ \tau = t^\star/N_t\}$ időbeli rácshálók direkt szorzataként, azaz $\overline{\omega}_{h,\tau} = \overline{\omega}_h \times \overline{\omega}_\tau$. Ezért $L_{h,\tau}$ és $b_{h,\tau}$ alakjának meghatározását két lépésben hajtjuk végre.

- Első lépésben a (11.3.23)-(11.3.25) feladatot egy elsőrendű, közönséges differenciálegyenletrendszer Cauchy-feladatával térben diszkretizáljuk az $\overline{\omega}_h$ rácshálón. (Ez az ún. *szemidiszkretizáció*.)
- Második lépésben ezen feladat időbeli diszkretizálásához alkalmazzuk az $\overline{\omega}_\tau$ rácshálón a 9. fejezetben megismert numerikus módszerek valamelyikét.

Az első lépés végrehajtásához mindegyik $x_i \in \overline{\omega}_h$ rácsponthoz hozzárendelünk egy $u_{h,i}(t)$ (ahol $t \in [0, t^\star]$) egyváltozós függvényt, mégpedig úgy, hogy azok lehetőleg közel legyenek a (11.3.23)-(11.3.25) feladat pontos megoldásának ezen rácspontbeli értékéhez, vagyis az $u(x_i, t)$ függvényhez. Mivel a (11.3.23)-(11.3.25) feladat az $x = x_i$ pontban a

$$\frac{\partial u(x_i,t)}{\partial t} - \frac{\partial^2 u(x_i,t)}{\partial x^2} = f(x_i,t), \quad i = 1, 2, \ldots, N_x - 1,\ t \in (0, t^\star] \tag{11.3.72}$$

illetve az

$$u(x_i, 0) = \mu_0(x_i), \tag{11.3.73}$$

és az

$$u(0,t) = \mu_1(t), \quad u(l,t) = \mu_2(t), \quad t \in (0, t^\star] \tag{11.3.74}$$

egyenlőségeket jelenti, ezért a csomópontokhoz tartozó ismeretlen $u_{h,i}(t)$ $(i = 0, 1, 2, \ldots, N_x)$ függvényeket úgy határozzuk meg, hogy azokra teljesüljenek az alábbi egyenletek:

$$u'_{h,i}(t) - \frac{u_{h,i+1}(t) - 2u_{h,i}(t) + u_{h,i-1}(t)}{h^2} = f_{h,i}(t), \quad i = 1, 2, \ldots, N_x,\ t \in (0, t^\star] \tag{11.3.75}$$

$$u_{h,i}(0) = \mu_0(x_i), \tag{11.3.76}$$

$$u_{h,0}(t) = \mu_1(t), \quad u_{h,N_x}(t) = \mu_2(t), \quad t \in (0, t^\star], \tag{11.3.77}$$

ahol $f_{h,i}(t) = f(x_i, t)$.

11.3.16. megjegyzés. A (11.3.75) egyenlet felírásánál a térváltozó szerinti második derivált szokásos

$$\frac{\partial^2 w}{\partial x^2}(x_i, t) \simeq \frac{w(x_{i+1}, t) - 2w(x_i, t) + w(x_{i-1}, t)}{h^2} \tag{11.3.78}$$

$\mathcal{O}(h^2)$ pontosságú véges differenciás approximációját alkalmaztuk. $\diamond$

Bevezetve a $\mathbf{Q} = \text{tridiag}[-1, 2, -1] \in \mathbb{R}^{(N_x-1)\times(N_x-1)}$ jelölést, ekkor az $u_{h,i}(t)$ koordinátafüggvényekkel rendelkező, ismeretlen $\mathbf{u}_h : [0, t^\star] \to \mathbb{R}^{N_x-1}$ függvényre - a (11.3.75)-(11.3.77) összefüggések alapján - a következő elsőrendű, közönséges differenciálegyenlet Cauchy-feladata írható fel:

$$\mathbf{u}_h'(t) + \left(\frac{1}{h^2}\mathbf{Q}\right)\mathbf{u}_h(t) = \mathbf{f}_h(t), \quad t \in (0, t^\star], \tag{11.3.79}$$

$$\mathbf{u}_h(0) = \boldsymbol{\mu}_0, \tag{11.3.80}$$

ahol

$$\begin{aligned} \mathbf{f}_h(t) =& [(f_{h,1}(t) + \mu_1(t)), f_{h,2}(t), \ldots, f_{h,N_x-2}(t), (f_{h,N_x-1}(t) + \mu_2(t))]^\top, \\ \boldsymbol{\mu}_0 =& [\mu_0(x_1), \mu_0(x_2), \ldots, \mu_0(x_{N_x-1})]^\top. \end{aligned} \tag{11.3.81}$$

A második lépés végrehajtásához numerikusan megoldjuk a (11.3.79)-(11.3.80) rendszer Cauchy-feladatát az $\overline{\omega}_\tau$ rácshálón. Ehhez választhatjuk a 9. fejezet bármely módszerét. Legyen $y_{h,\tau}(x_i, t_n)$ a megoldás i-edik koordinátafüggvényének közelítése a $t = t_n$ rácspontban. Így megkonstruáltunk egy $y_{h,\tau} \in \mathbb{F}(\overline{\omega}_{h,\tau})$ rácsfüggvényt, amelyről *be kell látni*, hogy h és τ nullához tartása esetén konvergál a pontos megoldáshoz, illetve a konvergencia rendjét is meg kell határoznunk.

A továbbiakban tekintsünk két példát.

11.3.17. példa. Válasszuk az explicit Euler-módszert a (11.3.79)-(11.3.80) rendszer Cauchy-feladatát megoldó numerikus módszernek.Ebben az esetben a numerikus megoldást az

$$\frac{\mathbf{y}_h^n - \mathbf{y}_h^{n-1}}{\tau} + \left(\frac{1}{h^2}\mathbf{Q}\right)\mathbf{y}_h^{n-1} = \mathbf{f}_h^{n-1}, \quad i = 1, 2, \ldots, N_x - 1, n = 1, 2, \ldots, N_t, \tag{11.3.82}$$

$$\mathbf{y}_h^0 = \boldsymbol{\mu}_0 \tag{11.3.83}$$

feladat megoldásával nyerjük, ahol $\mathbf{f}_h^{n-1} = \mathbf{f}_h(t_{n-1})$ és

$$\mathbf{y}_h^n = [y_{h,\tau}(x_1, t_n), y_{h,\tau}(x_2, t_n), \ldots, y_{h,\tau}(x_{N_x-1}, t_n)]^\top.$$

Könnyen látható, hogy az így definiált $y_{h,\tau} \in \mathbb{F}(\overline{\omega}_{h,\tau})$ rácsfüggvény megegyezik a (11.3.34) és (11.3.35) képletekkel definiált (11.3.29) egyenlet $y_{h,\tau} \in \mathbb{F}(\overline{\omega}_h)$ megoldásaiból álló rácsfüggvénysorozattal. Tehát a Cauchy-feladat numerikus megoldásának ezen megválasztása mellett a korábbi explicit módszerünket kapjuk. $\diamond$

11.3.18. példa. Válasszuk az implicit Euler-módszert a (11.3.79)-(11.3.80) rendszer Cauchy-feladatát megoldó numerikus módszernek. Ebben az esetben a numerikus megoldást a korábbi jelöléseink mellett az

$$\frac{\mathbf{y}_h^n - \mathbf{y}_h^{n-1}}{\tau} + \left(\frac{1}{h^2}\mathbf{Q}\right)\mathbf{y}_h^n = \mathbf{f}_h^n, \quad n = 1, 2, \ldots, N_t, \tag{11.3.84}$$

$$\mathbf{y}_h^0 = \boldsymbol{\mu}_0 \tag{11.3.85}$$

feladat megoldásával nyerjük. Könnyen látható, hogy az így definiált $y_{h,\tau} \in \mathbb{F}(\overline{\omega}_{h,\tau})$ rácsfüggvény megegyezik a (11.3.59) és (11.3.58) képletekkel definiált (11.3.29) egyenlet $y_{h,\tau} \in \mathbb{F}(\overline{\omega}_h)$ megoldásaiból álló rácsfüggvénysorozattal. Tehát a Cauchy-feladat numerikus megoldásának ezen megválasztása mellett a korábbi implicit módszerünket kapjuk. ⋄

A fenti példák alapján természetes, hogy a továbbiakban egy olyan módszert vizsgálunk meg, amelyet a 9.. fejezetben az explicit és implicit Euler-módszerek általánosításaként nyertünk: ez a θ-módszer volt. Tekintsük tehát a (11.3.23)-(11.3.25) feladat numerikus megoldására azt a numerikus módszert, amelyet a (11.3.79)-(11.3.80) feladat θ-módszerrel való megoldásával nyerünk, azaz amikor az $y_{h,\tau} \in \mathbb{F}(\overline{\omega}_{h,\tau})$ numerikus megoldást jelentő rácsfüggvényt az előre rögzített $\theta \in [0,1]$ melletti

$$\frac{\mathbf{y}_h^n - \mathbf{y}_h^{n-1}}{\tau} + \left(\frac{1-\theta}{h^2}\mathbf{Q}\right)\mathbf{y}_h^{n-1} + \left(\frac{\theta}{h^2}\mathbf{Q}\right)\mathbf{y}_h^n = \mathbf{f}_h^{n,\theta}, \quad n = 1, 2, \ldots, N_x, \tag{11.3.86}$$

$$\mathbf{y}_h^0 = \boldsymbol{\mu}_0 \tag{11.3.87}$$

feladat megoldásával kapjuk, ahol $\mathbf{f}_h^{n,\theta} = \mathbf{f}_h(t_{n-1} + \theta\tau)$. (Nyilvánvalóan $\theta = 0$ és $\theta = 1$ esetén a korábban vizsgált sémáinkat kapjuk.)

Célunk a fenti módszer rácsegyenletként való megfogalmazása. Ehhez definiáljunk egy újabb rácsoperátort. Legyen $L_{h,\tau}$ olyan, amely egy $w_{h,\tau}$ rácsfüggvényhez egy olyan rácsfüggvényt rendel hozzá, amelynek értéke az $(x_i, t_n) \in \overline{\omega}_{h,\tau}$ pontban a következő:

$$\begin{cases} \dfrac{w_i^n - w_i^{n-1}}{\tau} - \theta\dfrac{w_{i+1}^n - 2w_i^n + w_{i-1}^n}{h^2} - (1-\theta)\dfrac{w_{i+1}^{n-1} - 2w_i^{n-1} + w_{i-1}^{n-1}}{h^2}, & \text{ha } (x_i, t_n) \in \omega_{h,\tau}; \\ w_i^n, & \text{ha } (x_i, t_n) \in \gamma_{h,\tau}. \end{cases} \tag{11.3.88}$$

A korábbi speciális esetekhez hasonlóan a (11.3.88) operátor invertálhatósága (és így a megfelelő numerikus séma realizálhatósága) is közvetlenül vizsgálható.

Legyen továbbá a jobb oldali rácsfüggvény a következő:

$$b_{h,\tau}(x_i, t_n) = \begin{cases} f_i^{n,\theta}, & \text{ha } (x_i, t_n) \in \omega_{h,\tau}; \\ \mu_i^n & \text{ha } (x_i, t_n) \in \gamma_{h,\tau}. \end{cases} \tag{11.3.89}$$

Ezen jelölésekkel a (11.3.86)-(11.3.87) feladat felírható a (11.3.29) alakú operátoregyenletként. A továbbiakban ezzel a feladattal foglalkozunk.

11.3.19. tétel.
A (11.3.88) operátorú (11.3.29) feladatnak a

$$q = \frac{\tau}{h^2} \leq \frac{1}{2(1-\theta)} \tag{11.3.90}$$

feltétel teljesülése mellett létezik egyértelmű megoldása.

Bizonyítás. A 11.3.4. megjegyzésnek megfelelően a (11.3.29) feladat egy (11.3.36) alakú lineáris algebrai egyenletrendszert jelent. Írjuk ki az egyes rácspontokhoz tartozó egyenleteket!

- Felhasználva az operátor (11.3.88) alakját, az $(x_i, t_n) \in \omega_{h,\tau}$ belső rácspontokban (11.3.36) a következő egyenletet jelenti:

$$\begin{aligned}&\left(\frac{1}{\tau} + \frac{2\theta}{h^2}\right) y_i^n - \frac{\theta}{h^2} y_{i-1}^n - \frac{\theta}{h^2} y_{i+1}^n + \\ &\left(-\frac{1}{\tau} + \frac{2(1-\theta)}{h^2}\right) y_i^{n-1} - \frac{1-\theta}{h^2} y_{i-1}^{n-1} - \frac{1-\theta}{h^2} y_{i+1}^{n-1} = f_i^{n,\theta}.\end{aligned} \tag{11.3.91}$$

- Az $(x_i, t_n) \in \gamma_{h,\tau}$ perempontokban (11.3.36) a következőt jelenti:

$$y_i^n = \mu_i^n. \tag{11.3.92}$$

Ezért az $\mathbf{L}_{h,\tau}$ mátrix főátlójában pozitív elemek, míg azon kívül (a (11.3.90) feltételt figyelembevéve) nempozitív elemek állnak.

Tekintsük ismét a $\mathbf{g}_i^n = 1 + ih(l - ih)$ $(i = 0, 1, \ldots, N_x; n = 0, 1, \ldots, N_t)$ vektort! Nyilvánvalóan $\mathbf{g}_i^n \geq 1$, és $i = 0$ illetve $i = N_x$ esetén $\mathbf{g}_i^n = 1$. Határozzuk meg az $\mathbf{L}_{h,\tau}\mathbf{g}$ vektort! A perempontokhoz tartozó koordinátái nyilván $\mathbf{g}_i^n$. Mivel a $\mathbf{g}$ vektor koordinátái az n indextől függetlenek, ezért a belső pontokhoz tartozó koordinátákra:

$$(\mathbf{L}_{h,\tau}\mathbf{g})_i^n = \frac{1}{\tau}\underbrace{\left(\mathbf{g}_i^n - \mathbf{g}_i^{n-1}\right)}_{=0} + \frac{\theta}{h^2}\left(-\mathbf{g}_{i-1}^n + 2\mathbf{g}_i^n - \mathbf{g}_{i+1}^n\right) + \frac{1-\theta}{h^2}\left(-\mathbf{g}_{i-1}^{n-1} + 2\mathbf{g}_i^{n-1} - \mathbf{g}_{i+1}^{n-1}\right) = 2, \tag{11.3.93}$$

mivel, mint ismeretes, $-\mathbf{g}_{i-1}^{n-1} + 2\mathbf{g}_i^{n-1} - \mathbf{g}_{i+1}^{n-1} = 2h^2$.
Tehát $\mathbf{L}_{h,\tau}\mathbf{g} > 0$. Ezért $\mathbf{L}_{h,\tau}$ M-mátrix, ami az állításunkat bizonyítja. ■

11.3.20. következmény. A szokásos módon ismét becslés adható a (11.3.88) operátornak megfeleltetett $\mathbf{L}_{h,\tau}$ mátrix inverzének maximumnormájára:

$$\|\mathbf{L}_{h,\tau}^{-1}\|_\infty \leq \frac{\|\mathbf{g}\|_\infty}{\min_{i,n}(\mathbf{L}_h\mathbf{g})_i^n}.$$

Az előzőekben megmutattuk, hogy $\min_{i,n}(\mathbf{L}_h\mathbf{g})_i^n = 1$. Másrészt a számtani-mértani közepek közötti összefüggés alapján $ih(l - ih) \leq l^2/4$, azaz $\|g_h\|_\infty \leq (l^2 + 4)/4$. Ezért tehát az $\mathbf{L}_{h,\tau}$ reguláris mátrix inverzére érvényes az

$$\|\mathbf{L}_{h,\tau}^{-1}\|_\infty \leq \frac{l^2 + 4}{4} \tag{11.3.94}$$

becslés. ◇

11.3.21. megjegyzés. A (11.3.90) feltétel jobb oldalát $\theta = 1$ esetén szokásosan ∞ értékként értelmezzük, azaz ebben az esetben nincs feltétel q megválasztására. (Ezt már korábban is megmutattuk, lásd a 11.3.11. tételt.) A $\theta \in [0, 1)$ esetén van korlátozó feltétel. Ugyanakkor megjegyezzük, hogy a 11.3.19. tételben (11.3.90) *elégséges* feltétel, tehát szükségességéről nem szól az állítás. $\diamond$

A továbbiakban áttérünk a módszer konvergenciájának vizsgálatára. Először a módszer konzisztenciáját vizsgáljuk.

11.3.22. tétel.
A (11.3.90) feltétel teljesülése esetén a (11.3.88) operátorú és (11.3.89) jobb oldalú (11.3.29) egyenlettel definiált módszer a maximumnormában konzisztens, és rendje

- $\mathcal{O}(\tau + h^2)$, ha $\theta \neq 0.5$,
- $\mathcal{O}(\tau^2 + h^2)$, ha $\theta = 0.5$.

Bizonyítás. A definíció alapján a lokális approximációs hiba

$$\Psi_{h,\tau}(x_i, t_n) = b_{h,\tau}(x_i, t_n) - (L_{h,\tau} P_{h,\tau} u)(x_i, t_n) = (b_{h,\tau} - (L_{h,\tau} u_{h,\tau}))(x_i, t_n).$$

Mivel az $(x_i, t_n) \in \gamma_{h,\tau}$ perempontokban $\Psi_{h,\tau}(x_i, t_n) = 0$, ezért elegendő a lokális approximációs hibát csak az $(x_i, t_n) \in \omega_{h,\tau}$ belső rácspontokban vizsgálnunk. Vezessük be a $t_n^\theta = t_{n-1} + \theta\tau$ jelölést. Ekkor nyilvánvalóan

$$t_{n-1} = t_n^\theta - \theta\tau, \quad t_n = t_n^\theta + (1-\theta)\tau. \tag{11.3.95}$$

A jobb oldali rácsfüggvény (11.3.89) definicója következtében ekkor

$$\Psi_{h,\tau}(x_i, t_n) = f(x_i, t_n^\theta) - (L_{h,\tau} u_{h,\tau})(x_i, t_n). \tag{11.3.96}$$

Felírva a (11.3.23) egyenletet a $t = t_n^\theta$ pontban, az

$$f(x_i, t_n^\theta) = \frac{\partial u(x_i, t_n^\theta)}{\partial t} - \frac{\partial^2 u(x_i, t_n^\theta)}{\partial x^2} \tag{11.3.97}$$

összefüggést kapjuk. Másrészt az $u_{h,\tau}(x_i, t_n) = u_i^n$ jelöléssel a (11.3.88) alapján

$$\begin{aligned}(L_{h,\tau} u_{h,\tau})(x_i, t_n) = &\frac{u_i^n - u_i^{n-1}}{\tau} - \theta \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} - \\ &-(1-\theta)\frac{u_{i+1}^{n-1} - 2u_i^{n-1} + u_{i-1}^{n-1}}{h^2}.\end{aligned} \tag{11.3.98}$$

A (11.3.97) és a (11.3.98) összefüggések alapján tehát a (11.3.96) lokális approximációs hiba felírható

$$\Psi_{h,\tau}(x_i, t_n) = \Psi_1 + \Psi_2 \tag{11.3.99}$$

alakban, ahol

$$\Psi_1 = \frac{\partial u(x_i, t_n^\theta)}{\partial t} - \frac{u(x_i, t_n) - u(x_i, t_{n-1})}{\tau}, \tag{11.3.100}$$

$$\begin{aligned}\Psi_2 = &\theta \frac{u(x_{i+1}, t_n) - 2u(x_i, t_n) + u(x_{i-1}, t_n)}{h^2} + \\ &(1-\theta)\frac{u(x_{i+1}, t_{n-1}) - 2u(x_i, t_{n-1}) + u(x_{i-1}, t_{n-1})}{h^2} - \frac{\partial^2 u(x_i, t_n^\theta)}{\partial x^2}.\end{aligned} \tag{11.3.101}$$

Felhasználva a (11.3.95) összefüggést,

$$\begin{aligned} u(x_i,t_n) =& u(x_i,t_n^\theta+(1-\theta)\tau) = u(x_i,t_n^\theta)+(1-\theta)\tau\frac{\partial u(x_i,t_n^\theta)}{\partial t}+\mathcal{O}(\tau^2),\\ u(x_i,t_{n-1}) =& u(x_i,t_n^\theta-\theta\tau) = u(x_i,t_n^\theta)-\theta\tau\frac{\partial u(x_i,t_n^\theta)}{\partial t}+\mathcal{O}(\tau^2). \end{aligned} \tag{11.3.102}$$

Ezért a (11.3.100) és a (11.3.102) összefüggések alapján

$$\Psi_1 = \mathcal{O}(\tau^2). \tag{11.3.103}$$

Másrészt a szokásos Taylor-sorba fejtéssel

$$\begin{aligned} \frac{u(x_i+h,t_{n-1})-2u(x_i,t_{n-1})+u(x_i-h,t_{n-1})}{h^2} &= \frac{\partial^2 u(x_i,t_{n-1})}{\partial x^2}+\mathcal{O}(h^2),\\ \frac{u(x_i+h,t_n)-2u(x_i,t_n)+u(x_i-h,t_n)}{h^2} &= \frac{\partial^2 u(x_i,t_n)}{\partial x^2}+\mathcal{O}(h^2). \end{aligned}$$

Ezért

$$\Psi_2 = \theta\frac{\partial^2 u(x_i,t_{n-1})}{\partial x^2}+(1-\theta)\frac{\partial^2 u(x_i,t_n)}{\partial x^2}+\mathcal{O}(h^2). \tag{11.3.104}$$

Felhasználva a (11.3.95) összefüggést, ismét sorbafejtéssel a

$$\begin{aligned} \frac{\partial^2 u(x_i,t_{n-1})}{\partial x^2} &= \frac{\partial^2 u(x_i,t_n^\theta-\theta\tau)}{\partial x^2} = \frac{\partial^2 u(x_i,t_n^\theta)}{\partial x^2}-\theta\tau\frac{\partial^3 u(x_i,t_n^\theta)}{\partial t\partial x^2}+\mathcal{O}(\tau^2),\\ \frac{\partial^2 u(x_i,t_n)}{\partial x^2} &= \frac{\partial^2 u(x_i,t_n^\theta+(1-\theta)\tau)}{\partial x^2} = \frac{\partial^2 u(x_i,t_n^\theta)}{\partial x^2}+(1-\theta)\tau\frac{\partial^3 u(x_i,t_n^\theta)}{\partial t\partial x^2}+\mathcal{O}(\tau^2) \end{aligned} \tag{11.3.105}$$

összefüggéseket kapjuk. Behelyettesítve (11.3.105) kifejezéseket a (11.3.104) összefüggésbe,

$$\Psi_2 = \left(-\theta^2\tau+(1-\theta)^2\tau\right)\frac{\partial^3 u(x_i,t_n^\theta)}{\partial t\partial x^2}+\mathcal{O}(\tau^2)+\mathcal{O}(h^2),$$

azaz

$$\Psi_2 = (1-2\theta)\tau\frac{\partial^3 u(x_i,t_n^\theta)}{\partial t\partial x^2}+\mathcal{O}(\tau^2+h^2). \tag{11.3.106}$$

A (11.3.99) jelölés figyelembevételével ekkor a (11.3.103) és a (11.3.106) képletek a tétel állítását bizonyítják. ■

Mindezek alapján könnyen belátható a módszer konvergenciájáról szóló alábbi állítás.

11.3.23. tétel.

A (11.3.90) feltétel teljesülése esetén a (11.3.88) operátorú és (11.3.89) jobb oldalú, (11.3.29) egyenlettel definiált módszer a maximumnormában konvergál a (11.3.28) feladat elegendően sima megoldásához, és rendje

- $\mathcal{O}(\tau+h^2)$, ha $\theta\neq 0.5$,
- $\mathcal{O}(\tau^2+h^2)$, ha $\theta = 0.5$.

Bizonyítás. Mivel a (11.3.28) feladat korrekt kitűzésű (lásd a 11.3.1. szakaszt), a maximumnormában konzisztens (lásd a 11.3.22. tételt) és stabil (lásd a 11.3.20. következményt), ezért állításunk közvetlenül következik a 11.2.8. tételünkből. ■.

Vizsgáljuk meg a módszer megoldási algoritmusát! Az egyszerűbb tárgyalás kedvéért feltesszük, hogy a (11.3.23)-(11.3.25) feladatban a peremfeltételek homogének, azaz $\mu_1(t) = \mu_2(t) = 0$. Megtartva ezen szakasz eddigi, valamint a 11.3.4. szakasz $\mathbf{y}^n \in \mathbb{R}^{N_x-1}$, $\mathbf{b}^n \in \mathbb{R}^{N_x-1}$, $\mathbf{0}, \mathbf{E}_h \in \mathbb{R}^{(N_x-1)\times(N_x-1)}$ jelöléseit, illetve bevezetve a $\mathbf{D}_h = \mathbf{E}_h + \theta q\mathbf{Q}$ és az $\mathbf{F}_h = \mathbf{E}_h - (1-\theta)q\mathbf{Q}$ $\mathbb{R}^{(N_x-1)\times(N_x-1)}$-beli tridiagonális mátrix jelöléseket, a (11.3.48) egyenletben a mátrix ekkor az

$$\tilde{\mathbf{L}}_{h,\tau} = \begin{pmatrix} \mathbf{D}_h & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{F}_h & \mathbf{D}_h & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & -\mathbf{F}_h & \mathbf{D}_h & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & -\mathbf{F}_h & \mathbf{D}_h \end{pmatrix} \tag{11.3.107}$$

N_t számú sorból álló hipermátrix. A $\mathbf{b}^n$ $(n = 1, 2, \dots, N_t)$ $(N_x - 1)$ dimenziós vektorok alakja a következő:

$$\mathbf{b}^1 = \tau\mathbf{f}_h^{1,\theta} + \mathbf{F}_h\boldsymbol{\mu}_0; \quad \mathbf{b}^n = \tau\mathbf{f}_h^{n,\theta}, \quad n = 2, 3, \dots, N_t. \tag{11.3.108}$$

Soronként kiírva a (11.3.48) egyenletet ekkor a következőt kapjuk:

$$\begin{aligned} \mathbf{D}_h\mathbf{y}^1 &= \mathbf{b}^1, \\ -\mathbf{F}_h\mathbf{y}^{n-1} + \mathbf{D}_h\mathbf{y}^n &= \mathbf{b}^n, \quad n = 2, 3, \dots, N_t, \end{aligned} \tag{11.3.109}$$

amelynek átrendezésével a következő megoldó algoritmust kapjuk:

$$\begin{aligned} \mathbf{D}_h\mathbf{y}^1 &= \mathbf{b}^1, \\ \mathbf{D}_h\mathbf{y}^n &= \mathbf{b}^n + \mathbf{F}\mathbf{y}^{n-1}, \quad n = 2, 3, \dots, N_t. \end{aligned} \tag{11.3.110}$$

A (11.3.110) algoritmus $\theta = 0$ esetén explicit, az összes többi esetben viszont implicit módszert jelent. Ugyanakkor a $\mathbf{D}_h$ mátrix speciális struktúrájának következtében az időrétegenkénti lineáris algebrai egyenletrendszer megoldására alkalmazhatjuk a korábban már ismertetett speciális Gauss-eliminációt.

Befejezésül vegyük észre, hogy az ebben a szakaszban alkalmazott diszkretizációs módszerek rendje a következő:

- a térbeli diszkretizáció a szokásos másodrendű véges differenciás közelítés volt, tehát $\mathcal{O}(h^2)$ pontosságú;
- az időbeli diszkretizáció a θ-módszer volt, amelyről megmutattuk a 9. fejezetben, hogy $\theta = 0.5$ esetén másodrendű, egyébként elsőrendű.

Az összetett módszer pontossága tehát ezen rendek összegéből adódik. Így magasabb rendű térbeli illetve időbeli approximációval magasabb rendben konzisztens módszerek állíthatók elő. Ugyanakkor az ilyen módszerek stabilitása (ami a konvergenciához szükséges) általában kérdéses, és ha igaz is, többnyire nehéz megmutatni.

Az elméleti rész befejezéseként megjegyezzük, hogy terjedelmoi okok miatt nem foglalkozunk a másodrendű hiperbolikus feladatok numerikus megoldási módszereivel. Ezek vizsgálatai módszere lényegében megegyezik a parabolikus feladatok módszerekkel, és így az Olvasó néhány meglévő szakirodalom tanulmányozásával (pl. [29, 23, 35]) önállóan is elvégezheti.

11.4. A parciális differenciálegyenletek numerikus megoldása MATLAB segítségével

Az előző szakaszban megadtuk a Laplace-egyenlet és a hővezetési egyenlet véges differenciás megoldási módszerét. Ezek az algoritmusok a MATLAB programrendszer segítségével megvalósíthatók. Készíthetünk önálló programokat, illetve alkalmazhatjuk a MATLAB programrendszer saját, már elkészített és beépített programját is. Ebben a szakaszban a két fenti feladatosztályra vonatkozó MATLAB programok ismertetésével foglalkozunk.

11.4.1. A Poisson-egyenlet megoldása első (Dirichlet-féle) peremfeltétellel

Tekintsük a

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = f(x,y) \tag{11.4.1}$$

egyenletet az inhomogén (első, avagy Dirichlet-féle) peremfeltétellel az egységnégyzeten. Numerikus megoldásként a véges differenciák 11.2.5. szakaszban ismertetett algorimusát alkalmazzuk.

Egy ekvidisztáns rácshálót generálunk, és mindegyik irányban ugyanannyi osztáspontot jelölünk ki.

Az elkészített POISSON11.M m-fájl egyetlen bemenő paramétere n, amely az irányonkénti osztásrészek száma. Ugyancsak egyetlen kimenő paramétere van, az *uapprox* vektor, amely a numerikus közelítést tartalmazza.
A programban belül a Poisson-egyenletre definiáljuk az f jobb oldali függvényt, illetve a peremfeltételeket leíró $ux0, ux1, u0y, u1y$ függvényeket, amelyek rendre az $y = 0$, $y = 1$, $x = 0$ és az $x = 1$ oldalakon megadott peremfeltételeket jelentik.

A rutin a következő:

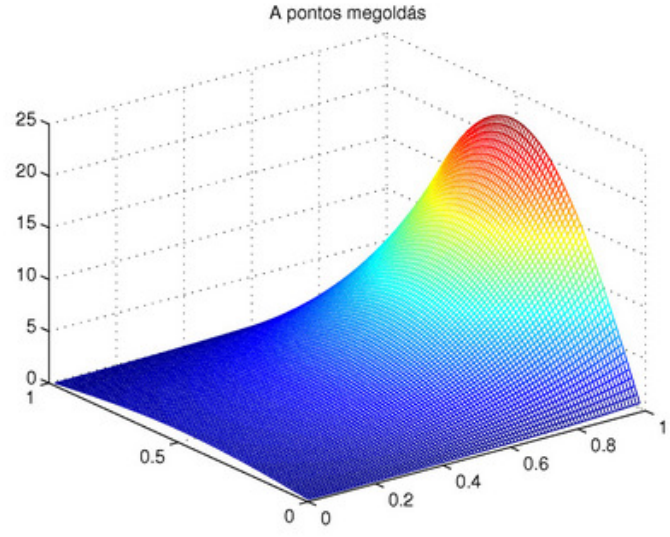
```
function[uapprox] = poisson11(n)
% a Poisson egyenlet megoldása egységnégyzeten
% inhomogén első peremfeltétellel
%
% n: az egyetlen bemenő adat, az egységnégyzet osztásrészeinek száma
% (mindkét irányban ugyanannyi)
% uapprox: az egyetlen kimenő paraméter, a numerikus megoldás a
% rácspontokban
%
% A feladatot definiáló függvények: jobb oldal: f
% peremfeltételek a négyzet odalain:(ux0,ux1,u0y,u1y)
% a pontos megoldás: u.
f = inline('x^2 + y^2');
ux0 = inline('0');
ux1 = inline('0.5*x^2');
u0y = inline('sin(pi*y)'); u1y = inline('exp(pi)*sin(pi*y) + 0.5*y^2');
u = inline('exp(pi*x)*sin(pi*y) + .5*(x^2)*(y^2)');
%
h = 1/n; N = (n-1)^2;
A = sparse(N,N); % az A mátrixot ritka mátrixként kezeljük
```

```
F = zeros(N,1);
% A továbbiakban felépítjük az A mátrixot és az F jobb oldali vektort
%
% A mátrix felépítése:
A = -4*sparse(eye(N,N));
for j=1:n-1,
for i=1:n-1,
k = (j-1)*(n-1)+i;
if j > 1, A(k,k-(n-1)) = 1;
end;
if j < n-1, A(k,k+(n-1)) = 1;
end;
if i > 1, A(k,k-1) = 1;
end;
if i < n-1, A(k,k+1) =1;
end; end;end;
A = (1/(h*h))*A;
%
% F vektor felépítése:
for j=1:n-1,
for i=1:n-1, k = (j-1)*(n-1)+i;
xi = i*h; yj = j*h;
F(k) = f(xi,yj);
if j==1, F(k) = F(k) - (1/(h*h))*ux0(xi);
end;
if j==n-1, F(k) = F(k) - (1/(h*h))*ux1(xi);
end;
if i==1,
F(k) = F(k) - (1/(h*h))*u0y(yj);
end;
if i==n-1, F(k) = F(k) - (1/(h*h))*u1y(yj);
end; end;end;
%
% Az Ax = F egyenlet megoldása
uapprox = A\F;
% A pontos megoldás kiszámolása
utrue = zeros(N,1);
for j=1:n-1,
for i=1:n-1,
k = (j-1)*(n-1)+i;
xi = i*h; yj = j*h;
utrue(k) = u(xi,yj);
end; end;
display('l_2 és a max. hiba')
err2 = h*norm(utrue-uapprox),
errinf = norm(utrue-uapprox,'inf')
% A pontos megoldás, a numerikus megoldás és a hiba kirajzolása
ugrid = reshape(utrue,n-1,n-1);
```

```
figure(1) mesh([h:h:(n-1)*h]',[h:h:(n-1)*h]',ugrid')
title('A pontos megoldás')
pause
apgrid = reshape(uapprox,n-1,n-1);
figure(12) mesh([h:h:(n-1)*h]',[h:h:(n-1)*h]',apgrid')
title('A közelítő megoldás')
pause
errgrid = reshape(utrue-uapprox,n-1,n-1);
figure(3)
mesh([h:h:(n-1)*h]',[h:h:(n-1)*h]',errgrid')
title('Hibafüggvény')
```

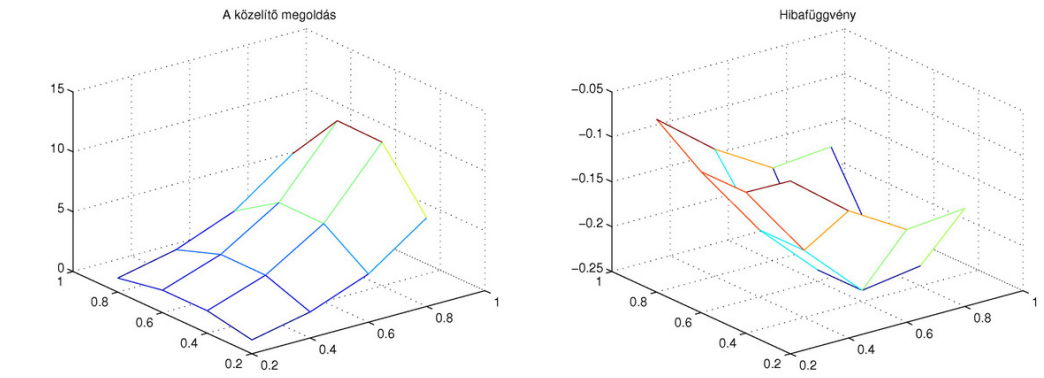
A rutin a pontos megoldás ismeretében a pontos megoldást, a közelítő megoldást és a hibát is kirajzolja. A véges differenciás lineáris algebrai egyenletrendszerben az együtthatómátrixot ritka mátrixként tároljuk, és a rendszert a MATLAB beépített (és általában optimális) $A\backslash F$ utasításával oldjuk meg.

Megjegyezzük, hogy a fenti programban a pontos megoldás az $u(x,y) = \exp(\pi x)\sin(\pi y) + 0.5x^2y^2$ függvény. Más feladatok esetén a programban a megfelelő függvénydefiníciókat értelemszerűen meg kell változtatnunk. Ha olyan feladatot oldunk meg, ahol a'priori nem ismerjük a megoldást, akkor a programból a megfelelő modulrészek (a pontos megoldás és a hibafüggvény ábrázolása illetve kiszámolása) kihagyásra kerül.

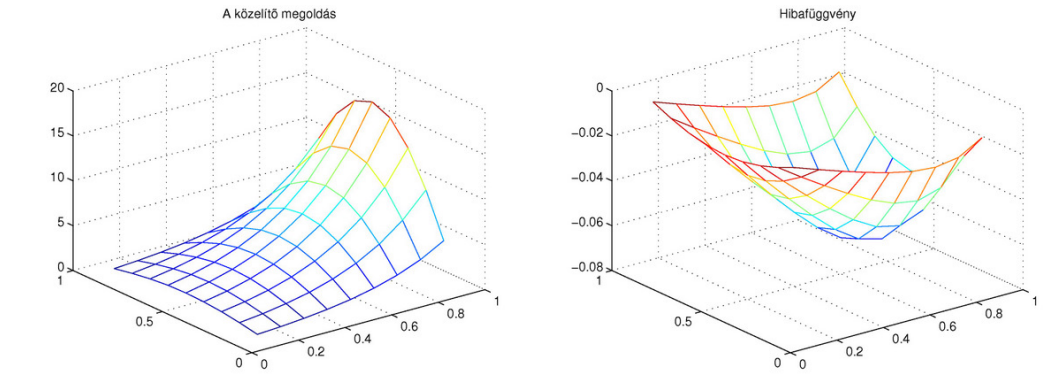


11.4.1. ábra: A Poisson-egyenlet $u(x,y) = \exp(\pi x)\sin(\pi y) + 0.5x^2y^2$ megoldása az egységnégyzeten.

A programban szereplő függvényekkel végeztünk néhány numerikus kísérletet. A pontos meg-



11.4.2. ábra: A numerikus megoldás és hibafüggvénye $n = 5$ esetén.



11.4.3. ábra: A numerikus megoldás és hibafüggvénye $n = 10$ esetén.

oldás a 11.4.1. ábrán látható.[9]

Különböző felosztásokra nézzük meg a hibákat!

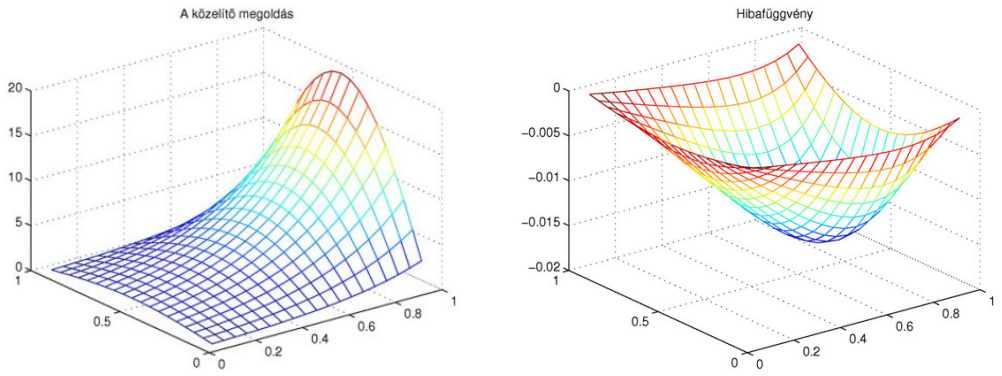
Az $n = 5$ esetre a numerikus megoldást és a hibafüggvényt a 11.4.2. ábra mutatja. Az $n = 10$ és $n = 20$ eseteket a 11.4.3. és a 11.4.4. ábrák mutatják.

Az $n = 80$ megválasztás esetén, mint a 11.4.5. ábrán is látható, a numerikus megoldás gyakorlatilag megegyezik a pontos megoldással.

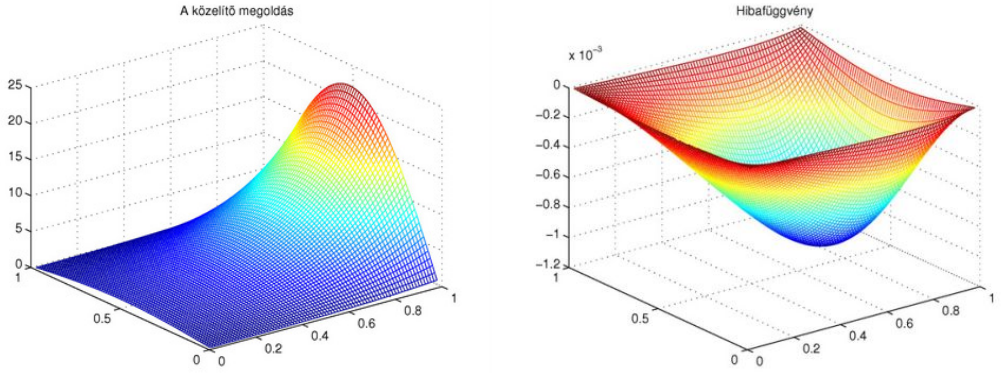
A 11.4.1. táblázatban összefoglaltuk a számításainkat, megadva a feleződő diszkretizációs lépésközök melletti maximumhibákat és a konvergenciarendet. (Mivel n értékét minden lépésben megkétszereztük, ezért a hibának negyedére kell csökkennie.) Mint az a harmadik sorból kiolvasható, a gyakorlatiban a hibák hányadosa ezzel lényegében megegyezik.

A módszer MATLAB realizálása során keletkezett számítási hibát jól jellemzi a következő példa. Legyen a Poisson-feladat megoldása az $u(x, y) = x + y$ lineáris függvény. (Ekkor a programban értelemszerűen $f = 0$, $ux0 = x$, $ux1 = x+1$, $u0y = y$ és $u1y = y+1$.) Ebben az esetben a módszer lokális approximációs hibája nulla. (Ugyanis a (11.2.52) szerinti $\Psi_h(x_i, y_j)$ kifejezésében

[9] Mivel a megoldást a peremen ismerjük, és ott a hibafüggvény nulla, ezért a továbbiakban többnyire csak a megoldási tartomány belsejében ábrázoljuk a hibát illetve a megoldásfüggvényt.



11.4.4. ábra: A numerikus megoldás és hibafüggvénye $n = 20$ esetén.



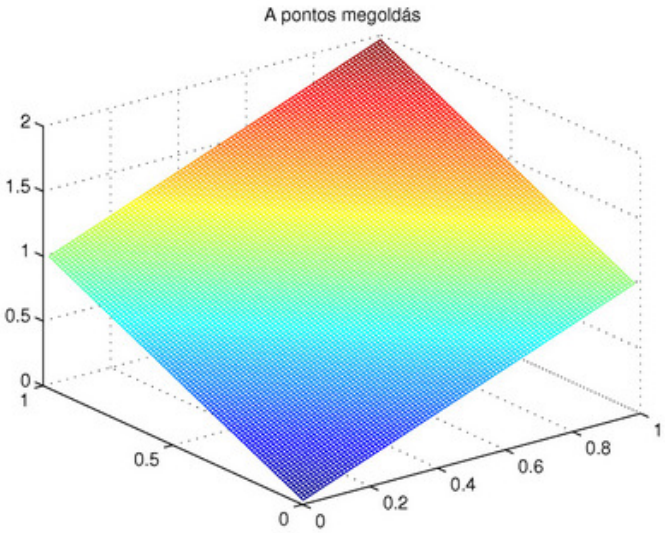
11.4.5. ábra: A numerikus megoldás és hibafüggvénye $n = 80$ esetén.

n	5	10	20	40	80
e_h	$2.3747e-001$	$6.6841e-002$	$1.6826e-002$	$4.2139e-003$	$1.0539e-003$
hányados		$2.8147e-001$	$2.5173e-001$	$2.5044e-001$	$2.5011e-001$

11.4.1. táblázat: A Poisson-feladat véges differenciás megoldásának hibája a maximumnormában és a hibák hányadosa.

a h^2-es vezető tag együtthatója $M_4/6$, és a mi esetünkben $M_4 = 0$. Mivel minden magasabb rendű derivált is eltűnik, ezért erre a példára a véges differenciás approximáció pontos.)

A fenti lineáris függvénnyel végeztünk néhány numerikus kísérletet. A pontos megoldás a 11.4.6. ábrán látható.

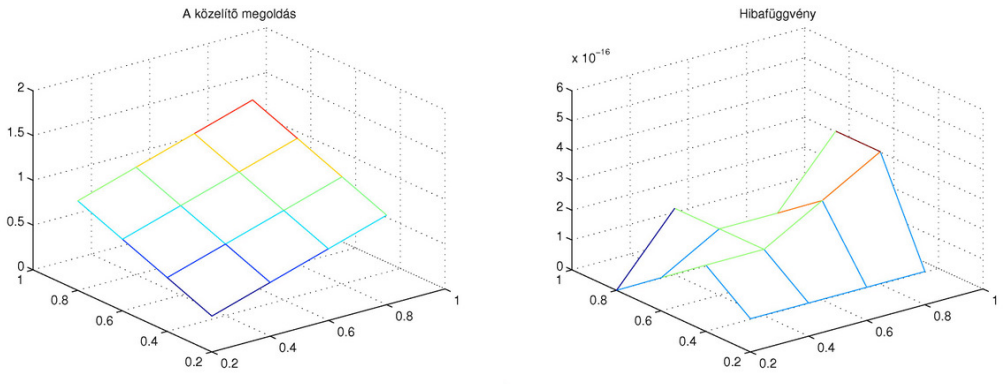


11.4.6. ábra: A Poisson-egyenlet $u(x, y) = x + y$ megoldása az egységnégyzeten.

A továbbiakban a különböző felosztásokra nézzük meg a hibákat!

Az $n = 5$ esetre a numerikus megoldást és a hibafüggvényt a 11.4.7. ábra mutatja.

A hibafüggvény nagyságából (10^{-16}) látható, hogy már ilyen, viszonylag durva felosztás esetén is a numerikus megoldás gyakorlatilag megegyezik a pontos megoldással, és hiba csak a gépi számítások miatt keletkezik. Ezért a rácsháló finomításával nem várható az eredmények további javulása, sőt, a műveletek számának növekedése miatt, azok romlása várható. A 11.4.2. táblá-



11.4.7. ábra: A numerikus megoldás és hibafüggvénye $n = 5$ esetén.

n	5	20	40	80	100
e_h	$4.4409e-016$	$4.4409e-016$	$6.6613e-016$	$2.2204e-015$	$1.8874e-015$

11.4.2. táblázat: A Poisson-feladat véges differenciás megoldásának hibája a maximumnormában az $u(x,y)=x+y$ megoldású tesztfeladatra.

zatban összegyűjtöttük a különböző lépésközökhöz tartozó maximumnorma-beli hibákat, és az eredmények alátámasztják ezt az elvárásunkat. (Értelemszerűen a hibahányadost nem szerepeltetjük a táblázatban.)

11.4.2. A hővezetési egyenlet megoldása véges differenciák módszerével

A hővezetési egyenletre viszonylag könnyen önállóan is elkészíthetünk MATLAB progamot. A továbbiakban megadjuk az explicit módszert realizáló HEATEXP.M nevű m-fájlt. Ez a program a

$$\frac{\partial u(x,t)}{\partial t}-\frac{\partial^2 u(x,t)}{\partial x^2}=0,\quad (x,t)\in(0,endx)\times(0,endt) \tag{11.4.2}$$

egyenletet oldja meg az

$$u(x,0)=init(x),\quad x\in(0,endx);\quad u(0,t)=bdry(1),\quad u(endx,t)=bdry(2),\quad t\in[0,endt] \tag{11.4.3}$$

kiegészítő feltételekkel.

Mindkét irányban egy-egy ekvidisztáns rácshálót generálunk, és ezen a rácshálón a (11.3.56) séma segítségével határozzuk meg a véges differenciás közelítéseket.

Az elkészített m-fájl bemenő paraméterei a következők:

- $endx$: a térbeli egydimenziós tartomány megadására szolgál, a térbeli változó a $[0,endx]$ intervallumon változik.
- $endt$: az időbeli tartomány megadására szolgál, az időbeli változó a $[0,endt]$ intervallumon változik.
- Nx: a diszkretizácó során a térbeli osztásrészek száma;
- q: a (11.3.37) képletben definiált τ/h^2 hányados. (Emlékeztetőül: a konvergencia feltétele a $q\leq 0.5$ feltétel volt, lásd a 11.3.5. tételt.)

A kezdeti feltételt a programon belül az *init* függvénnyel definiáljuk. A peremfeltételt a kétdimenziós *bdry* vektorral, ugyancsak a programon belül definiáljuk. (A két térbeli végpontban első peremfeltétel adott: $u(x(1),t)=bdry(1)$, $u(x(end),t)=bdry(2)$.)
A program kimenő paramétere a numerikus megoldást jelentő u.

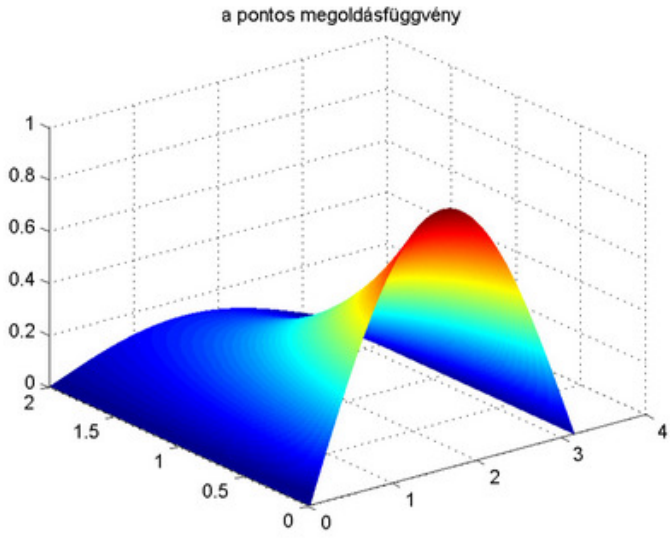
A rutin a következő:

```
function u = heatexp(endx, endt, Nx,q)
%
% Az egydimenziós hővezetési feladatot oldjuk meg
% A program bemenő adatai:
```

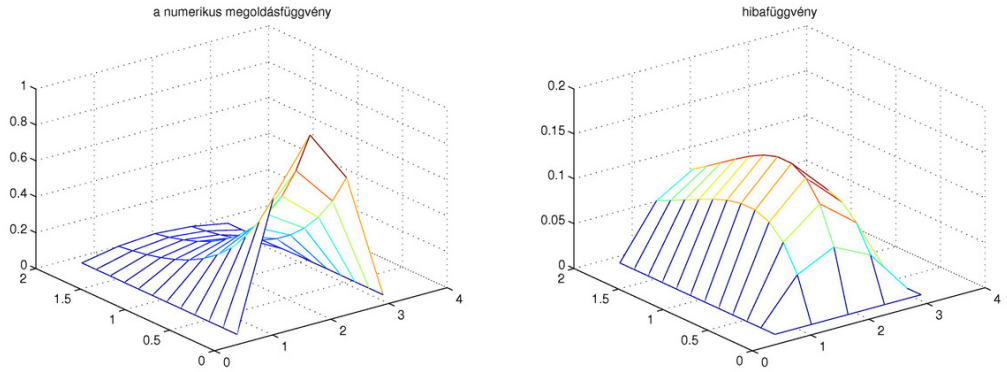
```
% endx: a térbeli változó [0,endx]intervallumon változik
% endt: az időváltozó [0,endt]intervallumon változik
% Nx: a térbeli osztásrészek száma
% q: az időbeli és a térbeli diszkretizációs lépésköz négyzetének hámyadosa
% (stabilitási feltétel: q ≤ 0.5)
% A két térbeli végpontban első peremfeltétel adott: u(x(1), t) = bdry(1), u(x(end), t) =
bdry(2).
% A kezdeti függvényt az init függvényben definiáljuk
%
x= linspace(0, endx, Nx);
dx = (endx/Nx) % a lépésközök definiálása
dt=(q*dx*dx)
%
Nt=fix(endt/dt) t = linspace(0, endt, Nt); s = dt/(dx * dx)
% ellenőrizzük a szabilitási feltételt
if s >0.5
'instabilitás!'
pause
end
%
init = sin(x); % a kezdeti és a peremfeltétel
bdry=[0 0];
%
J = length(t); N = length(x); u = zeros(N,J); u(:, 1) = init; for n = 1:J-1
u(2:N-1,n+1) = s*(u(3:N,n) + u(1:N-2,n)) + (1 - 2*s)*u(2:N-1, n);
u(1, n+1) = bdry(1);
u(N, n+1) = bdry(2);
end
% az eredmények ábrázolása
apprgrid = reshape(u,N,J);
figure(1) mesh([dx:dx:(N)*dx]',[dt:dt:(J)*dt]',apprgrid')
title('a numerikus megoldásfüggvény')
pause
% ha ismerjük a pontos megoldást, ide írjuk be!
upontos=zeros(N,J); for i=1:N
for n=1:J
upontos(i,n)=exp(-t(n))*sin(x(i));
end
end
% hibamatrix=upontos-apprgrid;
figure(3)
mesh([dx:dx:(N)*dx]',[dt:dt:(J)*dt]',hibamatrix')
title('hibafüggvény')
```

A programban szereplő függvényekkel végeztünk néhány numerikus kísérletet. A pontos megoldás a 11.4.8. ábrán látható.

Különböző felosztásokra nézzük meg a hibákat! Legyen mindvégig $q = 0.4$!



11.4.8. ábra: A hővezetési egyenlet $u(x,t) = e^{-t}sinx$ megoldása a $(0,\pi) \times (0,2)$ tartományon.



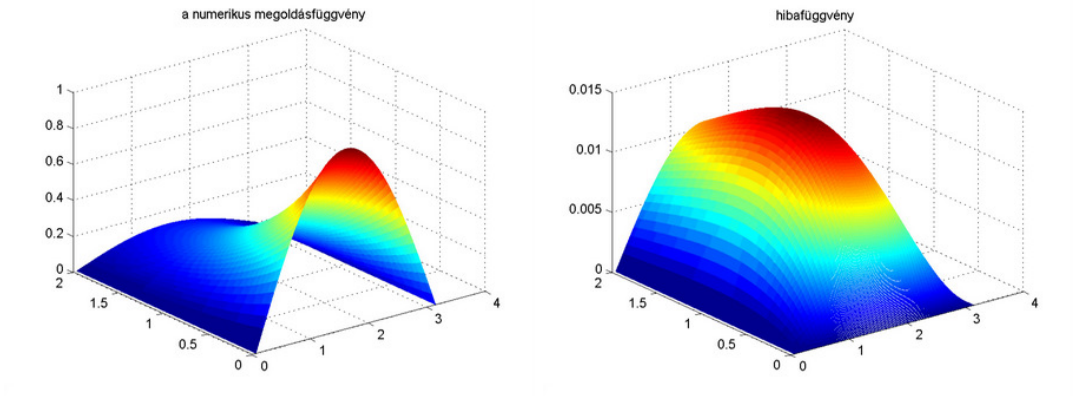
11.4.9. ábra: A numerikus megoldás és hibafüggvénye $Nx = 5$ esetén.

Az $Nx = 5$ esetre a numerikus megoldást és a hibafüggvényt a 11.4.9. ábra mutatja. Az $Nx = 50$ és az $Nx = 100$ eseteket a 11.4.10. és 11.4.11. ábrák mutatják.

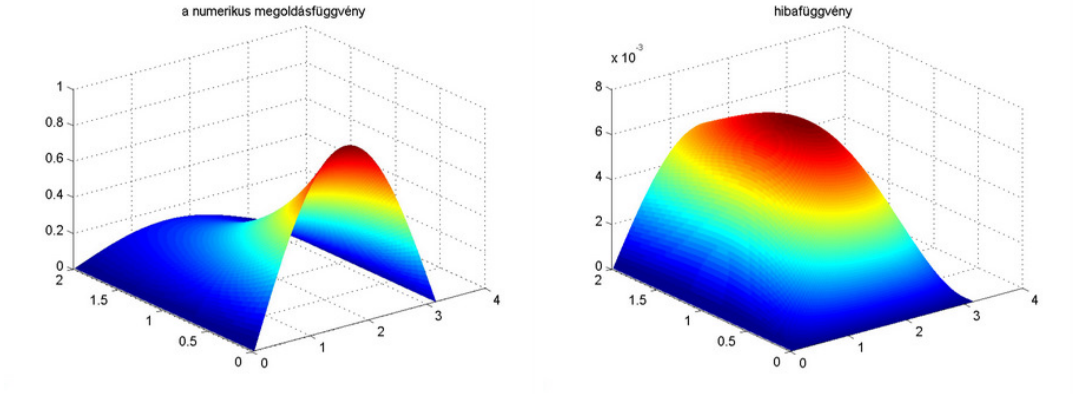
A 11.4.3. táblázatban összefoglaltuk a számításainkat, megadva a csökkenő diszkretizációs lépésközök melletti maximumhibákat.

Vizsgáljuk meg, mit jelent ugyanezen tesztfeladatra a $q \leq 0.5$ stabilitási feltétel megsértése! Legyen mindvégig $Nx = 50$. A $q = 0.5$ megválasztás esetén jó közelítést kapunk, lásd a 11.4.12. ábrát). Azonban q nagyobb értékei mellett az eredmények használhatatlanok! A $q = 0.51$ megválasztás esetén még ugyan korlátos marad a numerikus megoldás (lásd a 11.4.13. ábrát), de $q = 0.52$ esetén már gyakorlatilag felrobban a rendszer (lásd a 11.4.14. ábrát).

A 11.4.4. táblázatban összefoglaltuk $Nx = 50$ mellett a kritikus $q = 0.5$ értékhez közeli megválasztásokhoz tartozó maximumhibákat.



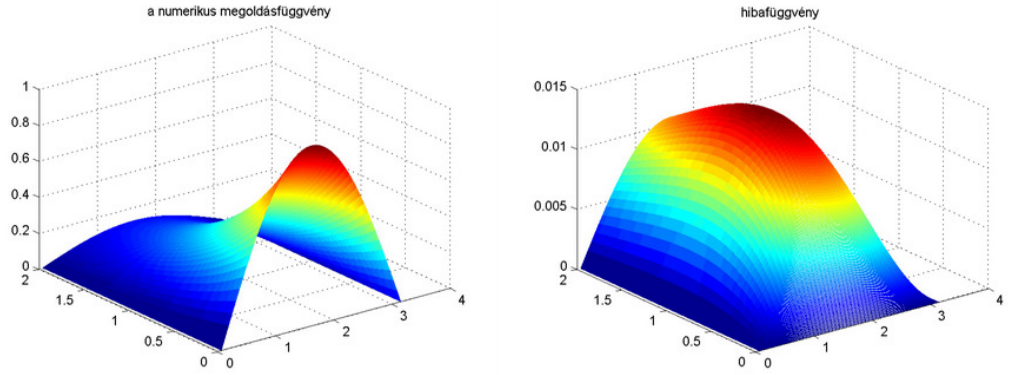
11.4.10. ábra: A numerikus megoldás és hibafüggvénye $Nx = 50$ esetén.



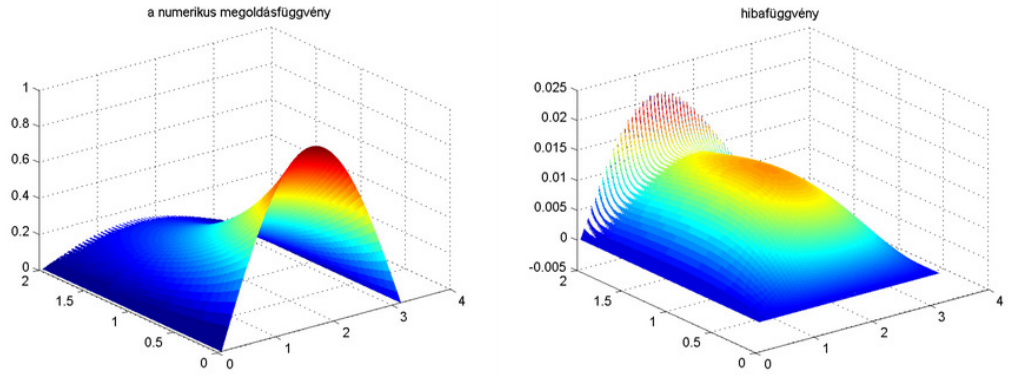
11.4.11. ábra: A numerikus megoldás és hibafüggvénye $Nx = 100$ esetén.

Nx	e_h
10	$6.9396e-002$
25	$2.8868e-002$
50	$1.4592e-002$
100	$7.3599e-003$
150	$4.8875e-003$
200	$3.6761e-003$
250	$2.9340e-003$
300	$2.4490e-003$
350	$2.1014e-003$
400	$1.8355e-003$

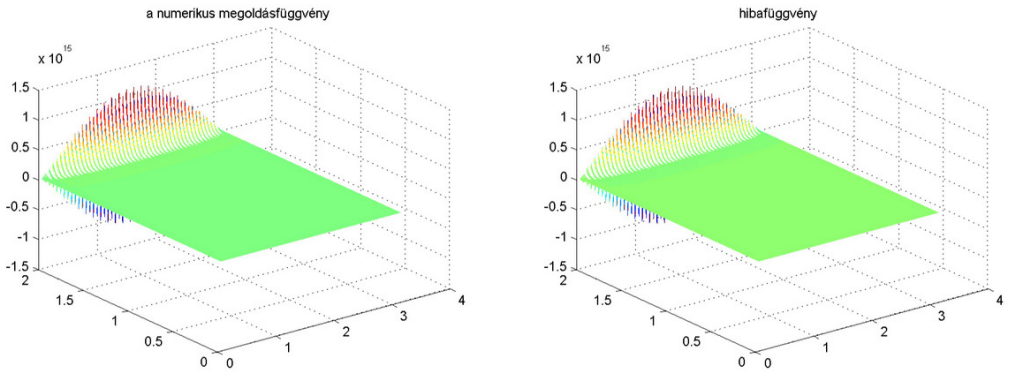
11.4.3. táblázat: A hővezetési feladat véges differenciás megoldásának hibája a maximumnormában.



11.4.12. ábra: A numerikus megoldás és hibafüggvénye $q = 0.5$ esetén.



11.4.13. ábra: A numerikus megoldás és hibafüggvénye $q = 0.51$ esetén.

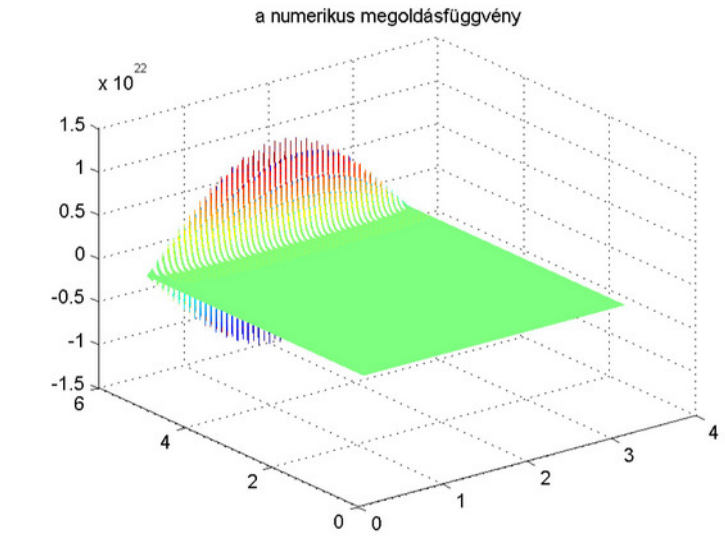


11.4.14. ábra: A numerikus megoldás és hibafüggvénye $q = 0.52$ esetén.

q	e_h
0.48	$1.4595e-002$
0.49	$1.4428e-002$
0.50	$1.4668e-002$
0.51	$2.1019e-002$
0.52	$1.1172e+015$
0.53	$1.1020e+029$
0.54	$2.3063e+042$
0.55	$4.1650e+054$
0.56	$4.7364e+066$
0.57	$4.9813e+077$

11.4.4. táblázat: A hővezetési feladat véges differenciás megoldásának hibája q függvényében, rögzített $Nx = 50$ esetén.

Megjegyezzük, hogy $q = 0.51$ esetén a hiba még elfogadhatónak látszik. Ezért azt gondolhatnánk, hogy ez a megválasztás is megengedhető. Az ezen esethez tartozó ábrát nézve (11.4.13. ábra) azt látjuk, hogy a numerikus megoldás az időintervallum végén ($t = 2$ körül) kezd elromlani. Tehát elképzelhető, hogy az időintervallum növelésével tovább romlik a megoldás. Ezért oldjuk meg a feladatot ugyanezen numerikus módszerrel ($Nx = 50$ és $q = 0.51$) a $[0, 5]$ időintervallumon, azaz legyen most $endt = 2$ helyett $endt = 5$! Sejtésünk beigazolódik: ahogy az a 11.4.15. ábrán is látható, a hiba gyakorlatilag nem marad korlátos.



11.4.15. ábra: A hővezetési feladat véges differenciás megoldása a $(0, \pi) \times (0, 5)$ tartományon $q = 0.51$ és $Nx = 50$ paraméterek esetén.

Befejezésül néhány szó a MATLAB programrendszer saját megoldó programjáról. A legálta-

lánosabb a PDEPE.M nevű program. Ez a program numerikusan megoldja a

$$c\left(x,t,u,\frac{\partial u}{\partial x}\right)\frac{\partial u}{\partial t}=x^{-m}\frac{\partial u}{\partial x}\left(x^m f(x,t,u,\frac{\partial u}{\partial x})\right)+s\left(x,t,u,\frac{\partial u}{\partial x}\right) \tag{11.4.4}$$

egyenletet, ahol c, f, s adott függvények. A rutin meghívása a
PDEPE(m,pdefun,icfun,bcfun,xmesh,tspan)
utasítással lehetséges, ahol az egyes bemenő paraméterek a következők.

- m a feladat térbeli tartományának szimmetriáját jellemző paraméter.
- *pdefun*: a parciális differenciálegyenletben szereplő függvények megadására szolgál.
- *icfun*: a kezdeti feltételt leíró függvény.
- *bcfun*: a peremfeltételeket leíró függvény.
- *xmesh*: megadjuk a térbeli változó szerinti rácspontokat az $[x0, x1, ..., xn]$ vektor segítségével.
- *tspan*: megadjuk az időbeli változó szerinti rácspontokat a $[t0, t1, ..., tf]$ vektor segítségével.

(Az *xmesh* és *tspan* vektorok minimális mérete 3.) A program alkalmazásának részletei megtanulhatók a MATLAB "help" funkciójának felhasználásával, továbbá számos más könyvből. A MATLAB néhány kidolgozott és a "help" funkciójában a forráskódját is ismertető programmal rendelkezik (PDEX1,PDEX2,PDEX3,PDEX4,PDEX5). Az érdeklődöknek javasoljuk ezek tanulmányozását és megismerését.

Javasoljuk a további MATLAB programokkal megismerkedni szándékozóknak a következő linkeket:
http://www.math.umd.edu/undergraduate/schol/matlab/pde.html,
http://people.sc.fsu.edu/~jburkardt/f_src/fd1d_heat_implicit/fd1d_heat_implicit.html
Megjegyezzük, hogy más típusú feltételek is megadhatók a feladat kiegészítő feltételeként. Ilyenek az ún. *periodikus feltételek*. Ezek nagy gyakorlati jelentőséggel rendelkeznek, ezért vizsgálatuk fontos. Az ilyen feltételek melletti hővezetési egyenletet megoldó MATLAB programok, kidolgozott példákkal megtalálhatók az alábbi linken:
http://www.math.toronto.edu/mpugh/Teaching/SamplePrograms/heat.html

11.5. Feladatok

Parciális differenciálegyenletek

11.5.1. feladat. Határozzuk meg, hogy $\mathbb{R}^2$ egyes részein milyen típusú az alábbi egyenlet:

$$x\frac{\partial^2 u(x,y)}{\partial x^2}+y\frac{\partial^2 u(x,y)}{\partial y^2}=0.$$

11.5.2. feladat. Határozzuk meg, hogy a folytonos a, b és c együtthatófüggvények milyen tulajdonságai mellett lesz elliptikus, parabolikus és hiperbolikus típusú az alábbi egyenlet:

$$a(x,y)\frac{\partial^2 u(x,y)}{\partial x^2}+2b(x,y)\frac{\partial^2 u(x,y)}{\partial x\partial y}+c(x,y)\frac{\partial^2 u(x,y)}{\partial y^2}=0!$$

11.5.3. feladat. Folytonos a, b és c függvények esetén érintkezhet-e egymással az elliptikus és hiperbolikus típusú pontok $\mathbb{R}^2$-bel halmaza az

$$(L_0u)(x,y) = a(x,y)\frac{\partial^2 u(x,y)}{\partial x^2} + 2b(x,y)\frac{\partial^2 u(x,y)}{\partial x \partial y} + c(x,y)\frac{\partial^2 u(x,y)}{\partial y^2} \tag{11.5.1}$$

operátorra?

11.5.4. feladat. Határozzuk meg a

$$\frac{\partial^2 u(x,t)}{\partial t^2} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \quad x \in \mathbb{R},\ t > 0$$

egyenlet megoldását az

$$u(x,0) = \varphi(x), \quad x \in \mathbb{R}$$
$$\frac{\partial u(x,0)}{\partial t} = \psi(x), \quad x \in \mathbb{R}$$

kiegészítő feltételek mellett!
(Útmutatás: a $\xi = x+t$, $\vartheta = x-t$ új változók bevezetésével transzformáljuk át az egyenletünket!)

11.5.5. feladat. Mutassuk meg, hogy a

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 0. \quad x \in \mathbb{R},\ y \geq 0$$

egyenlet

$$u(x,0) = 0, \quad x \in \mathbb{R}$$
$$\frac{\partial u(x,0)}{\partial t} = 0, \quad x \in \mathbb{R}$$

feltételek megadása esetén instabil feladat!
(Útmutatás: Használjuk fel, hogy tetszőleges $n \in \mathbb{N}$ esetén az

$$\frac{\partial^2 u_n(x,y)}{\partial x^2} + \frac{\partial^2 u_n(x,y)}{\partial y^2} = 0. \quad x \in \mathbb{R},\ y \geq 0,$$

$$u_n(x,0) = 0, \quad x \in \mathbb{R}$$
$$\frac{\partial u_n(x,0)}{\partial t} = \frac{1}{n}\cos(nx) \quad x \in \mathbb{R}$$

feladat megoldása az $u_n(x,y) = \frac{1}{n^2}\cos(nx)\sinh(ny)$ függvény!)

11.5.6. feladat. Oldjuk meg a változók szétválasztásának módszerével a hővezetési egyenletet a $u(x,0) = u_0(x)$ kezdeti feltétel és a $\dfrac{\partial u(0,t)}{\partial x} = u(l,t) = 0$ peremfeltétel esetén.

11.5.7. feladat. Írjuk fel az (x,y,z) változókra a Laplace-egyenletet $\mathbb{R}^3$-ban! Állítsuk elő a homogén, első peremfeltételű feladat megoldását a változók szétválasztásának módszerével.

11.5.8. feladat. Bizonyítsuk be a maximumelvet a Laplace-egyenletre $\mathbb{R}^2$-ben!

11.5.9. feladat. Bizonyítsuk be a maximumelvet a hővezetési egyenletre!

Elliptikus feladatok numerikus megoldása véges differenciákkal

11.5.10. feladat. Tekintsük az egységnégyzeten a

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = e^y(x^2+2)$$

egyenletet az $u(x,0) = x^2$, $u(x,1) = ex^2$, $u(0,y) = 0$, $u(1,y) = e^y$ peremfeltétellel. Írjuk fel a feladat véges differenciás approximációját jelentő lineáris algebrai egyenletrendszert, amikor mindkét irányban $N_x = N_y = 3$ osztásrészt veszünk! Oldjuk meg a rendszert és hasonlítsuk össze a numerikus megoldást az $u(x,y) = x^2e^y$ pontos megoldással.

11.5.11. feladat. Tekintsük az előző feladatot a

$$\frac{\partial u(x,0)}{\partial x} = 2x,\ u(x,1) = ex^2,\ u(0,y) = 0,\ u(1,y) = e^y$$

peremfeltétellel. Írjuk fel a feladat véges differenciás approximációját jelentő lineáris algebrai egyenletrendszert, amikor mindkét irányban $N_x = N_y = 3$ osztásrészt veszünk! Oldjuk meg a rendszert, és hasonlítsuk össze a numerikus megoldást az $u(x,y) = x^2e^y$ pontos megoldással!

11.5.12. feladat. Tekintsük az egységnégyzeten a

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 4$$

egyenletet az $u(x,0) = x^2$, $u(x,1) = x^2+1$, $u(0,y) = y^2$, $u(1,y) = y^2+1$ peremfeltétellel! Írjuk fel a feladat véges differenciás approximációját jelentő lineáris algebrai egyenletrendszert, amikor mindkét irányban $N_x = N_y = 5$ osztásrészt veszünk! Oldjuk meg a rendszert, és hasonlítsuk össze a numerikus megoldást az $u(x,y) = x^2+y^2$ pontos megoldással. Elemezzük a kapott pontosságot!

11.5.13. feladat. Tekintsük az egységnégyzeten a

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} - x\frac{\partial u(x,y)}{\partial x} - y\frac{\partial u(x,y)}{\partial y} + 2u(x,y) = 4$$

egyenletet az $u(x,0) = x^2$, $u(x,1) = x^2+1$, $u(0,y) = y^2$, $u(1,y) = y^2+1$ peremfeltétellel. Írjuk fel a feladat véges differenciás approximációját jelentő lineáris algebrai egyenletrendszert, amikor mindkét irányban $N_x = N_y = 3$ osztásrészt veszünk! Oldjuk meg a rendszert, és hasonlítsuk össze a numerikus megoldást az $u(x,y) = x^2+y^2$ pontos megoldással!

11.5.14. feladat. Oldjuk meg a 11.5.10. feladatot a POISSON11.M m-fájl segítségével! Készítsünk táblázatot, amely az osztásrészek száma és a maximumnormabeli pontosság közötti kapcsolatot mutatja! Ábrázoljuk a MATLAB segítségével ezt a kapcsolatot!

11.5.15. feladat. Írjuk át a POISSON11.M m-fájlt arra az esetre, amikor a megoldandó egyenlet

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} + cu = f(x,y)$$

alakú!

11.5.16. feladat. Írjuk át a POISSON11.M m-fájlt arra az esetre, amikor az $x = 0$ oldal mentén második peremfeltétel van adva! Az új program segítségével oldjuk meg a 11.5.11. feladatot!

11.5.17. feladat. Írjuk fel az egységkörön kitűzött Laplace-egyenlet véges differenciás megoldását az első peremfeltétel esetén!

Parabolikus feladatok numerikus megoldása véges differenciákkal

11.5.18. feladat. Tekintsük a $(0,1)\times(0,1)$ tartományon a

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \quad (x,t)\in(0,1)\times(0,1]$$

egyenletet az $u(x,0)=e^x, \quad x\in[0,1]$ kezdeti és az $u(0,t)=e^t, \quad u(1,t)=e^{1+t}, \quad t\in(0,1]$ peremfeltétellel. Írjuk fel a feladat explicit Euler-módszeres véges differenciás approximációját jelentő lineáris algebrai egyenletrendszerét, amikor az x irányban $N_x=3$ osztásrészt veszünk! Milyen időbeli felosztást választhatunk? Oldjuk meg a rendszert, és hasonlítsuk össze a numerikus megoldást az $u(x,t)=e^{x+t}$ pontos megoldással.

11.5.19. feladat. Írjuk fel a 11.5.18. feladat megoldását implicit Euler-módszeres véges differenciás módszerrel! Legyen ismét $N_x=3$. Milyen időbeli felosztást választhatunk? Oldjuk meg a rendszert, és hasonlítsuk össze a numerikus megoldást az $u(x,t)=e^{x+t}$ pontos megoldással!

11.5.20. feladat. Módosítsuk az HEATEXP.M nevű m-fájlt úgy, hogy explicit Euler-módszerrel állítsa elő a

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = f(x,t), \quad (x,t)\in(0,endx)\times(0,endt)$$

egyenlet megoldását a

$$u(x,0)=init(x), \quad x\in(0,endx); \quad u(0,t)=bdry(1), \quad u(endx,t)=bdry(2), \quad t\in[0,endt]$$

kiegészítő feltételekkel! Teszteljük programunkat a $(0,1)\times(0,1)$ tartományon a

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = xe^t, \quad (x,t)\in(0,1)\times(0,1]$$

egyenletű, az $u(x,0)=x, \quad x\in[0,1]$ kezdeti és az $u(0,t)=0, \quad u(1,t)=e^t, \quad t\in(0,1]$ peremfeltételű feladaton, amelynek a pontos megoldása $u(x,t)=xe^t$!

11.5.21. feladat. Készítsünk egy HEATIMP.M nevű m-fájlt, amely az implicit Euler-módszerrel állítja elő az

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \quad (x,t)\in(0,endx)\times(0,endt)$$

egyenlet megoldását a $u(x,0)=init(x), \quad x\in(0,endx)$ kezdeti feltétellel és az

$$u(0,t)=bdry(1), \quad u(endx,t)=bdry(2), \quad t\in[0,endt]$$

peremfeltételekkel! Teszteljük programunkat a 11.5.18. feladaton!

11.5.22. feladat. Készítsünk egy HEATCN.M nevű m-fájlt, amely a Crank–Nicolson-módszerrel állítja elő a

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \quad (x,t)\in(0,endx)\times(0,endt)$$

egyenlet megoldását az $u(x,0)=init(x), \quad x\in(0,endx)$ kezdeti feltétellel és az

$$u(0,t)=bdry(1), \quad u(endx,t)=bdry(2), \quad t\in[0,endt]$$

peremfeltételekkel! Teszteljük programunkat a 11.5.18. feladaton!

11.5.23. feladat. Készítsünk egy HEATTHETA.M nevű m-fájlt, amely a θ-módszerrel állítja elő a

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \quad (x,t) \in (0, endx) \times (0, endt)$$

egyenlet megoldását az $u(x,0) = init(x), \quad x \in (0, endx)$ kezdeti feltétellel és az

$$u(0,t) = bdry(1), \quad u(endx,t) = bdry(2), \quad t \in [0, endt]$$

peremfeltételekkel! Teszteljük programunkat a 11.5.18. feladaton! Vizsgáljuk meg a módszer pontosságát a $\theta \in [0.47, 0.53]$ értékekre, θ értékét századonként változtatva!

11.5.24. feladat. Tekintsük a

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = f(x,t), \quad (x,t) \in (0, endx) \times (0, endt)$$

egyenletet az $u(x,0) = init(x), \ x \in (0, endx)$ kezdeti és a

$$\frac{\partial u(0,t)}{\partial x} = bdry(1), \quad u(endx,t) = bdry(2), \quad t \in [0, endt]$$

peremfeltételekkel! Írjuk fel az explicit Euler-módszeres megoldást erre a feladatra!

11.5.25. feladat. Vizsgáljuk meg az explicit Euler-módszer konvergenciáját változó hosszúságú idő- és térbeli felosztásokra! Általánosítsuk vizsgálatunkat a θ-sémára!

Ellenőrző kérdések

1. Mit nevezünk parciális differenciálegyenletnek?
2. Mi a szerepe a kiegészítő feltételeknek? Mi a különbség a kezdeti feltétel és a peremfeltétel között?
3. Hogyan lehet operátoregyenletként felírni a Laplace-egyenletet illetve a hővezetési egyenletet a különböző kiegészítő feltételekkel?
4. Adja meg az explicit Euler-módszer algoritmusát!
5. Mutassa meg, hogy az explicit Euler-módszer konvergens!
6. Hogyan írható fel operátoregyenletként az explicit Euler-módszer?
7. Mikor nevezünk egy numerikus módszert konvergensnek? Mi a konzisztencia és a stabilitás? Mi a kapcsolat közöttük?
8. Mi a kapcsolat az M-mátrixok és a numerikus sémák stabilitása között?
9. Hogyan mutatható meg a véges differenciás módszer konvergenciája a Laplace-egyenletre?
10. Mutassa meg, hogy az implicit Euler-módszer konvergens!
11. Mutassa meg, hogy a θ-módszer konvergens!
12. Ismertessük az explicit Euler-módszer előnyeit és hátrányait!

13. Ismertessük az implicit Euler-módszer előnyeit és hátrányait!
14. Ismertessük az Crank–Nicolson-módszer előnyeit és hátrányait!
15. Milyen módszerek esetén szükséges lineáris algebrai egyenletrendszert megoldani a hővezetési feladat numerikus megoldásánál? Milyen tulajdonságú a rendszer mátrixa?
16. Mi a kapcsolat a rácsegyenletek és a lineáris algebrai egyenletrendszerek között?
17. Milyen MATLAB programokat ismer az elliptikus feladatok megoldására?
18. Milyen MATLAB programokat ismer a parabolikus feladatok megoldására?

Tárgymutató

Irodalomjegyzék

[1] Asher, U., Petzold, L. Computer Methods for Ordinary Differential Equations. SIAM, Philadelphia, 1998.

[2] Benoit, Note sur une méthode de resolution des équations normales provenant de l'application de la méthode des moindres carrés a un systeme d'équations lineaires en nombre inferieure a celui des inconnues, Application de la méthode a la resolution d'un systeme defini d'équations lineaires (Procede du Commandant Cholesky), Bulletin géodesique,1924.

[3] Serge Bernstein, Quelques remarques sur l'interpolation, Math. Ann. 79 (1918), 1–12.

[4] Jean-Paul Berrut, Lloyd N. Trefethen, Barycentric Lagrange interpolation, SIAM Review 46(4) (2004) 501–517.

[5] Bruce, G.H., Peaceman, D.W., Rachford, H.H., and Rice, J.D., Trans. Am. Inst. Min. Engrs (Petrol Div.) 198(79) (1953).

[6] Kurt Bryan, Tanya Leise, The $25,000,000,000 eigenvector – The linear algebra behind Google, http://www.rose-hulman.edu/~bryan/googleFinalVersionFixed.pdf.

[7] Coddington, E. A., Levinson, N. Theory of differential equations. New York, McGraw-Hill, 1955.

[8] Cooper, Jeffery M. Introduction to Partial Differential Equations with MATLAB (in Series: Applied and Numerical Harmonic Analysis), Birkhäuser, Boston, 1998.

[9] Czách L, Simon L. Parciális differenciálegyenletek, ELTE egyetemi jegyzet, Tankönyvkiadó, Budapest, 1980.

[10] J. W. Daniel, The conjugate gradient method for linear and nonlinear operator equations, SIAM J. Numer. Anal., 4 (1967), pp. 10–26.

[11] Erdős P., Vértesi P.: On the almost everywhere divergence of Lagrange interpolation polynomials for arbitrary systems of nodes, Acta Math. Acad. Sci. Hungar. 36 (1980), 71–89.

[12] Galántai A., Jeney A.: Numerikus módszerek, Miskolci Egyetemi Kiadó, 2002.

[13] David Goldberg, What Every Computer Scientist Should Know About Floating-Point Arithmetic, Computing Surveys, 1991.
http://docs.sun.com/source/806-3568/ncg_goldberg.html.

[14] Jacques Hadamard, Sur les problemes aux dérivées partielles et leur signification physique. Princeton University Bulletin, (1902), 49–52.

[15] P. Henrici, Numerikus módszerek, Műszaki Könyvkiadó, Budapest, 1985.

[16] Steve Hollasch, IEEE Standard 754 Floating Point Numbers.
http://steve.hollasch.net/cgindex/coding/ieeefloat.html

[17] IEEE-754 Floating-Point Conversion.
http://babbage.cs.qc.cuny.edu/IEEE-754/Decimal.html.

[18] Gene H. Golub, Charles van Loan, Matrix Computations, The Johns Hopkins University Press 1996.

[19] W. Kahan, Gauss–Seidel methods of solving large systems of linear equations, Doctoral Thesis, University of Toronto, Toronto, Canada, 1958.

[20] Kincaid, D, Cheney, W. Numerical Analysis. Mathematics of Scientific Computing,American Mathematical Society, 2009.

[21] Knabner, P., Angermann, L. Numerical Methods for Elliptic and Parabolic Partial Differential Equations, Texts in Applied Mathematics 44, Springer, Ney-York, 2003.

[22] M. M. Lavrent'ev, V. G. Romanov, S. P. Shishatski', Ill-posed problems of mathematical physics and analysis, translated by J. R. Schulenberger, Translations of Mathematical Monographs, vol. 64, American Mathematical Society, Providence, R. I., 1986.

[23] LeVeque, Randall J. Finite Difference Methods for Ordinary and Partial Differential Equations. Steady State and Time Dependent Problems, SIAM, Philadelphia,, 2007.

[24] J. Marcinkiewicz, Sur l'interpolation, Studia Math., 6 (1936), pp. 1-17.

[25] Cleve Moler, Numerical Computing with MATLAB, SIAM, 2004
(http://www.mathworks.com/moler/chapters.html).

[26] A.M. Ostrowski, On the linear iteration procedures for symmetiric matrices, Rend. Mat. Appl. 14 (1954) 140–163.

[27] E. Reich, On the convergence of the classical iterative method of solving linear simultaneous equations, Ann. Math. Statist. 20 (1949) 448–451.

[28] C. Runge, Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten, Zietschrift für Mathematik und Physik, 46 (1901), 224–243.

[29] Quarteroni, Sacco, Saleri, Numerical Mathematics, Springer, 2000 (újabb kiadás 2007-ben).

[30] Simon L., Baderko E. A. Másodrendű lineáris parciális differenciálegyenletek, Tankönyvkiadó, Budapest, 1983.

[31] Simon P., Tóth J. Differenciálegyenletek. Bevezetés az elméletben és az alkalmazásokba. TypoTech, Budapest, 2005.

[32] Stephenson, G. Partial Differential Equations for Scientists and Engineers, Longman, London-New-York, 1986.

[33] Stoyan Gisbert, Matlab - frissített kiadás, Typotex Kiadó, 2005.

[34] Stoyan G., Takó G. Numerikus módszerek 2., TypoTeX, Budapest, 1995.

[35] Stoyan Gisbert, Takó Galina: Numerikus módszerek 3., TypoTex Kiadó, Budapest, 1997.

[36] L.H. Thomas, Elliptic problems in linear difference equations over a network, Watson Sci. Comput. Lab. Rept., Columbia University, New York, 1949.

[37] Horst Zuse, The Life and Work of Konrad Zuse. http://www.epemag.com/zuse/.

[38] Joseph L. Zachary, Floating-Point Number Tutorial.
http//www.cs.utah.edu/~zachary/isp/applets/FP/FP.html

[39] K. Weierstrass, Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen, Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin, Erste Mitteilung (1885) 633–639.