Parciális differenciálegyenletek numerikus módszerei számítógépes alkalmazásokkal

Horváth Róbert, Izsák Ferenc, Karátson János

2013. február 5.

Tartalomjegyzék

Bevezetés

A parciális differenciálegyenletek a természettudományok által vizsgált folyamatok matematikai leírásában az egyik legalapvetőbb modellt jelentik. Néhány speciális esettől eltekintve ezek explicit képlettel való megoldása nem lehetséges, ezért közelítő, vagyis numerikus módszert kell alkalmaznunk. Ebben a jegyzetben áttekintünk néhány gyakran használatos ilyen eljárást.

Jegyzetünket elsősorban az ELTE, SZTE, DE, PTE és BME matematikus és alkalmazott matematikus mesterszakos hallgatói számára írtuk, akik tantervének része ez a témakör.

Érdemesnek látjuk konkretizálni a jegyzet megjelentetésének célját annak tükrében, hogy a témakörhöz kapcsolódóan természetesen létezik elérhető más magyar nyelvű anyag, elsősorban az igen tág területet átölelő, kézikönyvként is használatos [27] szakkönyv. Fő törekvéseink a következők voltak:

- Célzottan a mesterképzés anyagát összefoglalni, hogy a felkészüléshez közvetlenül használhassák a hallgatók. A jegyzet felépítése megegyezik az ELTE alkalmazott matematikus mesterszakán a két ide tartozó tárgy tematikájával.
- Az érthetőség kedvéért több helyen részletesebben tárgyalni egy-egy konkrétabb esetet és rövidebben foglalkozni az általánosabb helyzetekkel, ezt azonban oly módon, hogy a hallgatók megismerkedhessenek a mélyebb elméleti megalapozással is.
- Az elektronikus megjelenéshez kapcsolódóan animációkkal is szemléltetni egyes módszerek viselkedését.

A vizsgált feladattípusok elsősorban lineáris parciális differenciálegyenletek, röviden érintjük csak nemlineáris egyenletek néhány esetét. A jegyzet elliptikus részét Karátson János, időfüggő részét Izsák Ferenc írta (ELTE, Matematikai Intézet), az ábrákat és a számítógépes alkalmazásokat pedig Horváth Róbert (BME, Matematikai Intézet) készítette. Az elliptikus részben a végeselem-módszer, az időfüggő részben pedig a véges differenciák módszere kap nagyobb hangsúlyt.

I. rész

Elliptikus parciális differenciálegyenletek numerikus módszerei

1. fejezet

Lineáris elliptikus feladatok háttere

1.1. Elliptikus feladatok eredete

Lineáris elliptikus feladatok számos fizikai jelenség matematikai leírásában megjelennek alapvető modellként. Röviden felvázolunk néhány alapvető összefüggést, amely gyakran ilyen típusú parciális differenciálegyenletekre (PDE-kre) vezet.

Tekintsünk először egy olyan funkcionált, amely *energia* típusú mennyiséget ír le:

$$E(h) := \int_\Omega \left(\frac{1}{2} m|\nabla h|^2 + mgh \right),$$

ahol h egy $\Omega \subset \mathbb{R}^n$ korlátos tartományon értelmezett, a peremen előre ismert értékkel rendelkező függvény lehet. (Ilyen függvény leírhat hőmérsékletet, elmozdulási vagy mágneses potenciált stb.) Tegyük fel először, hogy h sima függvények körében mozoghat: ekkor tehát

$$h \in C^1(\overline{\Omega}), \quad h_{|\partial\Omega} = \varphi,$$

ahol $\varphi \in C(\partial\Omega)$ adott függvény, és $m, g \in C(\overline{\Omega})$ is adott, pozitív függvények vagy speciálisan konstansok. Vezessük be az $f := -mg$ jelölést:

$$E(h) = \int_\Omega \left(\frac{1}{2} m|\nabla h|^2 - fh \right).$$

Az ilyen modellekben a fizikai állapotot az a h írja le, amelyen az $E(h)$ energiafunkcionál értéke minimális. Ez azt jelenti, hogy ha $\tilde{h} \in C^1(\overline{\Omega})$ másik olyan függvény, melyre $\tilde{h}_{|\partial\Omega} = \varphi$, akkor $E(h) \leq E(\tilde{h})$. Ekkor h és $\tilde{h}$ eltérése a peremen 0, tehát a h függvényt homogén peremfeltételű perturbációival kell összehasonlítanunk. Legyen

$$v \in C^1(\overline{\Omega}), \quad v_{|\partial\Omega} = 0$$

tetszőleges, és $t \in \mathbb{R}$. Ekkor

$$0 \le E(h+tv) - E(h) = \int_\Omega \Big(\frac{1}{2} m\big(|\nabla h + t\nabla v|^2 - |\nabla h|^2 \big) - f(h+tv) + fh \Big)$$
$$= t \int_\Omega \big(m\, \nabla h \cdot \nabla v - fv \big) + \frac{t^2}{2} \int_\Omega m |\nabla v|^2.$$

Utóbbi (rögzített v esetén) t-nek másodfokú függvénye, amely csak akkor lehet nemnegatív bármely t-re, ha annak együtthatója 0. Így azt kaptuk, hogy az energiát minimalizáló h függvényre

$$\int_\Omega m\, \nabla h \cdot \nabla v = \int_\Omega fv \qquad (\forall v \in C^1(\overline{\Omega}),\ v_{|\partial\Omega} = 0). \tag{1.1}$$

Ha itt m és h elég sima, azaz $m\,\nabla h \in C^1(\overline{\Omega})$, és a perem is szakaszonként sima, akkor a Green-formula alapján h teljesíti a

$$\begin{cases} -\operatorname{div}(m\,\nabla h) = f, \\ h_{|\partial\Omega} = \varphi \end{cases} \tag{1.2}$$

peremérték-feladatot. (Ha h vagy m nem elég sima, akkor az integrálegyenlőség lényegében ennek gyenge megoldását jelentil, ezt pontosabban a következő szakaszban idézzük fel.) Speciálisan, konstans $m > 0$ és $\hat{f} := f/m$ esetén ez a Poisson-egyenlet:

$$\begin{cases} -\Delta h = \hat{f}, \\ h_{|\partial\Omega} = \varphi. \end{cases}$$

Olyan fizikai törvényszerűségek is elliptikus egyenletre vezethetnek, amelyeket nem energiával fogalmaztak meg. Gyakran modellezhetők egyes mennyiségek (pl. áramlások vagy erőterek) valamely

$$W \in C^1(\mathbb{R}^n, \mathbb{R}^n)$$

vektormezővel. Ha az $f \in C(\mathbb{R}^n)$ számértékű függvény sűrűség típusú mennyiséget ír le, akkor a megfelelő *megmaradási törvény* alakja

$$\operatorname{div} W = f. \tag{1.3}$$

Ekkor a Gauss–Osztrogradszkij-tétel szerint tetszőleges sima peremű $D \subset \mathbb{R}^n$ korlátos tartományon

$$\int_{\partial D} W \cdot \nu = \int_D f,$$

ahol a bal oldal a W mező fluxusát, a jobb oldal az f sűrűségnek megfelelő D-re eső összmennyiséget (tömeg, töltés stb.) írja le. Az $f \equiv 0$ esetben $\int_{\partial D} W \cdot \nu = 0$, azaz a ki- és

belépő fluxus kiegyenlíti egymást: ez a modellnek megfelelően jelenthet összenyomhatatlanságot, forrásmentességet stb., és a $\operatorname{div} W = 0$ egyenlethez vezet.

A W mezőre vonatkozó másik alapvető összefüggés a Fick-törvény, amely szerint a mező arányos egy alkalmas skalármennyiség (ún. skalárpotenciál) változásával. (Pl. az anyagáramlás a sűrűségkülönbséggel, a hőáram a hőmérséklet változásával stb.) Ilyenkor a W mező a skalárpotenciál változása által létrehozott *diffúziót* jelenti. Ezt a törvényt a

$$W = -k \nabla u \tag{1.4}$$

egyenlőség írja le, ahol a $k > 0$ függvény vagy állandó az arányossági tényező és $u \in C^1(\mathbb{R}^n)$ a skalárpotenciál. Ha k állandó, akkor W potenciálos mező, azaz $\operatorname{rot} W = 0$.

Az (1.3) és az (1.4) egyenlőségek is végül a

$$-\operatorname{div}(k \nabla u) = f$$

elliptikus egyenletre vezetnek.

A fenti $\operatorname{div} W = f$ és $W = -k \nabla u$ helyett néha más kiindulási egyenlőségből kapjuk ugyanazt az elliptikus egyenletet. Például a Maxwell-egyenletek speciális stacionárius esete a síkbeli

$$\begin{cases} \operatorname{rot} H = \rho \\ \operatorname{div} B = 0 \end{cases}$$

rendszer, ahol H a mágneses mező és B a mágneses indukció, valamint a kettő egymással egyenesen arányos, azaz van olyan $k > 0$, hogy

$$H = k\, B. \tag{1.5}$$

Ekkor a második egyenletben $\operatorname{div} B \equiv \partial_1 B_1 + \partial_2 B_2 = 0$ átírható $-\partial_2 B_2 = \partial_1 B_1$ alakba, így a

$$W := (W_1, W_2) := (-B_2, B_1)$$

mezőre $\operatorname{rot} W = 0$, azaz $W = -\nabla u$ alkalmas u potenciál mellett, vagyis (1.4) teljesül. Emellett (1.5)-ből

$$\operatorname{rot} H = \partial_2 H_1 - \partial_1 H_2 = k\,(\partial_2 B_1 - \partial_1 B_2) = k\,(\partial_2 W_2 + \partial_1 W_1) = k \operatorname{div} W = -\operatorname{div}(k \nabla u),$$

vagyis a $\operatorname{rot} H = \rho$ egyenletből

$$-\operatorname{div}(k \nabla u) = \rho.$$

A fenti példák lineáris és csak főrészből álló másodrendű egyenletre vezettek. Nemlineáris elliptikus egyenletek adódhatnak a fenti modellek általánosabb eseteiből (ilyen feladatokra a 7.1. szakaszban térünk ki, ahol pl. látni fogjuk (1.5) egy nemlineáris megfelelőjét), nemlineáris kémiai vagy biológiai reakciókból, melyek nulladrendű (azaz deriváltat nem tartalmazó) tagokkal írhatók le, stb. (lásd pl. a [12] könyv példáit). Egyéb lineáris esetek származhatnak pl. a nemlineáris modellek linearizáltjából, ill. elsőrendű tagokat kapunk konvekció típusú jelenségek modellezéséből, ilyen egyenletekkel a 3.7. fejezetben foglalkozunk, akárcsak a negyedrendű esettel. Nemlineáris negyedrendű egyenletekre, ill. elliptikus rendszerekre e jegyzet nem tér ki, erről is pl. a [12] könyvben olvashatunk.

1.2. Megoldhatóság

1.2.1. Szoboljev-terek

A továbbiakban gyakran használunk majd Szoboljev-tereket, elsősorban a végeselem-módszer esetén támaszkodunk a gyenge megoldás fogalmára és Szoboljev-térbeli becslésekre. A Szoboljev-terek fogalmáról és tulajdonságairól a [25] könyvben olvashatunk részletesen. Itt csak az alkalmazott jelöléseket és felhasznált állításokat foglaljuk össze.

1.1. Definíció. Legyen $\Omega \subset \mathbb{R}^n$ korlátos tartomány. Azt mondjuk, hogy Ω pereme *szakaszonként sima*, ha $\partial\Omega$ előáll véges sok folytonosan differenciálható felület egyesítéseként és az egész $\partial\Omega$ Lipschitz-folytonos. ♢

A továbbiakban végig feltesszük, hogy Ω korlátos tartomány szakaszonként sima peremmel. Azokat a Szoboljev-tereket használjuk majd, amelyek Hilbert-terek is, és ezeket valósnak értelmezzük: ha $k \in \mathbb{N}^+$, akkor

$$H^k(\Omega) := \{u \in L^2(\Omega) : \ \partial^\alpha u \in L^2(\Omega) \ (\forall |\alpha| \leq k)\},$$

ahol $\partial^\alpha u$ általánosított deriváltakat jelent disztribúció-értelemben. A $H^k(\Omega)$ tér skalárszorzata és az indukált norma

$$\langle u, v\rangle_{H^k(\Omega)} := \int_\Omega \sum_{|\alpha|\leq k} (\partial^\alpha u)(\partial^\alpha v), \qquad \|u\|^2_{H^k(\Omega)} = \int_\Omega \sum_{|\alpha|\leq k} \big(\partial^\alpha u\big)^2.$$

Ha nem félreérthető, hogy nem tüntetjük fel az Ω tartományt, akkor az

$$\|u\|_k := \|u\|_{H^k(\Omega)}$$

jelölést írjuk. Gyakran használjuk a csak legmagasabbrendű deriváltakat tartalmazó félnormát:

$$|u|^2_{H^k(\Omega)} := \int_\Omega \sum_{|\alpha|=k} \big(\partial^\alpha u\big)^2,$$

ill. Ω feltüntetése nélkül

$$|u|_k := |u|_{H^k(\Omega)}.$$

A fő speciális eset $k = 1$, ekkor

$$H^1(\Omega) := \{u \in L^2(\Omega) : \ \partial_i u \in L^2(\Omega) \ (\forall i = 1, \ldots, n)\},$$

ahol $\partial_i u$ általánosított deriváltakat jelent disztribúció-értelemben. Itt is, ha nem félreérthető, hogy nem tüntetjük fel az Ω tartományt, akkor

$$\|u\|_1 := \|u\|_{H^1(\Omega)}.$$

A $H^1(\Omega)$ tér fontos altere a megfelelő homogén peremfeltételt teljesítő függvényekből álló

$$H_0^1(\Omega) := \{u \in H^1(\Omega) : \quad u_{|\partial\Omega} = 0\}$$

tér, ahol $u_{|\partial\Omega}$ nyom-értelemben tekintendő. Ennek szokásos skalárszorzata és az indukált norma megegyezik a csak elsőrendű deriváltakat tartalmazó kifejezéssel, amely a norma esetén éppen a H^1-félnorma, így ott megtartjuk a korábbi jelölést:

$$\langle u, v\rangle_{H_0^1(\Omega)} := \int_\Omega \sum_{i=1}^n (\partial_i u)(\partial_i v), \qquad |u|^2_{H^1(\Omega)} = \int_\Omega \sum_{i=1}^n \big(\partial_i u\big)^2.$$

A peremfeltétel miatt a $H_0^1(\Omega)$ téren ez már norma. Most Ω feltüntetése nélkül az

$$\langle u, v\rangle_{1.0} := \langle u, v\rangle_{H_0^1(\Omega)}, \qquad |u|_1 := |u|_{H^1(\Omega)}$$

jelöléseket használjuk. A $H^1(\Omega)$ tér további fontos altere egyrészt a

$$H_D^1(\Omega) := \{u \in H^1(\Omega) : \quad u_{|\Gamma_D} = 0\} \tag{1.6}$$

tér, ahol a $\Gamma_D \subset \partial\Omega$ ún. Dirichlet-perem a perem pozitív mértékű és ún. szakaszonként sima részfelülete (utóbbin azt értjük, hogy relatíve – azaz a peremre nézve – nyílt halmaz szakaszonként sima határral) és $u_{|\Gamma_D}$ nyom-értelemben tekintendő, valamint a konstansfüggvények merőleges kiegészítője, azaz a nulla átlagú függvényekből álló

$$\dot H^1(\Omega) := \{u \in H^1(\Omega) : \int_\Omega u = 0\} \tag{1.7}$$

tér. Végül, a $k = 0$ esetben a $H^k(\Omega)$ tér megfelel az $L^2(\Omega)$ térnek, ezért szokásos az

$$\|u\|_0 := \|u\|_{L^2(\Omega)}$$

jelölés, ill. ha kell, a tartományt gyakran az

$$\|u\|_{0,\Omega} := \|u\|_{L^2(\Omega)}$$

módon tüntetjük fel.

A Szoboljev-terek egyik alapvető becslése a *Poincaré–Friedrichs-egyenlőtlenség*: van olyan $C_\Omega > 0$ konstans, hogy

$$\|u\|_{0,\Omega} \le C_\Omega |u|_1 \qquad (\forall u \in H_D^1(\Omega)), \tag{1.8}$$

azaz

$$\int_\Omega u^2 \le C_\Omega^2 \int_\Omega |\nabla u|^2 \qquad (\forall u \in H_D^1(\Omega)).$$

(A konstans értékének jól használható becslését lásd a 3.3. szakaszban.) Ebből következik többek közt a $\|\,.\,\|_1$ és $|\,.\,|_1$ normák ekvivalenciája a $H^1_D(\Omega)$ téren, hisz ez egyik irányban triviális, a másikban a Poincaré–Friedrichs-egyenlőtlenség révén

$$\|u\|_1^2 \le (1 + C_\Omega^2) \int_\Omega |\nabla u|^2 \qquad (\forall u \in H^1_D(\Omega)).$$

(A fentieket gyakran $H^1_0(\Omega)$ esetén használjuk, de Γ_D pozitív mértéke miatt igazak $H^1_D(\Omega)$-ben is, lásd [31].) Az egész $H^1(\Omega)$ téren a Poincaré–Friedrichs-egyenlőtlenség megfelelője a *Poincaré–Neumann-egyenlőtlenség*:

$$\|u\|_1^2 \le C_0\Big(\int_\Omega |\nabla u|^2 + \Big|\int_\Omega u\Big|^2\Big) = C_0\Big(|u|_1^2 + \Big|\int_\Omega u\Big|^2\Big) \qquad (\forall u \in H^1(\Omega)), \tag{1.9}$$

ebből következik, hogy a $\dot H^1(\Omega)$ téren

$$\|u\|_1^2 \le C_0 \int_\Omega |\nabla u|^2 \qquad (\forall u \in \dot H^1(\Omega)),$$

és így $\|\,.\,\|_1$ és $|\,.\,|_1$ itt is ekvivalens normák. Az (1.8) magasabbrendű megfelelője az *r-edrendű Poincaré–Neumann-egyenlőtlenség*:

$$\|u\|_r^2 \le C_r\Big(|u|_r^2 + \sum_{|\alpha|\le r-1} \Big|\int_\Omega \partial^\alpha u\Big|^2\Big) \qquad (\forall u \in H^r(\Omega)). \tag{1.10}$$

Végül egy $u \in H^1(\Omega)$ függvénynek a perem valamely $\Gamma \subset \partial\Omega$ (pozitív mértékű és szakaszonként sima) részfelületén való $u_{|\Gamma}$ nyomának az $\|u\|_{0,\Gamma} := \|u\|_{L^2(\Gamma)}$ normájára is érvényes hasonló becslés, ami épp a nyom-operátor folytonosságát jelenti: van olyan $\tilde C_\Gamma > 0$ konstans, hogy

$$\|u\|_{0,\Gamma} \le \tilde C_\Gamma \|u\|_1 \qquad (\forall u \in H^1(\Omega)). \tag{1.11}$$

A $H^1_D(\Omega)$ téren az ekvivalencia miatt a $|\,.\,|_1$ norma is írható:

$$\|u\|_{0,\Gamma} \le C_\Gamma |u|_1 \qquad (\forall u \in H^1_D(\Omega)), \tag{1.12}$$

1.2.2. Lax–Milgram-elmélet, gyenge és reguláris megoldás

Röviden összefoglaljuk a peremérték-feladatok gyenge megoldásáról tudnivalókat, a részleteket lásd pl. a [16, 25] könyvekben.

Tekintsünk először egy

$$\begin{cases} Lu := -\operatorname{div}(p\,\nabla u) = f, \\ u_{|\partial\Omega} = 0 \end{cases} \tag{1.13}$$

peremérték-feladatot, ahol $\Omega \subset \mathbb{R}^N$ korlátos tartomány, és alkalmas $m > 0$ esetén $p(x) \geq m > 0$ $(\forall x \in \Omega)$. A fenti feladatot homogén peremfeltétellel nézzük, az inhomogén eset könnyen visszavezethető erre, lásd majd az 1.6. megjegyzést.

Az (1.13) feladat megoldásától a klasszikus esetben az $u \in C^2(\Omega)$ simaságot várjuk; ha $p \in C^1(\overline{\Omega})$, akkor ez esetben az Lu kifejezés valóban értelmes, és $f \in C(\Omega)$ mellett lehet egyenlőség. Ha azonban a p és f adatok nem ilyen simák (pl. p lehet szakaszonként konstans anyagállandó), akkor általában csak a gyenge megoldás fogalma értelmezhető.

A gyenge megoldás fogalmához az (1.13) egyenletet a Green-formula segítségével átalakítjuk, éppen ellenkező irányban, mint ahogy (1.1)-ből kaptuk (1.2)-t, valamint észrevesszük, hogy ez értelmes akkor is, ha u is csak $H_0^1(\Omega)$-ból való. Az (1.13) feladat gyenge megoldása tehát olyan $u \in H_0^1(\Omega)$ függvény, melyre

$$\int_\Omega p\ \nabla u \cdot \nabla v = \int_\Omega f v \qquad (\forall v \in H_0^1(\Omega)). \tag{1.14}$$

A gyenge megoldás létezése és egyértelműsége a Lax–Milgram-elmélet alapján Hilbert-térbeli bilineáris formák segítségével igazolható, a részleteket lásd pl. a [16, II.7.2] könyvben. Esetünkben ennek leginkább használt változatát, a koercív esetet alkalmazhatjuk:

1.2. Tétel. (Lax–Milgram-lemma) *Legyen H valós Hilbert-tér, $a : H \times H \to \mathbb{R}$ korlátos, koercív bilineáris forma, azaz*

(i) létezik $M > 0$, hogy

$$|a(u,v)| \leq M\|u\|\,\|v\| \qquad (\forall u, v \in H); \tag{1.15}$$

(ii) létezik $m > 0$, hogy

$$a(u,u) \geq m\|u\|^2 \qquad (\forall u \in H). \tag{1.16}$$

Ekkor bármely $\ell : H \to \mathbb{R}$ korlátos lineáris funkcionálhoz létezik egyetlen olyan $u \in H$, melyre

$$a(u,v) = \ell v \qquad (\forall v \in H). \tag{1.17}$$

Az (1.17) egyenlőséget szokás variációs feladatnak hívni. Az (1.14) gyenge alakú feladat az (1.17) egyenlőség speciális esete. A p függvény pozitív alsó határa és korlátossága révén adódik az

$$a(u,v) := \int_\Omega p\ \nabla u \cdot \nabla v \qquad (\forall u, v \in H_0^1(\Omega)) \tag{1.18}$$

forma koercivitása és korlátossága a $H := H_0^1(\Omega)$ Szoboljev-térben, itt a korlátossághoz elég $p \in L^\infty(\Omega)$ is: $M := \|p\|_{L^\infty}$ mellett bármely $u, v \in H_0^1(\Omega)$ esetén

$$|a(u,v)| \leq \|p\|_{L^\infty}\|\nabla u\|_{L^2}\|\nabla v\|_{L^2} = M\,|u|_1\,|v|_1, \tag{1.19}$$

$$a(u,u) \geq m\|\nabla u\|_{L^2}^2 = m\,|u|_1^2. \tag{1.20}$$

Emellett $f \in L^2(\Omega)$ esetén (1.13) jobboldala korlátos lineáris funkcionálja v-nek.

1.3. Következmény. *Ha $p \in L^\infty(\Omega)$ és alkalmas $m > 0$ esetén $p(x) \geq m > 0$ (m-m. $\forall x \in \Omega$), akkor bármely $f \in L^2(\Omega)$ esetén az (1.13) feladatnak létezik egyetlen gyenge megoldása.*

A gyenge megoldás gondolatmenete ugyanígy alkalmazható akkor is, ha az egyenletben nulladrendű tag szerepel, és/vagy vegyes a peremfeltétel:

$$\begin{cases} -\operatorname{div}(p\,\nabla u) + qu = f, \\ u_{|\Gamma_D} = 0, \quad (p\,\partial_\nu u + su)_{|\Gamma_N} = \gamma, \end{cases} \tag{1.21}$$

ahol $s, q \geq 0$, $q \in L^\infty(\Omega)$, $s \in L^\infty(\Gamma_N)$, $\gamma \in L^2(\Gamma_N)$, ill. Γ_D és Γ_N a perem szakaszonként sima részfelületekre való felbontását alkotják az (1.6) után definiált értelemben, és Γ_D pozitív mértékű. Ekkor a gyenge alak a $H^1_D(\Omega)$ Szoboljev-térben

$$\int_\Omega (p\ \nabla u \cdot \nabla v + quv) + \int_{\Gamma_N} suv = \int_\Omega fv + \int_{\Gamma_N} \gamma v \qquad (\forall v \in H^1_D(\Omega)). \tag{1.22}$$

Hasonlóan kezelhetők a Neumann-feladatok is, ekkor azonban a

$$\begin{cases} -\operatorname{div}(p\,\nabla u) = f, \\ \partial_\nu u_{|\partial\Omega} = 0 \end{cases} \tag{1.23}$$

feladatban a megoldás additív konstans erejéig nem egyértelmű, ha a teljes $H^1(\Omega)$ térben nézzük, ezért itt a megoldást a $\dot{H}^1(\Omega)$ Szoboljev-térben keressük. A gyenge alak megegyezik az (1.14) egyenlőséggel, és a $\dot{H}^1(\Omega)$ térben a bal oldal már koercív bilineáris forma, így az elmélet alkalmazható.

1.4. Megjegyzés. Az (1.18) bilineáris forma, ill. ennek az (1.22) bal oldalán álló módosítása szimmetrikus is. A Lax–Milgram-lemma ezt nem követeli meg; a 3.7. szakaszban nem szimmetrikus (elsőrendű tagot is tartalmazó) elliptikus feladatra látni fogjuk ezen általánosabb eset alkalmazását is. ♢

A későbbiekben gyakran fontos szerepe van a megoldás regularitásának, azaz az $u \in H^1(\Omega)$-nél többszöri deriválhatóságnak.

Az egész zárt tartományon való $u \in C^2(\overline{\Omega})$ klasszikus simaság már egyszerű esetben sem teljesülhet: könnyen látható, hogy ha Ω az egységnégyzet, akkor a homogén peremfeltétel miatt az (1.13) feladat megoldására $\Delta u(0,0) = 0$ kell fennálljon. Ha például $f \equiv 1$, akkor $u \in C^2(\overline{\Omega})$ esetén $\Delta u(0,0) = 1$ lenne, így ez nem teljesülhet. A gondot itt a sarok okozza. Ha $f(0,0) = 0$, akkor lehet bármilyen sima megoldás is, pl. $u(x,y) = \sin \pi x \sin \pi y$.

A H^2-beliség viszont elég tág esetben igaz Dirichlet-feladatok esetén:

1.5. Tétel. (Kadlec, [15]) *Ha az Ω tartomány C^2-diffeomorf egy konvex tartománnyal, és $p \in Lip(\overline{\Omega})$ (pl. $p \in C^1(\overline{\Omega})$), akkor az (1.13) feladat megoldására $u \in H^2(\Omega)$. Sőt, van olyan $c > 0$ állandó, hogy*

$$\|u\|_{H^2} \leq c\|f\|_{L^2}. \tag{1.24}$$

Ha $\partial\Omega$ szakaszonként sima, akkor a C^2-diffeomorfizmus feltétele azt jelenti, hogy Ω lokálisan konvex a perem töréspontjaiban.

Ha ez nem áll fenn, akkor lehet $u \notin H^2(\Omega)$, és a gondot a konkáv sarok okozza: lásd pl. a [27] könyv III. 15.2. fejezetét, ha Ω egy egységkörből kivágott negyedkör.

1.6. Megjegyzés. A fentiekben homogén peremfeltétellel tekintettük az elliptikus feladatokat. Az inhomogén eset könnyen visszavezethető erre az ismert módszerrel. Tekintsünk egy

$$\begin{cases} Lu = f, \\ u_{|\partial\Omega} = g \end{cases} \tag{1.25}$$

Dirichlet-feladatot és legyen $\widetilde{g} \in D(L)$, melyre $\widetilde{g}_{|\partial\Omega} = g$. Ha megoldjuk az

$$\begin{cases} Lz = f - L\widetilde{g}, \\ z_{|\partial\Omega} = 0 \end{cases} \tag{1.26}$$

homogén peremfeltételű segédfeladatot, akkor $u := z + \widetilde{g}$ az eredeti feladat megoldása. Utóbbi gyenge alakban is felírható (ekkor az $\widetilde{g} \in D(L)$ feltétel helyett elég $\widetilde{g} \in H^1(\Omega)$, melyre $\widetilde{g}_{|\partial\Omega} = g$ nyom-értelemben, ez az ún. Dirichlet-beterjesztés):

$$\int_\Omega p\ \nabla z \cdot \nabla v = \int_\Omega (fv - p\ \nabla\widetilde{g} \cdot \nabla v) \qquad (\forall v \in H_0^1(\Omega)).$$

Ha ebben a jobboldali $\widetilde{g}$-os tagot balra rendezzük és felhasználjuk, hogy $u = z + \widetilde{g}$, akkor megkapjuk az inhomogén Dirichlet-feladat szokásos gyenge alakját: az $u \in H^1(\Omega)$ megoldás olyan függvény, melyre

$$\int_\Omega p\ \nabla u \cdot \nabla v = \int_\Omega fv \qquad (\forall v \in H_0^1(\Omega))$$

(azaz a tesztfüggvények csak homogén peremfeltételűek), és

$$u_{|\partial\Omega} = g \ \text{ nyom-értelemben} \qquad (\Leftrightarrow \quad u - \widetilde{g} \in H_0^1(\Omega)).$$

(Mindez értelemszerűen átvihető Neumann-, ill. vegyes peremfeltételre is, lásd [25].) $\diamondsuit$

2. fejezet

A véges differenciák módszere (FDM)

A véges differenciák módszere (angolul „finite difference method”, betűszóként az ennek megfelelő FDM-et fogjuk használni) arra alapul, hogy az egyenletben szereplő deriváltakat alkalmas különbségi hányadosokkal (véges differenciákkal) közelítjük. Ezt véges sok kiválasztott pontban (ún. csomópontban) tesszük, és a megoldást is csak ezekben a pontokban keressük. Ezzel a PDE közelítő megoldását lineáris algebrai egyenletrendszerre vezetjük vissza.

Részletesen megvizsgáljuk a téglalap-tartomány esetét, ahol egyszerű a módszer konstrukciója, a továbbiakban emellett homogén Dirichlet-peremfeltétellel tekintjük az elliptikus feladatokat. Az inhomogén és más peremfeltételeket, ill. általánosabb tartomány esetét a szakasz végén a 2.5. pontban érintjük.

2.1. Néhány elméleti segédeszköz

2.1.1. Alapvető diszkretizációs sémák

Idézzük fel az alapvető egydimenziós diszkretizációs sémákat a közönséges differenciálegyenletek véges differenciás megoldásának elméletéből.

- jobb oldali differencia :
$$D_+v(x) = v(x+h) - v(x)$$
- bal oldali differencia :
$$D_-v(x) = v(x) - v(x-h)$$
- centrális differencia :
$$D_0v(x) = \frac{1}{2}(v(x+h) - v(x-h)).$$

Az első derivált közelítését ezek alapján a következőképp írhatjuk fel adott rácspontokban. Tekintsünk egy $I = [0, a]$ intervallumot és abban az

$$x_i := ih \qquad (i = 0, \dots, n+1)$$

pontokat, ahol $n \in \mathbb{N}^+$ és $h := a/(n+1)$. Ha $u \in C(I)$, akkor legyen

$$u_i := u(x_i) \qquad (i = 0, \dots, n+1).$$

Mivel homogén peremfeltételű függvényekkel foglalkozunk, itt ekkor $u_0 = u_{n+1} = 0$ lesz.

Ekkor $u'(x_i)$ közelítését a belső pontokban a fenti differenciák h-adrészével értelmezhetjük: ha $i = 1, \dots, n$, akkor legyen

- *haladó differenciaséma*:
$$u_{x,i} := \frac{u_{i+1} - u_i}{h},$$
- *retrográd differenciaséma*:
$$u_{\overline{x},i} := \frac{u_i - u_{i-1}}{h},$$
- *centrális differenciaséma*:
$$u_{\dot{x},i} := \frac{u_{i+1} - u_{i-1}}{2h}.$$

Az $u''(x_i)$ második deriváltakra adódik ebből:

- *másodrendű centrális differenciaséma*:
$$u_{\overline{x}x,i} = u_{x\overline{x},i} = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} \qquad (i = 1, \dots, n).$$

Megjegyzendő, hogy az $u_{\dot{x}\dot{x},i}$ kétszeres elsőrendű centrális séma viszont nem ezt adja, hanem a $2h$ rácstávolsághoz tartozó megfelelő közelítést. Ha azonban bevezetjük a $h/2$ rácstávolsághoz tartozó centrális sémát:

$$u_{\tilde{x},i} := \frac{u_{i+1/2} - u_{i-1/2}}{h},$$

akkor ennek kétszeres alkalmazása már visszaadja a másodrendű centrális differenciasémát:

$$u_{\tilde{x}\tilde{x},i} = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} \qquad (i = 1, \dots, n).$$

2.1. Megjegyzés. A fenti sémák rendje az u függvény kellő simasága esetén a következő: az elsőrendű haladó, retrográd és centrális differenciasémákra

$$u_{x,i} - u'(x_i) = O(h), \qquad u_{\overline{x},i} - u'(x_i) = O(h), \qquad u_{\dot{x},i} - u'(x_i) = O(h^2),$$

a másodrendű centrális differenciasémára

$$u_{\overline{x}x,i} - u''(x_i) = O(h^2).$$

Konkrétabb megfogalmazásra az utóbbi esetén lesz szükségünk, lásd (2.5). ♢

2.1.2. M-mátrixok és alaptulajdonságaik

A későbbiekben szükségünk lesz e fontos mátrixosztály fogalmára és néhány tulajdonságára.

2.2. Definíció. Egy $A \in \mathbb{R}^{N\times N}$ mátrix *M-mátrix*, ha

(i) $a_{ij} \le 0, \quad \text{ha} \quad i \neq j;$

(ii) van olyan

$$g > 0 \quad \text{vektor, melyre} \quad Ag > 0,$$

ahol az egyenlőtlenségeket koordinátánként értjük. ♢

2.3. Definíció. Egy $A \in \mathbb{R}^{N\times N}$ mátrix *diagonálisan domináns*, ha

$$\sum_{\substack{j=1\\ j\neq i}}^{N} |a_{ij}| < |a_{ii}| \quad (\forall i = 1, \dots, N).$$

♢

2.4. Definíció. Egy $A \in \mathbb{R}^{N\times N}$ mátrix *megengedett előjeleloszlású*, ha

(i) $a_{ii} \ge 0 \quad (\forall i = 1, \dots, N);$

(ii) $a_{ij} \le 0, \qquad \text{ha} \quad i \neq j.$ ♢

A következő három állítás bizonyításához lásd a 9.1.–9.6. feladatokat.

2.5. Állítás. *Egy $A \in \mathbb{R}^{N\times N}$ megengedett előjeleloszlású mátrix pontosan akkor diagonálisan domináns, ha M-mátrix a $g := e$ vektor mellett, ahol e a csupa 1-esből álló oszlopvektor.*

2.6. Állítás.

(1) Ha az A megengedett előjeleloszlású mátrix diagonálisan domináns, akkor A reguláris és $A^{-1} \ge 0$.

(2) Ha A M-mátrix, akkor A reguláris és $A^{-1} \ge 0$.

(Itt az egyenlőtlenségeket elemenként értjük.)

2.7. Következmény. *Ha A M-mátrix, akkor A^{-1} monoton mátrix, azaz ha $x \le y$, akkor $A^{-1}x \le A^{-1}y$.*

2.8. Állítás. *Legyen A M-mátrix egy $g > 0$ vektorral. Ekkor $\|A^{-1}\|_\infty \le \frac{\max g}{\min Ag}$, ahol max és min koordinátánként értendő.*

2.2. Az FDM konstrukciója Poisson-egyenletre téglalapon

A feladat.

Tekintsük a Poisson-egyenletet az $\Omega :=]0, a[\times]0, b[$ téglalapon homogén Dirichlet-peremfeltételel, vagyis a

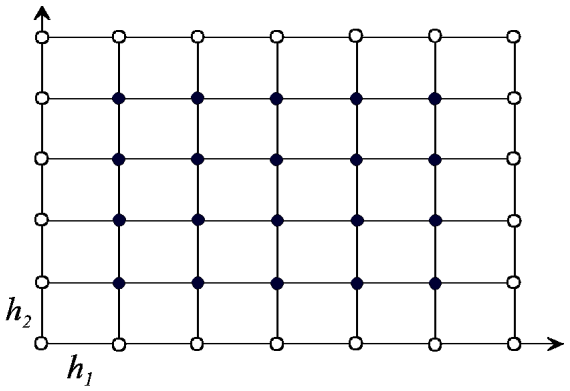
$$\begin{cases} -\Delta u = f, \\ u_{|\partial\Omega} = 0 \end{cases} \tag{2.1}$$

peremérték-feladatot.

Fedjük koordinátairányonként ekvidisztáns ráccsal Ω belsejét, azaz tekintsük az alábbi rácsot:

$$\omega_h := \{(ih_1,\, jh_2) :\ i = 1, \dots, n,\ j = 1, \dots, m\},$$

ahol $h_1, h_2 > 0$ és $n, m \in \mathbb{N}^+$ adott számok, $h_1 := a/(n+1)$ és $h_2 := b/(m+1)$, lásd 2.1. ábra. (A homogén peremfeltétel miatt a rácsba nem vesszünk bele perempontokat.)



2.1. ábra. Rácspontok Ω-ban.

2.9. Definíció. Az $(ih_1,\, jh_2)$ pontokat az ω_h rács *csomópontjainak* hívjuk. A rács *finomsága* a $h := \max\{h_1, h_2\}$ szám. ♢

(Megjegyezzük, hogy $\mathbf{h} = (h_1, h_2)$ jelöléssel precízebb lenne az $\omega_{\mathbf{h}}$ jelölés, mint a csak a rácsfinomságra utaló fenti ω_h, ez azonban nem okoz majd félreértést.)

Az u megoldást a rács csomópontjaiban szeretnénk közelíteni. A rács

$$N := nm$$

csomópontból áll, melyeket a helyzettől függően kétféleképp fogunk indexelni. Ha a fenti koordinátánkénti elhelyezkedés is számít, akkor az (i, j) indexpárral indexeljük. Emellett sorbarendezzük a csomópontokat a bal alsótól soronként rendre jobbra haladva a jobb felsőig, és erre az

$$x_k := (ih_1,\, jh_2) \qquad (k = 1, \dots, N)$$

indexeket használjuk. Ezzel a csomópontok egy N-dimenziós vektorba írhatók.

A fentieknek megfelelően értelmezzük Ω-n értelmezett függvények csomóponti értékeit is:

2.10. Definíció. Ha $z \in C(\overline{\Omega})$, akkor

$$z_{ij} := z(ih_1,\, jh_2) \qquad (i = 1, \dots, n,\ j = 1, \dots, m),$$

vagyis az egyindexű írásmóddal

$$z_{ij} = z(x_k) \qquad (k = 1, \dots, N).$$

A z függvény csomóponti értékeiből álló vektor:

$$z^h \in \mathbb{R}^N, \quad (z^h)_k := z(x_k) \qquad (k = 1, \dots, N). \qquad \diamondsuit$$

A z^h vektor tehát a z függvénynek az ω_h rácsra való leszűkítésével, azaz $z_{|\omega_h}$-val azonosítható, ha az $\mathbb{R}^N$-beli koordinátákat a csomóponti értékeknek feleltetjük meg.

A (2.1) egyenlet csomóponti közelítése e definíció szerint azt jelenti, hogy az u^h vektort szeretnénk közelítőleg kiszámítani.

A Laplace-operátor közelítése differenciasémával

Az egyváltozós másodrendű differenciaséma alapján értelemszerűen közelíthető a Laplace-operátor a csomópontokban a

$$\Delta u \approx u_{x_1\overline{x}_1} + u_{x_2\overline{x}_2}$$

sémával. Azaz, mivel

$$\begin{aligned} \partial_1^2 u(x_k) &\approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h_1^2} \qquad (i = 1, \dots, n), \\ \partial_1^2 u(x_k) &\approx \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h_2^2} \qquad (i = 1, \dots, m), \end{aligned} \tag{2.2}$$

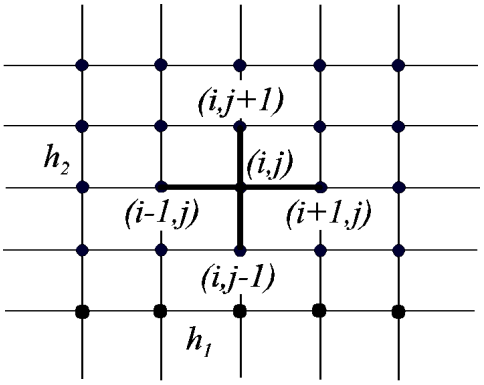
így $\Delta u(x_k)$ közelítése

$$(\Delta_h u^h)_k := \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h_1^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h_2^2} \tag{2.3}$$

minden $k = 1, \ldots, N$, azaz $i = 1, \ldots, n$ és $j = 1, \ldots, m$ esetén (vagyis minden $x_k = (ih_1,\, jh_2)$ pontban). Speciálisan, ha $h_1 = h_2 =: h$, akkor

$$(\Delta_h u^h)_k := \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}}{h^2}.$$

A felhasznált pontokat a 2.2. ábra mutatja, ez az elrendezés a séma ún. *differenciacsillagja* vagy stencilje.



2.2. ábra. Az 5-pontos differenciacsillag.

Vegyük észre, hogy a Δ_h leképezés csak az u függvény csomóponti értékeit használja fel, így értelmes mint

$$\Delta_h : \mathbb{R}^N \to \mathbb{R}^N, \qquad u^h \mapsto \Delta_h u^h \tag{2.4}$$

leképezés.

A $\Delta u(x_k) \approx (\Delta_h u^h)_k$ közelítésekből már akkor is származna minden csomópontban egy kiindulási hiba, ha u a pontos megoldás lenne. Ezért először ezzel a hibával foglalkozunk általános függvényre.

2.11. Definíció. Legyen $u \in C^2(\overline{\Omega})$ adott függvény. A (2.3) differenciaséma *képlethibáján* a

$$\psi^h := (\Delta u)^h - \Delta_h u^h$$

vektort értjük. ♢

Itt Δu^h az u függvény csomóponti értékeiből alkotott vektor, és ψ^h egy hibavektor, melynek koordinátái

$$\psi^h_k := \Delta u(x_k) - (\Delta_h u^h)_k \qquad (k = 1, \ldots, N).$$

általában ψ^h maximum-normáját szeretnénk felülről becsülni.

2.12. Állítás. *Ha $u \in C^4(\overline{\Omega})$, akkor a képlethibára*

$$\|\psi^h\|_\infty \le \frac{1}{6}|u|_{C^4}\, h^2$$

teljesül, ahol $h := \max\{h_1, h_2\}$ és $|u|_{C^4} := \max\limits_{|\alpha|=4} \|\partial^\alpha u\|_\infty$.

Bizonyítás. A közönséges differenciálegyenleteknél ismert analóg állítás nyomán következik: ha $y \in C^4(I)$ egyváltozós függvény, akkor Taylor-sorfejtéssel

$$y(x+h) + y(x-h) - 2y(x) = y''(x)h^2 + R(x), \qquad \text{ahol} \quad |R(x)| \le \frac{1}{12}\max_I |y^{(4)}|\, h^4,$$

így

$$\left|\frac{y(x+h) + y(x-h) - 2y(x)}{h^2} - y''(x)\right| \le \frac{1}{12}\max_I |y^{(4)}|\, h^2. \tag{2.5}$$

Ezt változónként alkalmazva, $x_k = (ih_1, jh_2)$ esetén

$$\begin{aligned}
|\psi_k^h| &= |(\Delta_h u^h)_k - \Delta u(x_k)| \\
&\le \left|\frac{u_{i+1,j} + u_{i-1,j} - 2u_{i,j}}{h_1^2} - \partial_1^2 u(x_k)\right| + \left|\frac{u_{i,j+1} + u_{i,j-1} - 2u_{i,j}}{h_2^2} - \partial_2^2 u(x_k)\right| \qquad \square \\
&\le \frac{1}{12}\big(\max_{\overline{\Omega}} |\partial_1^{(4)} u|\, h_1^2 + \max_{\overline{\Omega}} |\partial_2^{(4)} u|\, h_2^2\big) \le \frac{1}{6}\max_{|\alpha|=4} \|\partial^\alpha u\|_\infty\, h^2.
\end{aligned}$$

2.13. Megjegyzés.

(i) Ha u csak kevesebbszer differenciálható: $u \in C^3(\overline{\Omega})$ esetén a Taylor-sorban $|R(x)|$ csak harmadrendben becsülhető, így a (2.3) differenciaséma csak $O(h)$ rendű, ill. $u \in C^2(\overline{\Omega})$ esetén a (2.5)-beli különbség $|y''(\xi) - y''(x)|$ alakú lesz, amely $|\xi| \le h \to 0$ miatt 0-hoz tart, így a séma konvergens, de rend nem adható meg.

(ii) Ha u többször differenciálható, akkor magasabbrendű differenciasémák is felírhatók, a közelítések magasabb rendjét a Taylor-sor több tagjának figyelembevételével igazolhatjuk. Pl. többféle negyedrendű séma konstruálható 9-pontos differenciacsillaggal, ha a (2.3)-ben szereplő öt függvényérték mellett az $u_{i\pm1,j\pm1}$ értékeket is figyelembe vesszük valamely súlyokkal, lásd pl. [23] és [27, 15.3.2. fejezet]. (A (2.3) differenciaséma rendje viszont többször differenciálható u esetén is csak $O(h^2)$.) $\diamondsuit$

A csomóponti értékek közelítő meghatározása.

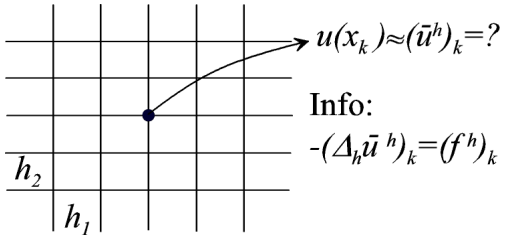
Tekintsük a (2.3) differenciasémával közelített Poisson-egyenletet, és jelöljük $\overline{u}^h$-sal az ebből kapott közelítő csomóponti megoldásvektort. Ekkor tehát a

$$-\Delta u(x_k) = f(x_k), \qquad \text{azaz} \qquad -\left(\Delta u\right)_k^h = \left(f^h\right)_k \qquad (k = 1, \ldots, N)$$

pontos csomóponti egyenlőségek helyett a

$$-\left(\Delta_h \overline{u}^h\right)_k = \left(f^h\right)_k \qquad (k = 1, \ldots, N) \tag{2.6}$$

egyenlőségeket oldjuk meg.



2.3. ábra. A csomóponti értékek keresése.

Mivel $\left(\Delta_h \overline{u}^h\right)_k$ az $\overline{u}^h$ vektor koordinátáinak lineáris kombinációja, ezért a (2.6) egyenlőségek egy lineáris algebrai egyenletrendszert határoznak meg. Jelöljük ennek mátrixát (vagyis a (2.4) leképezés mátrixát) A_h-val, azaz

$$A_h \overline{u}^h = -\Delta_h \overline{u}^h.$$

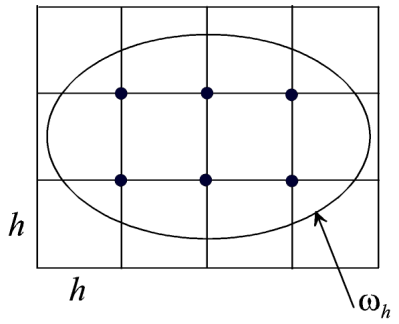
Ezzel a (2.6) egyenlőségek az

$$A_h \overline{u}^h = f^h$$

lineáris algebrai egyenletrendszer alakjában írhatók.

Vizsgáljuk meg, milyen alakú az A_h mátrix! Tekintsünk először példaként egy 3×2 belső pontból álló, a két irányban egyforma lépésközű rácsot, azaz $\omega_h := \{(ih,\, jh) :\ i = 1, 2, 3,\ \ j = 1, 2\}$, lásd 2.4. ábra. Könnyen látható (lásd 9.8. feladat), hogy ekkor

$$A_h = \frac{1}{h^2} \begin{pmatrix} 4 & -1 & & -1 & & \\ -1 & 4 & -1 & & -1 & \\ & -1 & 4 & & & -1 \\ -1 & & & 4 & -1 & \\ & -1 & & -1 & 4 & -1 \\ & & -1 & & -1 & 4 \end{pmatrix}.$$



2.4. ábra. Példa: 3×2 belső pontos négyzetrács.

A fentihez hasonlóan igazolható, hogy általában $n \times m$ belső pontból álló, a két irányban egyforma h lépésközű rács esetén az

$$A_h = \frac{1}{h^2} \begin{pmatrix} B & -I & & & & \\ -I & B & -I & & & \\ & -I & B & -I & & \\ & & \ddots & & \ddots & \\ & & & -I & B & -I \\ & & & & -I & B \end{pmatrix} \in \mathbb{R}^{N \times N} = \mathbb{R}^{nm \times nm}$$

blokk-tridiagonális mátrixhoz jutunk, ahol I az $n \times n$-es identitásmátrix és

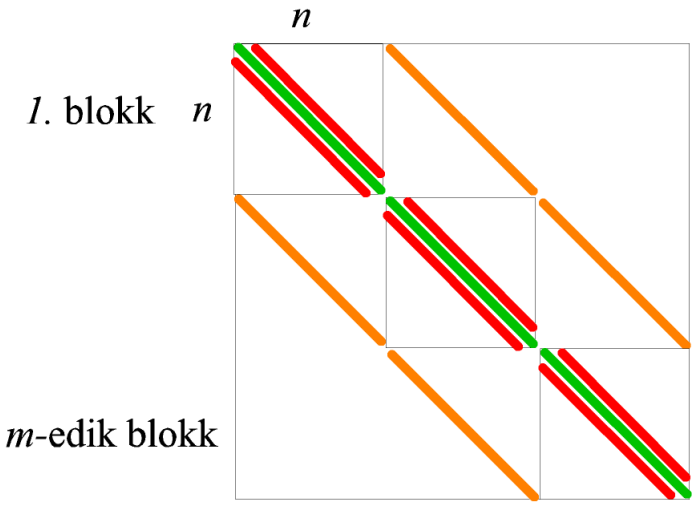
$$B = \begin{pmatrix} 4 & -1 & & & & \\ -1 & 4 & -1 & & & \\ & -1 & 4 & & & \\ & & \ddots & & \ddots & \\ & & & -1 & 4 & -1 \\ & & & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

tridiagonális mátrix, melyből m db szerepel. Ha az $n \times m$ belső pontból álló rács a két irányban különböző h_1 és h_2 lépésközű, akkor nem emelhető ki az $\frac{1}{h^2}$-hez hasonló tényező, ekkor a mátrix a fenti helyett

$$A_h = \begin{pmatrix} \widetilde{B} & -\frac{1}{h_2^2} I & & & & \\ -\frac{1}{h_2^2} I & \widetilde{B} & -\frac{1}{h_2^2} I & & & \\ & -\frac{1}{h_2^2} I & \widetilde{B} & -\frac{1}{h_2^2} I & & \\ & & \ddots & & \ddots & \\ & & & -\frac{1}{h_2^2} I & \widetilde{B} & -\frac{1}{h_2^2} I \\ & & & & -\frac{1}{h_2^2} I & \widetilde{B} \end{pmatrix} \tag{2.7}$$

lesz, ahol

$$\widetilde{B} = \begin{pmatrix} \frac{2}{h_1^2}+\frac{2}{h_2^2} & -\frac{1}{h_1^2} & & & & & \\ -\frac{1}{h_1^2} & \frac{2}{h_1^2}+\frac{2}{h_2^2} & -\frac{1}{h_1^2} & & & & \\ & -\frac{1}{h_1^2} & \frac{2}{h_1^2}+\frac{2}{h_2^2} & & & & \\ & & \ddots & \ddots & & & \\ & & & -\frac{1}{h_1^2} & \frac{2}{h_1^2}+\frac{2}{h_2^2} & -\frac{1}{h_1^2} & \\ & & & & -\frac{1}{h_1^2} & \frac{2}{h_1^2}+\frac{2}{h_2^2} & \end{pmatrix}.$$



2.5. ábra. Az A_h mátrix szerkezete. Zöld: $2/h_1^2 + 2/h_2^2$, piros: $-1/h_1^2$, narancs: $-1/h_2^2$.

Az $\overline{u}^h$ közelítő megoldás előállításához tehát feladatunk az

$$A_h \overline{u}^h = f^h \tag{2.8}$$

lineáris algebrai egyenletrendszer megoldása.

Az $A_h\overline{u}^h = f^h$ lineáris algebrai egyenletrendszer

Ennek vizsgálatához felhasználjuk a 2.1.2. szakaszból az M-mátrixok fogalmát és néhány tulajdonságát.

2.14. Állítás. *Az A_h mátrix M-mátrix.*

Bizonyítás. (i) Az $a_{ij} \leq 0 \quad (i \neq j)$ előjelfeltétel teljesül.

(ii) Tekintsük a $w(x_1, x_2) := x_1(a - x_1) + x_2(b - x_2)$ (ha $(x_1, x_2) \in \Omega$) függvényt, és legyen $g := w^h$ (a w csomóponti értékeinek vektora). Ekkor $g > 0$, mert $w(x_1, x_2) > 0$ minden $(x_1, x_2) \in \Omega$ esetén. Másrészt $-\Delta w \equiv 4$, és így

$$(A_h g)_k = -(\Delta_h w^h)_k = -\Delta w(x_k) = 4 > 0 \qquad (\forall k = 1, \dots, N),$$

mivel a Δ_h-ban szereplő másodrendű differenciaséma pontos a w másodrendű polinomon. (Utóbbi következik az egy dimenzióban ismert analóg tulajdonságból, de közvetlenül látható a 2.12. állításból is, hiszen $|w|_{C^4} = \max_{|\alpha|=4} \|\partial^\alpha w\|_\infty = 0$, így a képlethiba 0.) □

2.15. Következmény. *A_h reguláris, így az $A_h \overline{u}^h = f^h$ lineáris algebrai egyenletrendszernek egyértelműen létezik megoldása.*

2.16. Megjegyzés. További következmények:

(i) Az M-mátrix-tulajdonság a lineáris algebrai egyenletrendszer hatékony iterációs megoldása szempontjából is kedvező, lásd a 4.3. fejezetben.

(ii) $A_h^{-1} \geq 0$ (elemenként). Itt A_h a $-\Delta$ operátor közelítése, így A_h^{-1} a $-\Delta$ inverzéé. Utóbbi felírható a Green-függvénnyel való szorzással és integrálással, így A_h^{-1} konstans szorzó erejéig a $G > 0$ Green-függvény közelítése: a $(h_1 h_2 A_h)^{-1}$ mátrixot szokás is diszkrét Green-függvénynek hívni. Az A_h^{-1} elemenkénti nemnegativitása tehát az általa közelített Green-függvény nemnegativitásának felel meg. ♢

2.17. Állítás. *Az A_h mátrixra $\|A_h^{-1}\|_\infty \leq \frac{a^2+b^2}{16}$. (A becslés nem függ h-tól.)*

Bizonyítás. Alkalmazzuk a 2.8. állítást a 2.14. tétel bizonyításában szereplő $g := w^h$ vektorral! Ekkor

$$\|A_h^{-1}\|_\infty \leq \frac{\max g}{\min A_h g} = \frac{\max_k (w^h)_k}{\min_k (-\Delta_h w^h)_k} = \frac{1}{4} \max_k w(x_k) \leq \frac{1}{4} \max_{x \in \overline{\Omega}} w(x) = \frac{a^2 + b^2}{16}.$$

□

2.3. Stabilitás és konvergencia

A közelítő megoldás stabilitása.

Maximum-normabeli stabilitást vizsgálunk, azaz a megoldás maximum-normáját becsüljük felülről a jobboldal maximum-normájával.

2.18. Állítás. *Tekintsük az $A_h \overline{u}^h = f^h$ lineáris algebrai egyenletrendszert. Ekkor*

$$\|\overline{u}^h\|_\infty \leq \frac{a^2 + b^2}{16} \|f^h\|_\infty.$$

Bizonyítás. Mivel $\overline{u}^h = A_h^{-1} f^h$, a 2.17. állítás alapján

$$\|\overline{u}^h\|_\infty \leq \|A_h^{-1}\|_\infty \|f^h\|_\infty \leq \frac{a^2+b^2}{16} \|f^h\|_\infty. \qquad \square$$

A stabilitás szemléletesen azt fejezi ki, hogy a jobboldal hibája korlátos mértékben öröklődik a megoldás hibájára, így ha előbbi kicsi, akkor az utóbbi is. Valóban:

2.19. Állítás. *Legyen az $A_h \overline{u}^h = f^h$ lineáris algebrai egyenletrendszer f_1^h és f_2^h jobboldalakhoz tartozó megoldása rendre $\overline{u}_1^h$ és $\overline{u}_2^h$. Ekkor*

$$\|\overline{u}_1^h - \overline{u}_2^h\|_\infty \leq \frac{a^2+b^2}{16} \|f_1^h - f_2^h\|_\infty.$$

Bizonyítás. A linearitás miatt $A_h(\overline{u}_1^h - \overline{u}_2^h) = f_1^h - f_2^h$, így alkalmazható az előző állítás a különbségekre. $\square$

Az FDM konvergenciája maximum-normában.

Legyen $u \in C^2(\overline{\Omega})$ a Poisson-egyenlet megoldása és $\overline{u}^h$ az FDM-es közelítő megoldás. A konvergenciavizsgálat során az

$$e^h := u^h - \overline{u}^h$$

hibavektort becsüljük valamilyen normában. Ez a csomópontokbeli eltérést méri, így *csomóponti hibavektornak* hívjuk.

A becslések alapja a következő összefüggés.

2.20. Állítás. *érvényes az alábbi ún. hibaegyenlet, amely szerint az A_h mátrix a csomóponti hibavektort a képlethibavektorba viszi:*

$$A_h e^h = \psi^h. \tag{2.9}$$

Bizonyítás. Mivel $A_h v = -\Delta_h v$ bármely vektorra, így a pontos megoldás u^h csomóponti vektorára is. Ebből és az $A_h \overline{u}^h = f^h$ egyenletből

$$A_h e^h = A_h u^h - A_h \overline{u}^h = -\Delta_h u^h - f^h.$$

Másrészt a pontos megoldásra fennálló $f = -\Delta u$ egyenlőséget a csomópontokban felírva az $f^h = -(\Delta u)^h$ vektoregyenlőséget kapjuk, így

$$A_h e^h = -\Delta_h u^h + (\Delta u)^h =: \psi^h. \qquad \square$$

A hibaegyenlet fő haszna, hogy a képlethibára már meglévő becslésünket érvényesíthetjük.

2.21. Tétel. *Ha $u \in C^4(\overline{\Omega})$, akkor a (2.3) differenciasémával az FDM másodrendben konvergál:*

$$\|e^h\|_\infty \leq \frac{a^2+b^2}{96}|u|_{C^4}\, h^2.$$

Bizonyítás. A hibaegyenlet, a 2.12. és a 2.17. állítások révén

$$\|e^h\|_\infty \leq \|A_h^{-1}\|_\infty \|\psi^h\|_\infty \leq \frac{a^2+b^2}{96}|u|_{C^4}\, h^2. \qquad \square$$

2.22. Megjegyzés. A fentiek és a 2.13. megjegyzések szerint a (2.3) differenciasémára az u simaságának függvényében az alábbiak mondhatók:

$$\begin{array}{lcl} u \in C^4(\overline{\Omega}) & \Rightarrow & \|e^h\|_\infty = O(h^2); \\ u \in C^3(\overline{\Omega}) & \Rightarrow & \|e^h\|_\infty = O(h); \\ u \in C^2(\overline{\Omega}) & \Rightarrow & \|e^h\|_\infty \to 0, \end{array}$$

másrészt magasabbrendű konvergencia is elérhető magasabbrendű differenciasémák és u nagyobb simasága esetén. $\diamondsuit$

Egy modellfeladat véges differenciás megoldását Poisson-egyenlet és homogén Dirichlet-peremfeltétel esetén a 8.2.1. *animáció* mutatja be.

2.4. Stabilitás és hibabecslések diszkrét L^2- és H_0^1-normában

Alapvető fogalmak és becslések

A rácsfüggvényeket mérhetjük az L^2-norma diszkrét megfelelőjével is. Mivel $u \in C(\overline{\Omega})$ esetén

$$\|u\|_0^2 := \|u\|_{L^2}^2 = \int_\Omega u(x)^2 dx_1 dx_2 \approx \sum_{k=1}^N u(x_k)^2 h_1 h_2,$$

ahol az x_k pontok a 2.2. szakaszban definiált rácspontok, így értelemszerű az alábbi fogalom:

2.23. Definíció. Egy $v : \mathbb{R}^N \to \mathbb{R}^N$ függvénynek az ω_h rácshoz tartozó *diszkrét L^2-normája*:

$$\|v\|_{0,h}^2 := \sum_{k=1}^N v_k^2\, h_1 h_2.$$

Ezt a normát a

$$\langle v, z\rangle_{0,h} := \sum_{k=1}^{N} v_k\, z_k\, h_1 h_2$$

skalárszorzat indukálja. ♢

Itt

$$\langle v, z\rangle_{0,h} = h_1 h_2\, v \cdot z, \qquad \|v\|_{0,h}^2 = h_1 h_2\, |v|^2,$$

vagyis az euklideszi skalárszorzat és hossz konstansszorosáról van szó. Megjegyezzük, hogy a diszkrét L^2-norma egy Riemann-összeg, amely folytonos függvény esetén a rács finomításával tart a függvény L^2-normájához. (Ez L^p-re is igaz, lásd a 10.20. lemmát.)

Az FDM (2.7)-beli A_h mátrixával indukálva értelmezzük a H_0^1-norma diszkrét megfelelőjét:

2.24. Definíció. Egy $v : \mathbb{R}^N \to \mathbb{R}^N$ függvénynek az ω_h rácshoz tartozó *diszkrét* H_0^1-*normája*:

$$|v|_{1,h}^2 := \langle A_h v, v\rangle_{0,h}.$$

Ezt a normát a

$$\langle v, z\rangle_{1,0,h} := \langle A_h v, z\rangle_{0,h}$$

skalárszorzat indukálja. ♢

2.25. Megjegyzés. A fenti definíció az $|u|_{1,h}^2 := \int_\Omega |\nabla u|^2 = -\int_\Omega (\Delta u) u \quad (\forall u \in H_0^1(\Omega))$ Green-formulát imitálja. Valóban igazolható, hogy

$$|v|_{1,h}^2 = \|v_{\bar{x}_1}\|_{0,h}^2 + \|v_{\bar{x}_2}\|_{0,h}^2, \tag{2.10}$$

azaz hogy a fent definiált diszkrét H_0^1-norma a változónkénti retrográd első differenciahányadosok diszkrét L^2-normájának összege, lásd 9.9. feladat. A (2.10) képletet szokás diszkrét Green-formulának is hívni. Nyilvánvaló, hogy a diszkrét H_0^1-normát a (2.10) jobboldalával is értelmezhettük volna, sőt ez lehetett volna a természetes definíció. A továbbiakban azonban a diszkrét H_0^1-normára az A_h mátrixszal való összefüggésekben lesz szükség, ezért vezettük be a fenti módon. ♢

Célunk a 2.18. stabilitási állítás megfelelőjének levezetése a fenti normákkal.

2.26. Állítás. *Az A_h mátrix szimmetrikus és pozitív definit.*

Bizonyítás. A szimmetria látható (2.7)-ben, és az is, hogy A_h sor- és oszlopösszegei nemnegatívak, azaz A_h gyengén diagonálisan domináns. Ismeretes (lásd 9.7. feladat), hogy az utóbbiból következik a pozitív szemidefinitség. Mivel a 2.15. állítás szerint A_h reguláris, így csak pozitív definit lehet. □

2.27. Állítás. *Tekintsük az $A_h\overline{u}^h = f^h$ lineáris algebrai egyenletrendszert. Ekkor*

$$|\overline{u}^h|_{1,h} \le \frac{1}{\sqrt{\lambda_0}}\|f^h\|_{0,h},$$

ahol λ_0 az A_h mátrix legkisebb sajátértéke.

Bizonyítás. Mivel A_h szimmetrikus és pozitív definit, így minden v-re $A_hv \cdot v \ge \lambda_0|v|^2$, amit h_1h_2-vel szorozva $\langle A_hv, v\rangle_{0,h} \ge \lambda_0\|v\|_{0,h}^2$, így

$$\|v\|_{0,h} \le \frac{1}{\sqrt{\lambda_0}}|v|_{1,h}$$

(ez a diszkrét Poincaré–Friedrichs-egyenlőtlenség). Ebből

$$|\overline{u}^h|_{1,h}^2 = \langle A_h\overline{u}^h, \overline{u}^h\rangle_{0,h} = \langle f^h, \overline{u}^h\rangle_{0,h} \le \|f^h\|_{0,h}\|\overline{u}^h\|_{0,h} \le \frac{\|f^h\|_{0,h}}{\sqrt{\lambda_0}}|\overline{u}^h|_{1,h},$$

amiből az állítás következik. □

A fenti becslés alapján tehát szükségünk van az A_h mátrix (azaz a diszkrét Laplace-operátor) legkisebb sajátértékére, vagy legalábbis alkalmas alsó becslésére. A téglalap-tartomány révén pontosan meghatározható az összes sajátérték és sajátvektor; mivel ezeket a később másutt is felhasználjuk, most külön foglalkozunk velük.

A diszkrét Laplace-operátor sajátértékei és sajátvektorai téglalapon.

Kiindulásként idézzük fel az egydimenziós esetet, azaz közönséges differenciálegyenletek peremérték-feladatainál FDM-es esetben felmerülő sajátértékeket és sajátvektorokat. Legyen tehát $A_h^{(1)}$ az egydimenziós diszkrét Laplace-operátor, azaz egy $[0, a]$ intervallum $x_i := ih$ ($i = 1, \dots, n$, ahol $h = 1/(n+1)$) belső csomópontjaiban

$$(A_h^{(1)}u^h)_i = -u_{\overline{x}x,i} \qquad (i = 1, \dots, n).$$

Mint ismeretes (lásd pl. [27]), ennek normált sajátvektorai éppen az $u \mapsto -u''$ operátor első n normált sajátfüggvényének csomópontokra vett megszorításaiból képzett vektorok, azaz a

$$v_\ell = \left\{\sqrt{\frac{2}{a}}\,\sin\frac{\ell\pi x_i}{a}\right\}_{i=1,\dots,n} \in \mathbb{R}^n \qquad (\ell = 1, \dots, n)$$

vektorok. Valóban, könnyen látható (lásd 9.10. feladat), hogy ezekre

$$A_h^{(1)}v_\ell = \lambda_\ell^h\, v_\ell, \qquad \text{ahol} \quad \lambda_\ell^h = \frac{4}{h^2}\,\sin^2\frac{\ell\pi h}{2a} \qquad (\ell = 1, \dots, n). \tag{2.11}$$

2.28. Megjegyzés. Érdemes összevetni a fenti λ_ℓ^h sajátértékeket a differenciáloperátor $\lambda_\ell = (\frac{\ell\pi}{a})^2$ „folytonos” sajátértékekeivel: könnyen látható, hogy $\lambda_\ell^h \le \lambda_\ell \ \ (\forall \ell = 1, \dots, n)$. Emellett a következők is fennállnak: amíg a „folytonos” sajátértékek a

$$\frac{\pi^2}{a^2} = \lambda_1 \le \lambda_\ell \le \lambda_n = \left(\frac{n\pi}{a}\right)^2 = \left(\frac{n}{n+1}\right)^2 \frac{\pi^2}{h^2} \approx \frac{\pi^2}{h^2}$$

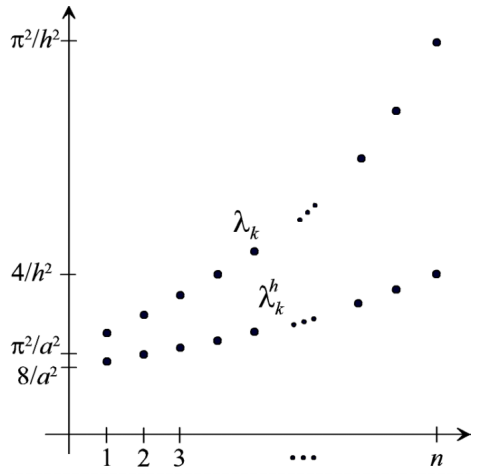
tartományban, addig a diszkrét sajátértékek a

$$\frac{8}{a^2} \lessapprox \frac{4}{h^2} \sin^2 \frac{\pi h}{2a} = \lambda_1^h \le \lambda_\ell^h \le \lambda_n^h = \frac{4}{h^2} \sin^2 \left(\frac{n}{n+1}\frac{\pi}{2}\right) \lessapprox \frac{4}{h^2}$$

tartományban mozognak, lásd 2.6. ábra. Az alábbiakban most főleg a

$$\lambda_1^h \ge \frac{8}{a^2} \tag{2.12}$$

becslést használjuk majd. ♢



2.6. ábra. Folytonos és diszkrét sajátértékek 1D-ben.

Az egydimenziósról a *kétdimenziós esetre* való áttérés teljesen analóg a „folytonos” esettel: az egydimenziós sajátértékek összeadódnak, a sajátvektorok szorzódnak. Ez amiatt igaz, mivel a diszkrét Laplace-operátor az eredetihez hasonlóan az egyváltozós tagok összege. Ebből a kétdimenziós $-\Delta_h$ operátor sajátértékei

$$\lambda_{rs}^h = \frac{4}{h_1^2} \sin^2 \frac{r\pi h_1}{2a} + \frac{4}{h_2^2} \sin^2 \frac{s\pi h_2}{2b} \qquad (r = 1, \dots, n, \ s = 1, \dots, m) \tag{2.13}$$

és normált sajátvektorai

$$v_{rs} = \left\{ \frac{2}{\sqrt{ab}} \sin\frac{r\pi x_i}{a} \sin\frac{s\pi x_j}{b} \right\}_{i=1,\dots,n,\ j=1,\dots,m} \in \mathbb{R}^N \qquad (r=1,\dots,n,\ s=1,\dots,m),$$

ahol a v_{rs} N-dimenziós vektorok ($N=nm$) koordinátáit a fenti nm számnak a kiinduláskor a rácson vett sorfolytonos rendezéséből kapjuk.

Az alábbiakban szükséges alsó becslést (2.12) kétszeri alkalmazásából kapjuk: a legkisebb sajátértékre

$$\lambda_0 := \lambda_{11}^h \geq \frac{8}{a^2} + \frac{8}{b^2} = \frac{8(a^2+b^2)}{a^2b^2}. \tag{2.14}$$

Ebből a 2.27. állítás alapján adódik a stabilitási becslés:

2.29. Állítás. *Tekintsük az $A_h\overline{u}^h = f^h$ lineáris algebrai egyenletrendszert. Ekkor*

$$|\overline{u}^h|_{1,h} \leq \frac{ab}{\sqrt{8(a^2+b^2)}} \|f^h\|_{0,h}.$$

2.30. Megjegyzés.

(i) A maximum-normabeli becsléssel való összevetés az $a=b$ esetben a legszemléletesebb. Ekkor a 2.18. és 2.29. állításokból

$$\|\overline{u}^h\|_\infty \leq \frac{a^2}{8}\|f^h\|_\infty \quad \text{és} \quad |u^h|_{1,h} \leq \frac{a}{4}\|f^h\|_{0,h}.$$

Látható, hogy a most kapott diszkrét Szoboljev-normabecslés kevésbé nő (azaz a stabilitási konstans kevésbé romlik el), amikor az a paramétert növeljük, tehát kevésbé érzékeny a tartomány méretére.

(ii) Ugyanezt a becslést (a maximum-normabeli esettel azonos módon) használhatjuk a konvergencia becslésére is, ha az $A_h\overline{u}^h = f^h$ egyenlet helyett az $A_he^h = \psi^h$ hibaegyenletre alkalmazzuk: ekkor

$$|e^h|_{1,h} \leq \frac{ab}{\sqrt{8(a^2+b^2)}} \|\psi^h\|_{0,h},$$

és itt is igaz, hogy a konstans kevésbé érzékeny a tartomány méretére. A $\|\psi^h\|_{0,h}$ hibanorma becsülhető a $\|\psi^h\|_\infty$ norma konstansszorosával, így szintén $O(h^2)$ rendű, ha $u \in C^4(\overline{\Omega})$: ekkor tehát

$$|e^h|_{1,h} = O(h^2).$$

A Taylor-sorfejtés integrál-maradéktaggal való módosításával itt az $u \in C^4(\overline{\Omega})$ kikötés az $u \in H^4(\Omega)$ feltételre enyhíthető. ♢

2.5. Az FDM általánosabb feladatokra

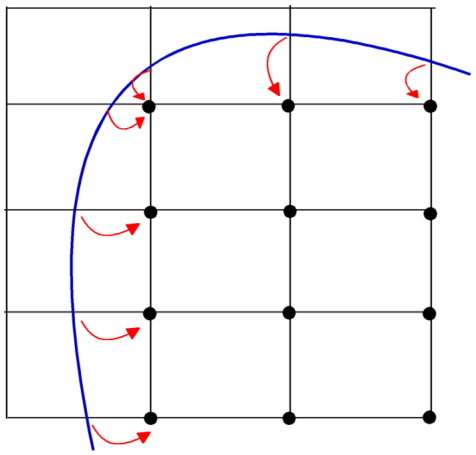
Ebben a szakaszban említés szintjén utalunk arra, hogyan módosítható az előzőekben bemutatott módszer általánosabb feladatokra.

Általános tartomány. Legyen $\Omega \subset \mathbb{R}^2$ szakaszonként sima peremű korlátos tartomány, és tekintsük ezen a (2.1) Poisson-egyenletet.

Kiindulásképpen készíthetünk egy olyan rácsot, amely egy koordinátairányonként ekvidisztáns síkbeli rács azon csomópontjaiból áll, amelyek $\bar{\Omega}$-ban vannak, azaz

$$\Omega_{\mathbf{h}} := \{(ih_1, jh_2) \in \bar{\Omega} : \ i, j \in \mathbb{Z}\}, \tag{2.15}$$

ahol $h_1, h_2 > 0$ adott számok. Ekkor azonban általában a perem nem tartalmaz csomópontokat, így a peremfeltételt más módon kell érvényesíteni. A két leginkább kézenfekvő lehetőség az alábbi. Rendelhetünk a peremhez legközelebb eső csomópontokhoz 0 (perem)értéket (2.7. ábra), ekkor azonban elvész a másodrendű pontosság. Másrészt



2.7. ábra. Peremértékek átvitele a rácspontokba.

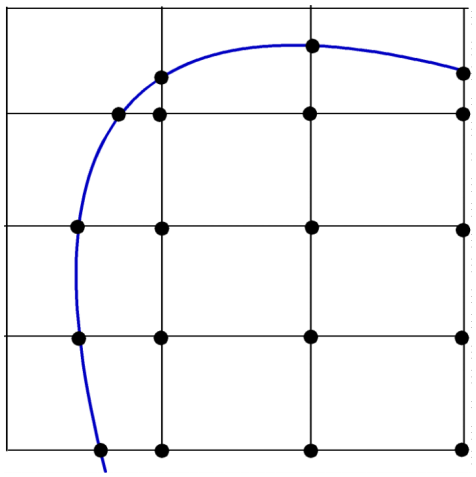
hozzávehetjük a rácshoz a perem és a rács egyeneseinek metszéspontját (2.8. ábra). Itt ez utóbbival foglalkozunk. Legyen tehát

$$\overline{\omega}_h := \Omega_{\mathbf{h}} \cup \{(\xi, \eta) \in \partial\Omega : \ \exists i \in \mathbb{Z},\ \xi = ih_1 \text{ vagy } \exists j \in \mathbb{Z},\ \eta = jh_2\}.$$

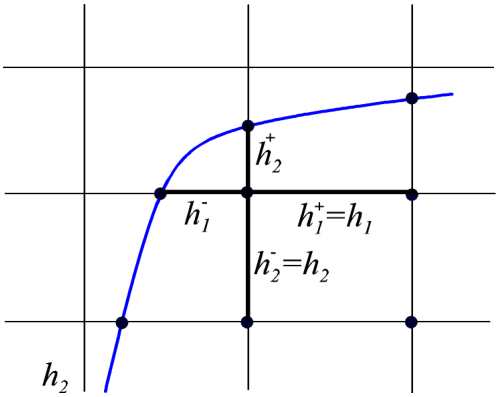
A peremmel nem szomszédos $x_k = (ih_1,\, jh_2)$ csomópontokban $\Delta u(x_k)$ közelítésére most is a (2.3) képlet használható, ami a következő alakban írható:

$$(\Delta_h u^h)_k := \frac{1}{h_1}\left(\frac{u_{i+1,j} - u_{i,j}}{h_1} - \frac{u_{i,j} - u_{i-1,j}}{h_1}\right) + \frac{1}{h_2}\left(\frac{u_{i,j+1} - u_{i,j}}{h_2} - \frac{u_{i,j} - u_{i,j-1}}{h_2}\right). \tag{2.16}$$

Tekintsük most a peremmel szomszédos rögzített rácspontot (lásd 2.9. ábra). Itt a rács



2.8. ábra. Peremértékek hozzávétele a rácshoz.



2.9. ábra. A Shortley–Weller-séma differenciacsillagja.

lokálisan nem ekvidisztáns; jelölje a rögzített pont melletti rácstávolságokat rendre h_1^- és h_1^+, ill. h_2^- és h_2^+. Ezekből az Ω belseje felé eső rácstávolságok a régiek, az ábrán $h_1^+ = h_1$ és $h_2^- = h_2$. Ekkor a $\Delta u(x_k)$ közelítésére alkalmas képlet a fenti helyett most

$$(\Delta_h u^h)_k := \frac{2}{h_1^+ + h_1^-}\left(\frac{u_{i+1,j} - u_{i,j}}{h_1^+} - \frac{u_{i,j} - u_{i-1,j}}{h_1^-}\right) + \frac{2}{h_2^+ + h_2^-}\left(\frac{u_{i,j+1} - u_{i,j}}{h_2^+} - \frac{u_{i,j} - u_{i,j-1}}{h_2^-}\right),$$

ez az ún. *Shortley–Weller-approximáció*. Ennek differenciacsillagját a 2.9. ábra mutatja.

A konvergencia becslésénél ekkor az a nehézség merül fel, hogy a peremközeli pontoknál a Taylor-sorfejtésben nem esnek ki az elsőrendű tagok, mivel $u_{i+1,j}$ és $u_{i-1,j}$ (ill. $u_{i,j+1}$ és $u_{i,j-1}$) együtthatói nem egymás (-1)-szeresei. Így ezekben a pontokban a képlethiba rendje csak $O(h)$ és nem $O(h^2)$. Ez azonban a csomóponti hibákra mégsem rontja el a másodrendű konvergenciát. A (2.9) egyenlőség révén ugyanis $e^h = A_h^{-1}\psi^h$, ahol A_h^{-1} a

Green-függvény közelítésének felel meg (lásd a 2.16. megjegyzést). Ezért A_h^{-1} peremközeli pontoknak megfelelő elemei kicsik, és kompenzálják ψ^h nagyobb értékeit; összességében ekkor is igazolható az $u \in C^4(\overline{\Omega})$ esetre az $O(h^2)$ konvergenciarend.

A fenti Shortley–Weller-séma mátrixa nem szimmetrikus. Jelölje pl. $x_k = (ih_1,\, jh_2)$ a peremmel szomszédos rögzített rácspontot és legyen $x_{k+1} := ((i+1)h_1,\, jh_2)$. Ekkor a mátrix $a_{k,k+1}$ és $a_{k+1,k}$ elemei különbözőek, ui. $u_{i+1,j}$ együtthatóját csak a peremközeli ponthoz tartozó sémában módosítottuk. Ezért szokás bevezetni az ún. szimmetrizált Shortley–Weller-sémát:

$$(\Delta_h u^h)_k := \frac{1}{h_1}\Big(\frac{u_{i+1,j}-u_{i,j}}{h_1^+} - \frac{u_{i,j}-u_{i-1,j}}{h_1^-}\Big) + \frac{1}{h_2}\Big(\frac{u_{i,j+1}-u_{i,j}}{h_2^+} - \frac{u_{i,j}-u_{i,j-1}}{h_2^-}\Big).$$

Inhomogén peremfeltétel. Tekintsük a Poisson-egyenletet inhomogén peremfeltétellel. Ha nem homogenizáljuk a feladatot az 1.6. megjegyzésben leírt módon, akkor a (2.8)-beli $A_h\overline{u}^h = f^h$ lineáris algebrai egyenletrendszer helyett egy nagyobb rendszerhez jutunk, melyben a peremre eső csomópontok is szerepelnek. Rendezzük úgy a csomópontokat, hogy előbb a belső, azután a perempontokat soroljuk fel. Ekkor a kapott lineáris algebrai egyenletrendszer az alábbi alakú lesz:

$$\begin{bmatrix} A_h & \tilde{A}_h \\ 0 & I \end{bmatrix} \begin{bmatrix} \overline{u}^h \\ \widetilde{u}^h \end{bmatrix} = \begin{bmatrix} f^h \\ g^h \end{bmatrix}, \tag{2.17}$$

ahol $\overline{u}^h$ tartalmazza a belső, $\widetilde{u}^h$ pedig a peremen lévő csomóponti értékeket, azaz a rendszer

$$\begin{cases} A_h\overline{u}^h + \tilde{A}_h\widetilde{u}^h = f^h, \\ \widetilde{u}^h = g^h. \end{cases}$$

Könnyen felírható az $\tilde{A}_h$ mátrix, lásd 9.12. feladat. A második egyenlőség révén valójában az

$$A_h\overline{u}^h = f^h - \tilde{A}_h g^h$$

feladatot kell megoldanunk, ez pedig épp a homogenizált Poisson-egyenletnek megfelelő lineáris algebrai egyenletrendszer , mátrixa a (2.8)-beli A_h mátrix. A diszkrét feladat így tehát természetes módon redukálható a homogenizált esetre.

Vegyes peremfeltétel. Tekintsük a Poisson-egyenletet az $\Omega :=]0,a[\times]0,b[$ téglalapon vegyes (Robin-féle) peremfeltételel, vagyis a

$$\begin{cases} -\Delta u = f, \\ \partial_\nu u_{\,|\partial\Omega} = \sigma(u_0 - u) \end{cases} \tag{2.18}$$

peremérték-feladatot. Ekkor a peremfeltétel normális irányú deriváltat is tartalmaz, ezt is differenciasémával közelítjük. A Dirichlet-peremfeltétel esetén elérhető $O(h^2)$ konvergenciarend megőrzésére a normális derivált differenciasémáját is $O(h^2)$ rendűre célszerű választani. Bizonyítás nélkül írjuk fel a megfelelő sémákat (lásd [27]):

- ha az $(n_1h_1,\, jh_2)$ pont a téglalap jobb oldali függőleges élének belső pontja, akkor a séma
$$\frac{u_{n_1,j}-u_{n_1-1,j}}{h_1}=\sigma_{n_1,j}((u_0)_{n_1,j}-u_{n_1,j})-\frac{h_1}{2}\left(-\frac{u_{n_1,j+1}-2u_{n_1,j}+u_{n_1,j_1}}{h_2^2}-f_{n_1,j}\right),$$
és ennek értelemszerű megfelelőit vesszük a többi él belső pontjában;

- a $(0,b)$ csúcsban
$$\left(-2\frac{u_{0,n_2}-u_{0,n_2-1}}{h_2^2}+2\frac{u_{1,n_2}-u_{0,n_2}}{h_1^2}\right)H=-\sigma_{0,n_2}((u_0)_{0,n_2}-u_{0,n_2})-Hf_{0,n_2},$$
ahol $b=n_2h_2$, $H:=\frac{h_1h_2}{2(h_1+h_2)}$, és ennek értelemszerű megfelelőit vesszük a többi csúcsban.

Két modellfeladat véges differenciás megoldását Poisson-egyenlet és vegyes (Robin-féle) peremfeltétel esetén a 8.2.2. és a 8.2.3. *animációk* mutatják be. Az egyik feladatnál a pontos megoldást is ismerjük, így méginkább szemléletes a konvergencia.

Függvényegyütthatós PDE. Tekintsük a
$$-\operatorname{div}(p\nabla u)=f$$
egyenletet homogén Dirichlet-peremfeltétellel, ahol $p\in C^1(\overline{\Omega})$, $p(x)\geq m>0$ $(\forall x\in\Omega)$. Ismét a Poisson-egyenlet közelítéséből indulunk ki a (2.16) alakban. Itt a zárójelekben lévő négy elsőrendű differenciahányadost a $\operatorname{div}(p\nabla u)$ képlet miatt most súlyozni kell p alkalmas értékeivel, és a cél a Poisson-egyenletre elérhető $O(h^2)$ konvergenciarend megőrzése. Tekintsük a négyből az első, azaz az (ih_1,jh_2) ponthoz tartozó haladó x_1 irányú differenciahányadost, ez csak elsőrendben konvergál az (ih_1,jh_2) pontban, ha $h_1\to 0$. Ugyanez tekinthető viszont az $((i+1/2)h_1,jh_2)$ ponthoz tartozó $h/2$ lépésközű centrális differenciahányadosnak:
$$\frac{u_{i+1,j}-u_{i,j}}{h_1}=\frac{u_{(i+1/2)+1/2,j}-u_{(i+1/2)-1/2,j}}{h_1},$$
amely az $((i+1/2)h_1,jh_2)$ pontban már másodrendben konvergál, ha $h_1\to 0$ (lásd 2.1. megjegyzés). Emiatt ezt a differenciahányadost a $p_{i+1/2,j}:=p((i+1/2)h_1,jh_2)$ függvényértékkel szokás súlyozni. Ugyanez elmondható a többi tagra is. Ezekből adódik $\operatorname{div}(p\nabla u)$ közelítésére az
$$\frac{1}{h_1}\left(p_{i+1/2,j}\frac{u_{i+1,j}-u_{i,j}}{h_1}-p_{i-1/2,j}\frac{u_{i,j}-u_{i-1,j}}{h_1}\right)$$
$$+\frac{1}{h_2}\left(p_{i,j+1/2}\frac{u_{i,j+1}-u_{i,j}}{h_2}-p_{i,j-1/2}\frac{u_{i,j}-u_{i,j-1}}{h_2}\right)$$
differenciaséma. Erre valóban igazolható elég sima p és u esetén az $O(h^2)$ konvergenciarend.

Végül megemlítjük, hogy a téglalapra vagy a fentiekben általánosabb síkbeli tartományra alkalmazott FDM értelemszerű módosításokkal átvihető a térbeli esetre.

3. fejezet

A végeselem-módszer (FEM)

A végeselem-módszer (angolul „finite element method”, betűszóként az ennek megfelelő FEM-et fogjuk használni) lényege, hogy a PDE megoldását „szakaszonként” polinommal közelítjük, azaz olyan függvénnyel, amely az Ω tartománynak véges sok résztartományra való alkalmas felbontása mellett a résztartományokra leszűkítve egy-egy polinom. A résztartományok általában sokszögek/poliéderek (ezen belül síkon háromszögek vagy téglalapok, ill. térben tetraéderek vagy téglatestek), és a közelítő megoldást folytonosnak konstruáljuk az egész tartományon, azaz a résztartományok találkozásánál is.

Mivel az ilyen közelítő megoldások nem differenciálhatóak az egész tartományon a triviális speciális esetektől eltekintve, így a gyenge megoldásból kiindulva konstruáljuk meg őket. Ezért érdemes először absztrakt esetben, bilineáris formával megadott feladatra felírni a módszert, mert ez megvilágítja elvi hátterét. Ezt nevezzük *Galjorkin-módszernek*.

A véges differenciákkal ellentétben az elméletben nem szükséges külön vizsgálnunk a Poisson-egyenletet, és a tartomány sem kell speciális (téglalap) legyen. Elsősorban (1.13) alakú függvényegyütthatós feladatokkal foglalkozunk majd:

$$\begin{cases} -\operatorname{div}(p\,\nabla u) = f, \\ u_{|\partial\Omega} = 0, \end{cases}$$

ahol $\Omega \subset \mathbb{R}^N$ korlátos tartomány, $p \in L^\infty(\Omega)$ és alkalmas $m > 0$ esetén $p(x) \geq m > 0$ (m-m. $\forall x \in \Omega$), valamint $f \in L^2(\Omega)$. Kitérünk majd emellett általánosabb, alsóbbrendű tago(ka)t tartalmazó egyenlet és vegyes, ill. Neumann-peremfeltétel esetére is.

Az egész fejezetben $\Omega \subset \mathbb{R}^N$ korlátos, szakaszonként sima peremű tartomány (lásd 1.1. definíció).

3.1. Az FEM elméleti alapja: a Galjorkin-módszer

3.1.1. A Galjorkin-módszer értelmezése bilineáris formára

Legyen H valós Hilbert-tér, $a : H \times H \to \mathbb{R}$ korlátos, koercív bilineáris forma, azaz

(i) létezik $M > 0$, hogy

$$|a(u, v)| \leq M\|u\| \, \|v\| \qquad (\forall u, v \in H); \tag{3.1}$$

(ii) létezik $m > 0$, hogy

$$a(u, u) \geq m\|u\|^2 \qquad (\forall u \in H), \tag{3.2}$$

valamint legyen $\ell : H \to \mathbb{R}$ korlátos lineáris funkcionál. Tekintsük az alábbi ún. variációs feladatot:

$$a(u, v) = \ell v \qquad (\forall v \in H). \tag{3.3}$$

Ennek a Lax–Milgram-lemma (lásd 1.2. tétel) szerint létezik egyetlen $u \in H$ megoldása. Tegyük fel, hogy az a bilineáris forma szimmetrikus is:

$$a(u, v) = a(v, u) \qquad (\forall u, v \in H).$$

Az 1.4. megjegyzés szerint ui. az (1.13) és (1.21) egyenletek gyenge alakjához tartozó bilineáris formák szimmetrikusak is.)

A Galjorkin-módszerben a fenti feladat közelítő megoldását alkalmasan választott véges dimenziós alterekben keressük. Ezeket az altereket V_h-val jelöljük, ahol $h > 0$ paraméter. (A h érték a végeselem-módszerben a rács finomságát fogja kifejezni. Az absztrakt szinten rögzített altér esetén ez egyelőre csak egy jelölés, később az alterek egy családja esetén azt tesszük majd fel, hogy $h \to 0$ esetén az alterek jól közelítik az egész teret, lásd 3.9. állítás.)

Legyen tehát most

$$V_h \subset H$$

adott véges dimenziós altér, ebben szeretnénk értelmezni a közelítő megoldást.

A közelítő megoldás értelmezése vetületi egyenlettel. Legyen $u_h \in V_h$ az az elem, amelyre (3.3) csak V_h-beli tesztfüggvények mellett teljesül, azaz

$$a(u_h, v_h) = \ell v_h \qquad (\forall v_h \in V_h). \tag{3.4}$$

A Lax–Milgram-lemma a V_h altérben is alkalmazható, mivel a (3.1) és (3.2) egyenlőtlenségek speciálisan $u := u_h \in V_h$ esetén is fennállnak. Így a fenti V_h-beli feladatnak, az ún. *vetületi egyenletnek* is létezik egyetlen $u_h \in V_h$ megoldása.

Az u_h elem konstrukcióját a V_h altér egy adott

$$\varphi_1, \dots, \varphi_n$$

bázisa segítségével végezhetjük el. Ekkor u_h-t a báziselemek lineáris kombinációjaként keressük:

$$u_h = \sum_{j=1}^{n} c_j \varphi_j.$$

A c_j együtthatók meghatározásához először írjuk fel a (3.4) vetületi egyenletet úgy, hogy tesztfüggvénynek a $v_h := \varphi_i$ $(i = 1, \dots, n)$ báziselemeket választjuk:

$$a(u_h, \varphi_i) = \ell\varphi_i \qquad (i = 1, \dots, n).$$

Ezután helyettesítsük be a fenti összeget: itt

$$a(u_h, \varphi_i) = a\Big(\sum_{j=1}^{n} c_j \varphi_j, \varphi_i\Big) = \sum_{j=1}^{n} a(\varphi_j, \varphi_i)\, c_j,$$

így

$$\sum_{j=1}^{n} a(\varphi_j, \varphi_i)\, c_j = \ell\varphi_i \qquad (i = 1, \dots, n).$$

Ez egy lineáris algebrai egyenletrendszer. Jelölje

$$a_{ij} := a(\varphi_j, \varphi_i), \qquad b_i := \ell\varphi_i \qquad (i, j = 1, \dots, n)$$

a megfelelő együtthatókat és jobboldalakat, itt az a forma szimmetriája miatt

$$a_{ij} = a(\varphi_j, \varphi_i) = a(\varphi_i, \varphi_j) = a_{ji}.$$

A fenti lineáris algebrai egyenletrendszer tömören

$$A_h c = b_h \tag{3.5}$$

alakban írható, ahol

$$A_h := \ \{a_{ij}\}_{i,j=1,\dots,n} = \ \big\{a(\varphi_i, \varphi_j)\big\}_{i,j=1,\dots,n}$$

az ún. merevségi mátrix (angolul 'stiffness matrix'), és

$$b_h := \{b_i\}_{i=1,\dots,n} = \ \big\{\ell\varphi_i\big\}_{i=1,\dots,n}.$$

3.1. Állítás. *Az A_h mátrix szimmetrikus és pozitív definit.*

Bizonyítás. A szimmetriát már beláttuk, hiszen $a_{ij} = a_{ji}$.

Az A_h mátrix és az a forma kapcsolatát meghatározó összefüggés a következő. Legyenek

$$u_h = \sum_{j=1}^{n} c_j \varphi_j, \qquad v_h = \sum_{j=1}^{n} d_j \varphi_j$$

tetszőleges elemei V_h-nak, és legyen $c, d \in \mathbb{R}^N$ rendre a c_i és d_i együtthatókból alkotott vektor. Ekkor

$$a(u_h, v_h) = a\Big(\sum_{j=1}^{n} c_j \varphi_j,\ \sum_{i=1}^{n} d_i \varphi_i\Big) = \sum_{i,j=1}^{n} a(\varphi_j,\ \varphi_i) c_j d_i = \sum_{i,j=1}^{n} a_{ij} c_j d_i = A_h c \cdot d.$$

Ebből speciálisan

$$A_h c \cdot c = a(u_h, u_h) > 0 \tag{3.6}$$

az a forma koercivitása miatt, így A_h pozitív definit. □

3.2. Következmény. *A (3.5) lineáris algebrai egyenletrendszernek egyértelműen létezik $c \in \mathbb{R}^N$ megoldása.*

A szimmetria és pozitív definitség révén ráadásul a lineáris algebrai egyenletrendszer hatékonyan megoldható, ezzel a 4. fejezetben foglalkozunk majd. A lineáris algebrai egyenletrendszer egyértelmű megoldhatóságából is következik a (3.4) vetületi egyenlet egyértelmű megoldhatósága, amit annak definíciójánál közvetlenül is indokoltunk.

A közelítő megoldás értelmezése minimalizáló funkcionállal. A V_h-beli Galjorkin-féle közelítő megoldás alternatív módon alkalmas minimalizáló funkcionál segítségével is bevezethető. Az 1.1. szakaszban definiált energiafunkcionál mintájára tekintsük a

$$\Phi : H \to \mathbb{R}, \qquad \Phi(u) := \frac{1}{2}\, a(u,u) - \ell u \quad (u \in H)$$

funkcionált. Ekkor érvényes az

3.3. Állítás. *Ha $u^* \in H$ jelöli a (3.3) variációs feladat megoldását, akkor*

$$\min_{u \in H} \Phi(u) = \Phi(u^*),$$

és ez szigorú minimumhely.

Bizonyítás. Mivel most

$$a(u^*, v) = \ell v \qquad (\forall v \in H),$$

így bármely $u \in H$ esetén

$$\Phi(u) - \Phi(u^*) = \frac{1}{2}\, a(u,u) - \ell u - \frac{1}{2}\, a(u^*,u^*) + \ell u^* = \frac{1}{2}\, a(u,u) - a(u^*,u) - \frac{1}{2}\, a(u^*,u^*) + a(u^*,u^*)$$

$$= \frac{1}{2}\, a(u,u) - a(u^*,u) + \frac{1}{2}\, a(u^*,u^*).$$

Felhasználva az a forma szimmetriáját,

$$a(u^*,u) = \frac{1}{2}\, a(u^*,u) + \frac{1}{2}\, a(u,u^*),$$

így

$$\Phi(u) - \Phi(u^*) = \frac{1}{2}\, a(u,u) - \frac{1}{2}\, a(u^*,u) - \frac{1}{2}\, a(u,u^*) + \frac{1}{2}\, a(u^*,u^*) = \frac{1}{2}\, a(u-u^*, u-u^*) > 0$$

bármely $u \neq u^*$ esetén a koercivitás miatt. □

Azaz a variációs feladat megoldása egyúttal az energiafunkcionál szigorú minimumhelye is. Ebből természetes módon adódik a V_h-beli közelítő megoldás értelmezése: tekintsük V_h azon elemét, amely szigorú minimumhelye a V_h-ra leszűkített energiafunkcionálnak. A H tér minimális energiájú tagját tehát a V_h altér minimális energiájú tagjával közelítjük.

Könnyen látható, hogy ez a definíció ugyanazt adja vissza, mint a vetületi egyenlet:

3.4. Állítás. *A (3.4) vetületi egyenlet megoldása egyben a $\Phi_{|V_h}$ leszűkített funkcionál szigorú minimumhelye.*

Bizonyítás. Megegyezik a 3.3. állítás bizonyításával, ha u^* helyére a vetületi egyenlet megoldását, tetszőleges $u \in H$ helyére pedig V_h tetszőleges elemét írjuk. □

A közelítő megoldás hibájának Galjorkin-ortogonalitása. A fentiekben értelmezett $u_h \in V_h$ elem egy további tulajdonsággal rendelkezik, amely ráadásul ekvivalens akár a vetületi egyenlettel, akár a minimális energia feltételével. Ez a tulajdonság az $u^* - u_h$ hibavektor V_h-ra való a-ortogonalitását mondja ki, vagyis itt az ortogonalitást az a (szimmetrikus és pozitív definit) bilineáris forma által definiált skalárszorzatra nézve értjük, és ezt Galjorkin-ortogonalitásnak hívjuk.

3.5. Állítás. *Jelölje $u^* \in H$ a (3.3) variációs feladat megoldását. Az $u_h \in V_h$ elem pontosan akkor a (3.4) vetületi egyenlet megoldása, ha*

$$a(u^* - u_h, v_h) = 0 \qquad (\forall v_h \in V_h). \tag{3.7}$$

Bizonyítás. A (3.3) egyenlőség speciálisan minden $v := v_h \in V_h$ elemre is fennáll, azaz

$$a(u^*, v_h) = \ell v_h \qquad (\forall v_h \in V_h). \tag{3.8}$$

Tegyük fel először, hogy $u_h \in V_h$ teljesíti a (3.7) ortogonalitási egyenlőséget. Ekkor a (3.8) egyenlet és az ortogonalitási egyenlőség különbsége épp a vetületi egyenlet, így u_h annak is megoldása. Megfordítva, ha u_h a vetületi egyenlet megoldása, akkor ezt a (3.8)-ből kivonva épp az ortogonalitási egyenlőséget kapjuk. □

A hibavektor V_h-ra való a-ortogonalitása azt jelenti, hogy u_h a V_h altérnek az u^* megoldáshoz legközelebbi eleme az a-skalárszorzatra nézve. Ez a tulajdonság és a vetületi egyenlet egyaránt azt tükrözi, hogy u_h éppen az u^* megoldásnak a V_h altérre való vetülete az a-skalárszorzatra nézve.

3.6. Megjegyzés. A továbbiakban a Galjorkin-módszer értelmezésénél, ill. konstrukciójánál a vetületi egyenletet használjuk. Ez a kézenfekvő mód, és a nem szimmetrikus esetre is ez vihető át.

Megemlítjük a fentiek egy lehetséges általánosítását, az ún. Petrov-Galjorkin-módszert, lásd pl. [2]: a közelítő megoldást és a tesztelemeket más altérből is lehet választani, azaz ha V_h és W_h adott alterek, akkor $u_h \in V_h$ az az elem, amelyre

$$a(u_h, v_h) = \ell v_h \qquad (\forall v_h \in W_h). \tag{3.9}$$

♢

3.1.2. Céa-lemma, kvázioptimalitás és konvergencia

Az imént kapott eredmény azt jelenti, hogy u_h teljesíti az

$$\|u^* - u_h\|_a = \min_{v_h \in V_h} \|u^* - v_h\|_a$$

optimalitási tulajdonságot az $\|u\|_a := a(u,u)^{1/2}$ normára nézve, hiszen ebben mérve u_h a V_h altérnek az u^* megoldáshoz legközelebbi eleme.

Gyakran azonban ehelyett az eredeti normában szeretnénk becslést látni a hibavektorra. Erre vonatkozik a Galjorkin-módszer konvergenciavizsgálatának alaptétele, a Céa-lemma néven nevezetessé vált állítás (amely szintén a Galjorkin-ortogonalitásból következik).

3.7. Állítás. ***(„Céa-lemma”)*** *Az $u_h \in V_h$ Galjorkin-megoldásra*

$$\|u^* - u_h\| \le \frac{M}{m} \inf_{v_h \in V_h} \|u^* - v_h\|.$$

Bizonyítás. Legyen $v_h \in V_h$ tetszőleges. Alkalmazzuk a koercivitást, a (3.7) ortogonalitási egyenlőséget v_h helyett $(u_h - v_h)$-ra, végül a korlátosságot:

$$\begin{aligned} m\|u^* - u_h\|^2 &\leq a(u^* - u_h, u^* - u_h) = a(u^* - u_h, u^* - v_h) \\ &\leq M\|u^* - u_h\|\|u^* - v_h\|. \end{aligned} \tag{3.10}$$

Ebből $\|u^* - u_h\| \leq \frac{M}{m}\|u^* - v_h\|$. Ez igaz minden v_h-ra, így az infimumra is. □

3.8. Megjegyzés. A fenti infimum valójában most is minimum, mert véges dimenziós altértől vett távolság. Ezért értelmes a

$$dist(u^*, V_h) := \min_{v_h \in V_h} \|u^* - v_h\|$$

jelölés, és ezzel a Céa-lemma állítása

$$\|u^* - u_h\| \leq \frac{M}{m}\, dist(u^*, V_h).$$

Az eredmény azt jelenti, hogy az eredeti normában a hibavektor nagysága a V_h-tól való távolság konstansszorosával becsülhető felülről. Ezt nevezzük kvázioptimalitási tulajdonságnak. Ez annyiban általánosabb is, mint az a-normára vett pontos optimalitás, hogy a Céa-lemma nem szimmetrikus bilineáris formára is ugyanígy igazolható. (A bizonyításban csak a koercivitást és korlátosságot használtuk, szimmetriát nem.) ♢

A konvergenciához a fentiek alapján olyan alterekre van szükség, melyek egyre közelebb vannak bármely előre megadott vektorhoz, tehát ebben az értelemben jól approximálják a H teret.

3.9. Állítás. *Legyen $\{V_h : \ h > 0\}$ a H tér véges dimenziós altereiből álló család. Ha fennáll az az approximációs tulajdonság, hogy*

$$\forall u \in H \quad dist(u, V_h) \to 0 \qquad (ha \quad h \to 0),$$

akkor a variációs egyenlet V_h alterekben vett Galjorkin-megoldásaira

$$\|u^* - u_h\| \to 0 \qquad (ha \quad h \to 0).$$

Bizonyítás. A Céa-lemma és az u^*-ra alkalmazott approximációs tulajdonság következménye. □

3.2. Az FEM konstrukciója

3.2.1. Általános konstrukció szimmetrikus elliptikus feladatra

Ebben a pontban alkalmazzuk elliptikus feladatra az előbbiekben ismertetett Galjorkin-módszert. Látható lesz, hogy a módszer tényleges megvalósítása attól függ, milyen V_h véges dimenziós altereket (és ezekben $\varphi_1, \dots, \varphi_n$ bázist) választunk. Ennek konkrét lehetőségeiről a következő pontban lesz szó.

Homogén Dirichlet-feladat. Tekintsük a

$$\begin{cases} -\operatorname{div}(p\,\nabla u) = f, \\ u_{|\partial\Omega} = 0 \end{cases} \tag{3.11}$$

Dirichlet-feladatot, ahol $\Omega \subset \mathbb{R}^N$ korlátos tartomány, $p \in L^\infty(\Omega)$ és alkalmas $m > 0$ esetén $p(x) \geq m > 0$ (m-m. $\forall x \in \Omega$), valamint $f \in L^2(\Omega)$. Az 1.2. szakasz szerint e feladatnak egyértelműen létezik gyenge megoldása: $u \in H_0^1(\Omega)$, melyre

$$\int_\Omega p\ \nabla u \cdot \nabla v = \int_\Omega f v \qquad (\forall v \in H_0^1(\Omega)). \tag{3.12}$$

Ennek háttere a megfelelő

$$a(u,v) := \int_\Omega p\ \nabla u \cdot \nabla v \qquad (u, v \in H_0^1(\Omega)) \tag{3.13}$$

bilineáris forma koercivitása és korlátossága a p-re tett feltételeknek köszönhetően, ill. az

$$\ell v := \int_\Omega f v \qquad (v \in H_0^1(\Omega)) \tag{3.14}$$

lineáris funkcionál korlátossága.

A fentiekben leírt Galjorkin-módszer azt jelenti, hogy tekintünk egy alkalmas

$$V_h \subset H_0^1(\Omega)$$

véges dimenziós alteret, és ebben keressük a vetületi egyenlet megoldását: $u_h \in V_h$, melyre

$$\int_\Omega p\ \nabla u_h \cdot \nabla v_h = \int_\Omega f v_h \qquad (\forall v_h \in V_h). \tag{3.15}$$

Ha $\varphi_1, \dots, \varphi_n$ bázis V_h-ban és a közelítő megoldást

$$u_h = \sum_{j=1}^{n} c_j \varphi_j \tag{3.16}$$

alakban keressük, akkor a megfelelő

$$A_h c = b_h$$

lineáris algebrai egyenletrendszerben

$$a_{ij} = \int_\Omega p\ \nabla\varphi_i \cdot \nabla\varphi_j \qquad \text{és} \qquad b_i = \int_\Omega f\varphi_i \qquad (i,j = 1,\dots,n).$$

A lineáris algebrai egyenletrendszer egyértelmű megoldhatóságát a 3.2. következményben láttuk. A $c = (c_1,\dots,c_n)$ együtthatóvektor kiszámítása után (3.16) megadja a V_h altérbeli végeselemes megoldást.

Vegyes peremfeltétel. Tekintsük most az (1.21)-beli vegyes feladatot:

$$\begin{cases} -\operatorname{div}(p\,\nabla u) + qu = f, \\ u_{|\Gamma_D} = 0, \quad (p\,\partial_\nu u + su)_{|\Gamma_N} = \gamma, \end{cases} \tag{3.17}$$

az ott tett feltételekkel. A gyenge megoldás (1.22) szerint olyan $u \in H^1_D(\Omega)$, melyre

$$\int_\Omega (p\ \nabla u \cdot \nabla v + quv) + \int_{\Gamma_N} suv = \int_\Omega fv + \int_{\Gamma_N} \gamma v \qquad (\forall v \in H^1_D(\Omega)).$$

Ennek létezését és egyértelműségét most az

$$a(u,v) := \int_\Omega (p\ \nabla u \cdot \nabla v + quv) + \int_{\Gamma_N} suv \qquad (\forall u, v \in H^1_D(\Omega)) \tag{3.18}$$

bilineáris forma koercivitása és korlátossága biztosítja, amely a feltételekből következik.

A megfelelő $A_h c = b_h$ lineáris algebrai egyenletrendszerben most

$$a_{ij} = \int_\Omega (p\ \nabla\varphi_i \cdot \nabla\varphi_j + q\varphi_i\varphi_j) + \int_{\Gamma_N} s\varphi_i\varphi_j) \qquad \text{és} \qquad b_i = \int_\Omega f\varphi_i \qquad (i,j = 1,\dots,n).$$

Neumann-feladat. Az (1.23) feladat esetén, mint ott láttuk, a gyenge alak és egyben a bilineáris forma megegyezik a (3.12) egyenlőséggel, így az A_h mátrix elemei is: a különbség, hogy az ebben szereplő alteret, ill. bázisfüggvényeket a $\dot{H}^1(\Omega)$ faktorizált térben vesszük.

Inhomogén Dirichlet-feladat. Ennek gyenge alakját az 1.6. megjegyzésben láttuk: az $u \in H^1(\Omega)$ megoldás olyan függvény, melyre

$$\int_\Omega p\ \nabla u \cdot \nabla v = \int_\Omega fv \qquad (\forall v \in H^1_0(\Omega))$$

(azaz a tesztfüggvények csak homogén peremfeltételűek), és

$$u_{|\partial\Omega} = g \quad \text{nyom-értelemben} \qquad (\Leftrightarrow \quad u - \widetilde{g} \in H_0^1(\Omega)).$$

Ekkor a V_h altérhez olyan bázist kell vennünk, mely homogén és inhomogén peremfeltételű tagokból áll, azaz (a korábbi $\varphi_1, \ldots, \varphi_n$ jelölést megtartva a $H_0^1(\Omega)$-beli bázisfüggvényekre) most további $\varphi_{n+1}, \ldots, \varphi_{n+n^\partial}$ bázisfüggvényeket is beveszünk a perem közelítésre. Így a bázis most

$$\varphi_1, \ldots, \varphi_n, \ \varphi_{n+1}, \ldots, \varphi_{n+n^\partial}.$$

Ekkor a végeselemes megoldást

$$u_h = \sum_{j=1}^{n} c_j \varphi_j + \sum_{j=n+1}^{n+n^\partial} c_j \varphi_j$$

alakban keressük, ahol a két tag megfelel az 1.6. megjegyzésbeli z és $\widetilde{g}$ függvényeknek, vagyis itt az $u = z + \widetilde{g}$ előállítást közelítjük. Itt tipikusan a $\varphi_{n+1}, \ldots, \varphi_{n+n^\partial}$ függvények nullák alkalmas perempontok egy környezetén kívül, így a szumma második része lényegében g közelítéséből adódik a peremen, amelyhez nincs szükség egy $\widetilde{g}$ beterjesztés kiszámítására.

3.2.2. Véges elemek és típusaik

Mint láttuk, a módszer tényleges megvalósítása attól függ, milyen V_h véges dimenziós altereket (és ezekben $\varphi_1, \ldots, \varphi_n$ bázist) választunk. A végeselem-módszer lényege, hogy ezek az alterek „szakaszonként" polinomokból állnak, a résztartományok sokszögek ill. poliéderek, és a közelítő megoldást folytonosnak konstruáljuk az egész tartományon.

A továbbiakban legyen $d = 2$ vagy 3, és feltesszük, hogy maga az $\Omega \subset \mathbb{R}^d$ adott korlátos tartomány is sokszög (2D-ben) ill. poliéder (3D-ben). (Ha ez nem áll fenn, akkor Ω közelíthető sokszöggel ill. poliéderrel, ami a felbontás finomítása során nem rontja el a konvergenciát.)

A következőképp értelmezzük a tartomány felbontását:

3.10. Definíció. Az Ω tartomány *triangulációjának* nevezzük a

$$\mathcal{T}_h := \{T_1, \ldots, T_M\}$$

halmazt, ahol

(i) minden $k = 1, \ldots, M$ esetén $T_k \subset \Omega$ sokszög (2D-ben) ill. poliéder (3D-ben);

(ii) $T_1, \ldots, T_M$ a tartomány nemátfedő felbontását alkotja, azaz az $\operatorname{int} T_k$ halmazok páronként diszjunktak és $\cup T_k = \overline{\Omega}$;

(iii) a felbontás konform, azaz a $T_k \cap T_\ell$ $(k \neq \ell)$ halmazok csak csúcsokból vagy teljes oldalakból (élekből vagy lapokból) állhatnak. ♢

(Megjegyezzük, hogy a résztartományok elvben értelmezhetők a sokszögeknél/poliédereknél általánosabban, görbe határral is, lásd 3.13. megjegyzés.)

3.11. Definíció. A $\mathcal{T}_h$ trianguláció *finomsága* a fellépő legnagyobb átmérő:

$$h := \max_{k=1,\ldots,M} diam(T_k).$$ ♢

A V_h altér „szakaszonként" polinomokból áll, melyek folytonosak az egész tartományon, azaz

$$V_h \subset \{u \in C(\overline{\Omega}) : \ u_{|T_k} \in P^{\ell_k}(T_k) \ \forall k = 1, \ldots, M\},$$

ahol $P^{\ell_k}(T_k)$ jelöli a legfeljebb ℓ_k-adfokú polinomok T_k-ra való leszűkítésének halmazát.

Általában az alábbi tulajdonságú altereket szokás használni:

- A T_k halmazok azonos típusúak, vagyis egy adott trianguláció csupa háromszögből/tetraéderből vagy csupa négyszögből/téglatestből áll (rendre a 2D/3D esetben).
- $\ell_k \equiv \ell$, vagyis minden résztartományon azonos fokú polinomokat tekintünk (legegyszerűbb az $\ell = 1$ eset, magasabb fokot a konvergenciarendnek vagy a közelítő megoldás simaságának növelésére használnak).
- Előfordulhat, hogy V_h-ban $u_{|T_k}$ nem az összes legfeljebb ℓ_k-adfokú (ill. ℓ-edfokú) polinomot veheti fel, ennek pontosítására használjuk a

$$P(T_k) := \{u_{|T_k} : \ u \in V_h\} \tag{3.19}$$

  jelölést.
- Az ℓ-edfokú polinomokat T_k-ban kijelölt *csomóponti értékek* határozzák meg. Ezek függvényértékek, és lehetnek még deriváltértékek is. Amikor a csomóponti értékek csak függvényértékek, akkor a *bázis* olyan polinomokból áll, melyek egy adott csomópontban 1-et, a többiben 0-t vesznek fel, azaz ha $x_1, \ldots, x_r$ jelöli a csomópontokat, akkor

$$\varphi_i(x_j) = \delta_{ij}.$$

Ha teljesül az első két tulajdonság, akkor értelmes az alábbi

3.12. Definíció.

(i) Tegyük fel, hogy a T_k halmazok azonos típusúak, és $\ell_k \equiv \ell$. Ekkor *elemnek* nevezzük a T_k halmaztípus és a használt polinomosztály együttesét. Az elem *rendje* p, ha ez a polinomosztály minden p-edfokú polinomot tartalmaz.

(ii) Ha a használt polinomok meghatározására előírt csomóponti értékek csak függvényértékek, akkor Lagrange-elemről, ha pedig deriváltértékek is lehetnek, akkor Hermite-elemről beszélünk. ♢

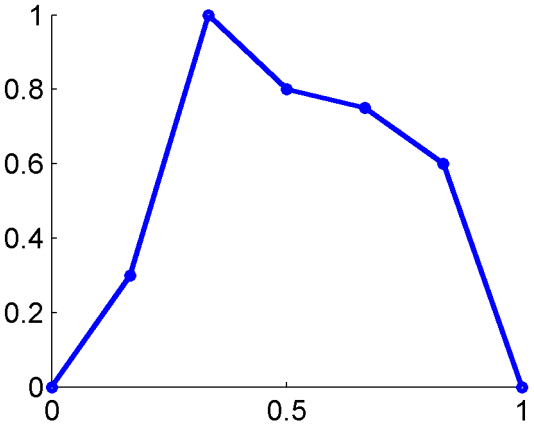
Az alábbiakban példákat mutatunk a leggyakrabban használt elemekre. Az 1-7. elemek Lagrange-típusúak, a 8-10. elemek pedig Hermite-típusúak.

Az elemek elnevezésére gyakran használt jelölés háromszög vagy tetraéder esetén a $\mathbf{T_s}$-elem, téglalap vagy téglatest esetén az $\mathbf{R_s}$-elem, ahol s a szabadsági fokok (Lagrange-elem esetén egyben a csomópontok) száma. (Ez az angol „triangle/tetrahedron”, ill. „rectangle” szavak kezdőbetűjéből jön.)

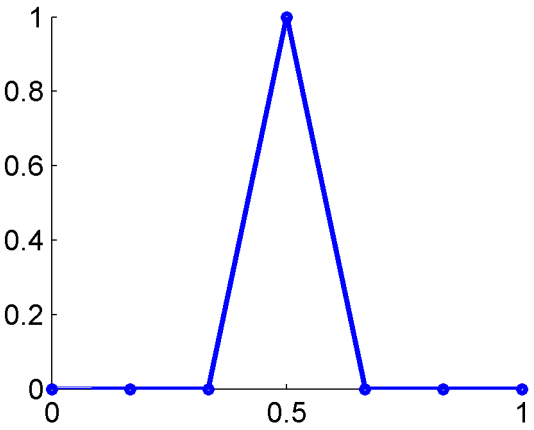
A használt polinomok s szabadsági fokát mindig s adattal fogjuk meghatározni. Az s adat függetlensége nem mindig nyilvánvaló a magasabbfokú esetekben, de ennek bizonyítását itt nem közöljük, pl. a [8] könyvben olvasható.

1. 1D példa.

   Bevezetésnek itt a KDE-k (másodrendű peremérték-feladatok) megoldásánál használt legegyszerűbb alteret írjuk fel. Ekkor a résztartományok is intervallumok, a V_h altér a folytonos, szakaszonként lineáris függvényekből áll (lásd 3.1. ábra), melynek bázisát a $\varphi_i(x_j) = \delta_{ij}$ feltétel alapján meghatározott „kalapfüggvények” alkotják (lásd 3.2. ábra).



3.1. ábra. Egy u_h szakaszonként lineáris függvény a [0,1] intervallumon.



3.2. ábra. Egy ϕ_i szakaszonként lineáris bázisfüggvény a [0,1] intervallumon.

2. 2D eset, $\mathbf{T_3}$-elem vagy Courant-elem.

   Ez a legegyszerűbb, gyakran használatos elem, amelyben

$$T_k \text{ háromszög}, \quad u_{|T_k} \text{ lineáris függvény}$$

   ($\forall k = 1, \ldots, M$, ahol háromszögön csak a nem elfajuló esetet értjük). Itt a hagyományos „lineáris függvény” (ill. az altér elemeire a „szakaszonként lineáris”) szóhasználat inhomogén lineáris függvényt, azaz elsőfokú polinomot jelöl (lásd 3.5. ábra). Emellett $u_{|T_k}$ bármely elsőfokú polinom lehet, azaz

$$P(T_k) = P^1(T_k).$$

   Ekkor a csomópontok a háromszögek csúcsai. Nyilvánvaló, hogy az itt felvett három érték minden háromszögön egyértelműen meghatározza az $u_{|T_k}$ lineáris függvényt, hiszen az

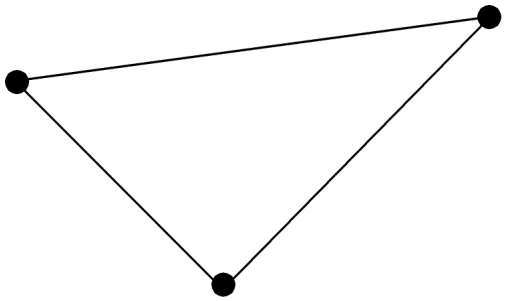
$$x, y \mapsto a_k + b_k x + c_k y$$

   függvényben három szabad paraméter van.

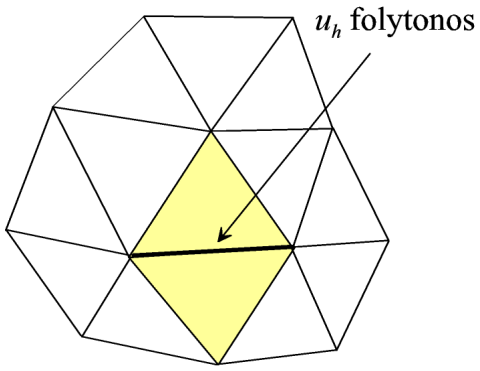
   Emellett az élek mentén, és így az egész $\overline{\Omega}$-on is értelmes folytonos függvényt kapunk, mivel két szomszédos háromszög közös élén a két csúcsbeli függvényérték egyértelműen meghatározza az élen vett egydimenziós lineáris függvényt, így mindkét háromszögről az élre vett leszűkítés megegyezik. Ezt a 3.4. ábrán szemléltetjük.

   Az altér tehát

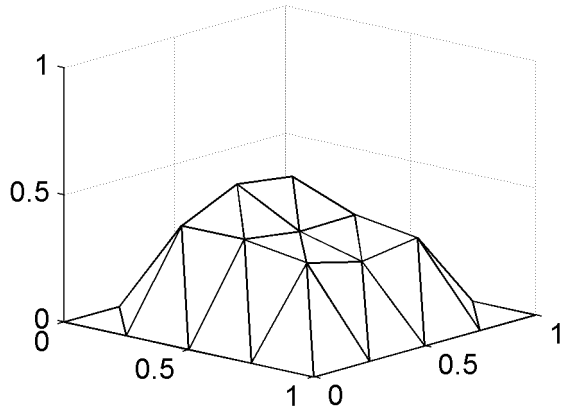
$$V_h = \{u \in C(\overline{\Omega}) : \ u_{|T_k} \in P^1(T_k) \ \forall k = 1, \ldots, M\}.$$



3.3. ábra. T_3-elem (Courant-elem).



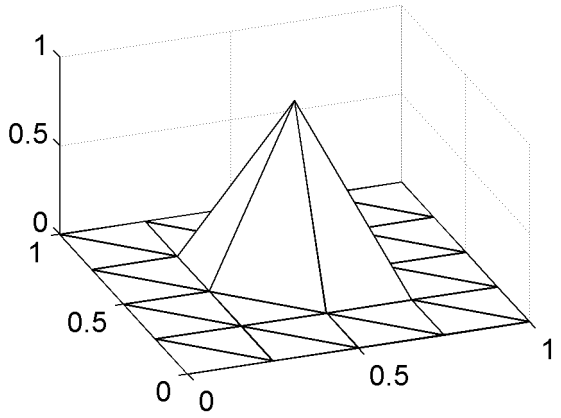
3.4. ábra. Folytonosság az egész tartományon.



3.5. ábra. Egy u_h szakaszonként lineáris függvény az egységnégyzeten.

A V_h altér bázisát (a síkbeli csomópontokat most (x_j, y_j)-vel jelölve) a $\varphi_i(x_j, y_j) = \delta_{ij}$ feltétel alapján meghatározott „sátorfüggvények” alkotják (3.6. ábra).

A Courant-elem a 3.3. ábrán látható, rendje 1.



3.6. ábra. Egy φ_i szakaszonként lineáris bázisfüggvény az egységnégyzeten.

3. 2D eset, $\mathbf{T_6}$-elem.

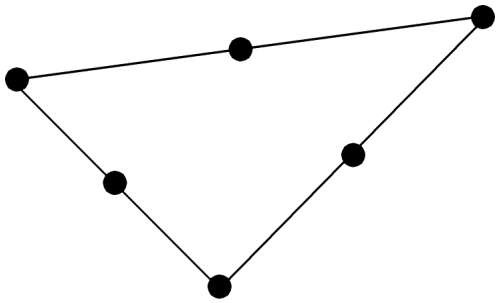
Legyen most

$$T_k \text{ háromszög}, \quad u_{|T_k} \text{ másodfokú polinom}$$

$(\forall k = 1, \ldots, M)$, ahol $u_{|T_k}$ bármely másodfokú polinom lehet, azaz

$$P(T_k) = P^2(T_k).$$

Csomópontoknak a háromszögek csúcsait és az élek felezőpontjait választjuk.



3.7. ábra. T_6-elem.

Az itt felvett hat érték minden háromszögön egyértelműen meghatároz egy másodfokú polinomot, hiszen az

$$x, y \mapsto \; a_k + b_k x + c_k y + d_k x^2 + e_k xy + f_k y^2 \tag{3.20}$$

függvényben hat szabad paraméter van. Emellett az élek mentén, és így az egész $\overline{\Omega}$-on is értelmes folytonos függvényt kapunk, mivel két szomszédos háromszög közös

élén a három adott függvényérték (a két csúcsbeli és a felezőpontbeli) egyértelműen meghatározza az élen vett egydimenziós másodfokú polinomot, így mindkét háromszögről az élre vett leszűkítés megegyezik.

Az altér tehát

$$V_h = \{u \in C(\overline{\Omega}) : \ u_{|T_k} \in P^2(T_k) \ \forall k = 1, \ldots, M\}.$$

A V_h altér bázisát a $\varphi_i(x_j, y_j) = \delta_{ij}$ feltétel alapján meghatározott függvények alkotják. Ezek hat együtthatója egy hatismeretlenes lineáris algebrai egyenletrendszerből kapható meg, amely a (3.20) képletnek a $\varphi_i(x_j, y_j) = \delta_{ij}$ egyenlőségbe való behelyettesítéséből származik.

A $\mathbf{T_6}$-elem a 3.7. ábrán látható, rendje 2.

4. 2D eset, $\mathbf{R_4}$-elem.

Legyen

$$T_k \text{ téglalap}, \quad u_{|T_k} \text{ bilineáris függvény}$$

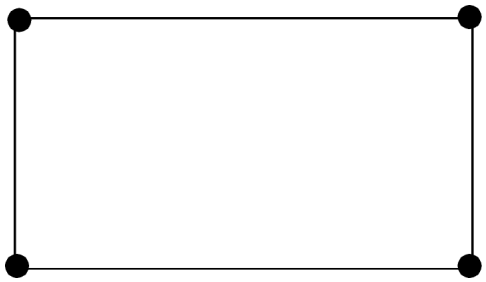
($\forall k = 1, \ldots, M$, ahol téglalapon csak a nem elfajuló esetet értjük)), ahol bilineáris függvénynek egy

$$x, y \mapsto \ a_k + b_k x + c_k y + d_k \, xy$$

alakú függvényt hívunk. E kifejezés oka, hogy ez mindkét változójára nézve különkülön elsőfokú polinom, azaz a fenti szóhasználattal lineáris függvény. A bilineáris függvények nem egyeznek meg egy rögzített fokú polinomosztállyal, hanem

$$P^1(T_k) \subsetneq P(T_k) \subsetneq P^2(T_k). \tag{3.21}$$

Csomópontoknak a téglalapok csúcsait választjuk.



3.8. ábra. R_4-elem.

Az itt felvett négy érték minden téglalapon egyértelműen meghatározza a négy szabad paraméterrel megadható bilineáris függvényt, emellett az élek mentén is értelmes folytonos függvényt kapunk, mivel két szomszédos téglalap közös élén

a két adott függvényérték egyértelműen meghatározza az élen vett egydimenziós elsőfokú polinomot.

A V_h altér bázisát a $\varphi_i(x_j, y_j) = \delta_{ij}$ feltétel alapján meghatározott függvények alkotják, és együtthatóik az előző ponthoz hasonlóan lineáris algebrai egyenletrendszerből határozhatók meg. Ugyanezt a további Lagrange-elemeknél már nem írjuk le.

Az $\mathbf{R_4}$-elem a 3.8. ábrán látható, rendje (3.21) révén 1.

5. 2D eset, $\mathbf{R_8}$-elem.

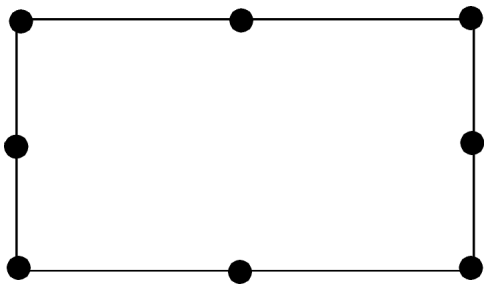
   Legyen T_k téglalap, $u_{|T_k}$ pedig olyan polinom, amely mindkét változójára nézve másodfokú, de nincs benne x^2y^2-es tag: azaz

$$x, y \mapsto \; a_k + b_k x + c_k y + d_k x^2 + e_k xy + f_k y^2 + g_k x^2 y + h_k x y^2.$$

   Ekkor

$$P^2(T_k) \subsetneq P(T_k) \subsetneq P^3(T_k).$$

   Csomópontoknak a téglalapok csúcsait és az élek felezőpontjait választjuk.



3.9. ábra. R_8-elem.

   Az itt felvett nyolc érték ismét minden téglalapon egyértelműen meghatározza a fenti típusú függvényt, az élek mentén való folytonosság pedig a T_6-elemével azonos módon adódik.

   Az $\mathbf{R_8}$-elem a 3.9. ábrán látható, rendje 2.

6. 3D eset, $\mathbf{T_4^3}$-elem.

   Első 3D-s példaként ismét a legegyszerűbb esetet nézzük, amely a 2D-beli $\mathbf{T_3}$-elem közvetlen megfelelője:

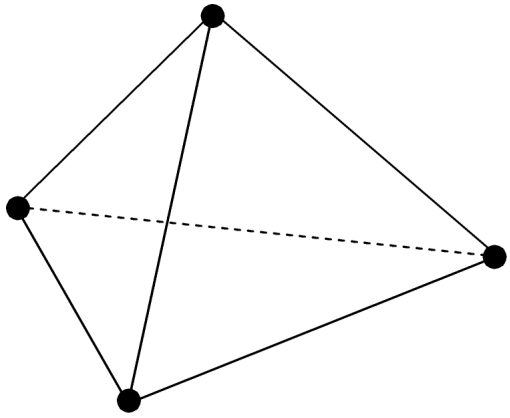
$$T_k \text{ tetraéder}, \quad u_{|T_k} \text{ lineáris függvény}$$

$(\forall k = 1, \ldots, M)$. A tetraéder nem elfajuló volta és a „lineáris függvény” szóhasználat azonos a $\mathbf{T_3}$ esetével.

A csomópontok a tetraéderek csúcsai, az itt felvett négy érték minden tetraéderen egyértelműen meghatározza az

$$x, y, z \mapsto a_k + b_k x + c_k y + d_k z$$

lineáris függvényt.



3.10. ábra. T_4^3-elem.

Emellett a lapok mentén is értelmes folytonos függvényt kapunk, mivel két szomszédos tetraéder közös lapját alkotó háromszögön a három csúcsbeli függvényérték egyértelműen meghatározza a lapon vett kétdimenziós lineáris függvényt.

Az altér tehát

$$V_h = \{u \in C(\overline{\Omega}) : \ u_{|T_k} \in P^1(T_k) \ \forall k = 1, \ldots, M\}.$$

A $\mathbf{T_4^3}$-elem a 3.10. ábrán látható, rendje 1.

7. 3D eset, $\mathbf{T_{10}^3}$-elem.

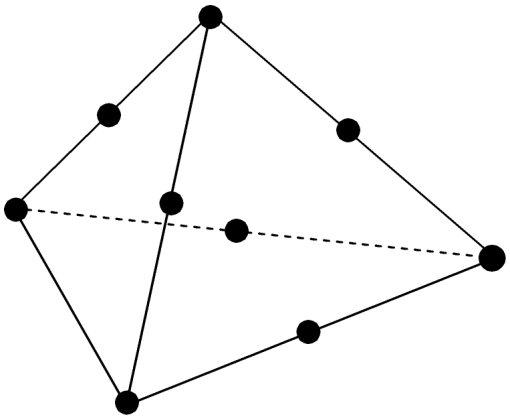
   Ez a 2D-beli $\mathbf{T_6}$-elem megfelelője:

$$T_k \text{ tetraéder}, \quad u_{|T_k} \text{ másodfokú polinom}$$

   $(\forall k = 1, \ldots, M)$. A csomópontok a tetraéderek csúcsai és az élek felezőpontjai, az itt felvett tíz érték minden tetraéderen egyértelműen meghatározza az

$$x, y, z \mapsto a_k + b_k x + c_k y + d_k z + e_k x^2 + f_k y^2 + g_k z^2 + h_k xy + i_k xz + j_k yz$$

   tíz együtthatós másodfokú polinomot.



3.11. ábra. T^3_{10}-elem.

Két szomszédos tetraéder közös lapját alkotó háromszögön a három csúcsban és három élfelezőben vett hat függvényérték egyértelműen meghatározza a lapon vett kétdimenziós másodfokú polinomot.

Az altér tehát

$$V_h = \{u \in C(\overline{\Omega}) : \ u_{|T_k} \in P^2(T_k) \ \forall k = 1, \ldots, M\}.$$

A $\mathbf{T^3_{10}}$-elem a 3.11. ábrán látható, rendje 2.

8. 3D eset, $\mathbf{R^3_8}$-elem.

Ez a 2D-beli $\mathbf{R_4}$-elem megfelelője:

$$T_k \text{ téglatest}, \quad u_{|T_k} \text{ trilineáris függvény}$$

$(\forall k = 1, \ldots, M)$, ahol trilineáris függvénynek egy

$$x, y, z \mapsto \ a_k + b_k x + c_k y + d_k z + e_k xy + f_k xz + g_k yz + h_k xyz$$

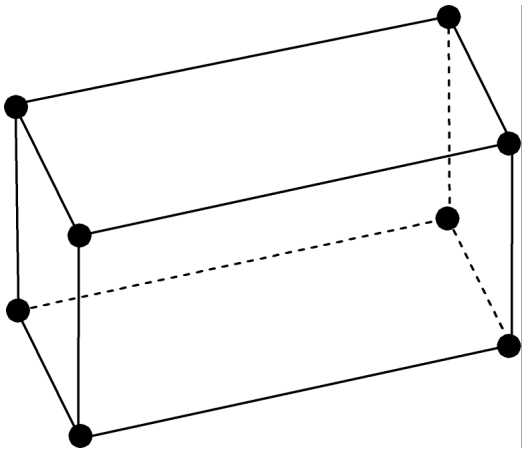
alakú függvényt hívunk, amely tehát mindhárom változójára nézve külön-külön elsőfokú polinom. Ekkor ismét

$$P^1(T_k) \subsetneq P(T_k) \subsetneq P^2(T_k).$$

Csomópontoknak a téglalapok csúcsait választjuk.

Ez a nyolc érték minden téglatesten egyértelműen meghatározza a trilineáris függvényt, két szomszédos téglatest közös lapját alkotó téglalapon pedig a négy adott függvényérték egyértelműen meghatározza a lapon vett kétdimenziós bilineáris függvényt.

Az $\mathbf{R^3_8}$-elem a 3.12. ábrán látható, rendje 1.



3.12. ábra. R_8^3-elem.

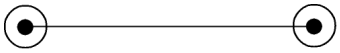
9. 1D köbös Hermite-elem.

   Első példaként Hermite-elemre ismét az 1D esetet (KDE másodrendű peremérték-feladata) idézzük fel, ahol látható, hogy ezzel az Hermite-elemmel az egész szakaszon nemcsak a folytonosság, hanem a folytonos deriválhatóság is elérhető. Ezért szokták C^1-elemnek is hívni.

   A résztartományok intervallumok, a V_h altér a folytonos, szakaszonként harmadfokú polinomokból áll. A csomópontok a végpontok, ahol nemcsak a függvényértékeket, hanem az első deriváltakat is felhasználjuk a polinomok meghatározásához.

   Ekkor egy részintervallumon négy szabadsági fokunk van, melyek egyértelműen meghatározzák a harmadfokú polinomot. Mivel a végpontokban a deriváltak is megegyeznek, a V_h-beli függvények az egész szakaszon folytonosan deriválhatóak.

   Az Hermite-elemek szabadsági fokainak szemléltetéséhez a szokásos jelölés, hogy a csomóponti függvényértékeket az előzőekhez hasonlóan vastag ponttal, az első deriváltakat karikával, a második deriváltakat kettős karikával jelöljük.



3.13. ábra. Egydimenziós köbös Hermite-elem.

   Az egydimenziós köbös Hermite-elemet a 3.13. ábra szemlélteti, rendje 3.

10. 2D eset, $\mathbf{H_{10}}$-elem (köbös Hermite-elem).

    Legyen most

$$T_k \text{ háromszög}, \quad u_{|T_k} \text{ harmadfokú polinom}$$

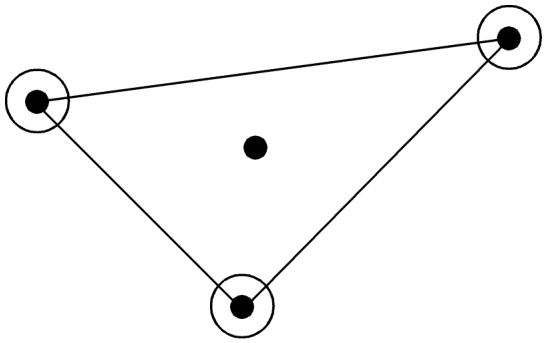
$(\forall k = 1, \ldots, M)$, azaz

$$x, y \mapsto \; a_k + b_k x + c_k y + d_k x^2 + e_k xy + f_k y^2 + g_k x^3 + h_k x^2 y + i_k xy^2 + j_k y^3.$$

Ekkor

$$P(T_k) = P^3(T_k).$$

Csomópontoknak a háromszög csúcsait és súlypontját választjuk, a csúcsokban a polinomok meghatározásához a függvényértékeket és a két első parciális deriváltat (tehát 3-3 adatot), a súlypontban csak a függvényértéket használjuk. Ez a tíz érték egyértelműen meghatározza a harmadfokú polinomot.



3.14. ábra. Kétdimenziós köbös Hermite-elem.

Két szomszédos háromszög közös élén a csúcsokban adott két függvényérték és a (parciális deriváltak által meghatározott) két élirányú derivált az 1D elemnél látottak alapján egyértelműen meghatározza az élen vett egydimenziós harmadfokú polinomot. Ezért az élek mentén, és így az egész $\overline{\Omega}$-on is értelmes folytonos függvényt kapunk.

Igazolható azonban, hogy az élek mentén a normális irányú (tehát az élekre merőleges) deriváltak nem feltétlenül egyeznek meg a két szomszédos háromszögre nézve, így az 1D esettel ellentétben itt nem következik a folytonosan deriválhatóság. Ehhez a következő pontbeli magasabb fokra van szükség, lásd [8].

Az altér tehát

$$V_h = \{u \in C(\overline{\Omega}) : \; u_{|T_k} \in P^3(T_k) \; \forall k = 1, \ldots, M\}.$$

A V_h altér bázisát olyan harmadfokú polinomok alkotják, melyeknél a fenti tíz szabadsági fok egyike 1, a többi mind 0. Ezek a korábbiakhoz hasonlóan egy (tízismeretlenes) lineáris algebrai egyenletrendszerből kaphatók meg.

A kétdimenziós köbös Hermite-elemet a 3.14. ábra szemlélteti, rendje 3.

11. 2D eset, $\mathbf{H_{21}}$-elem (ún. Argyris-elem).

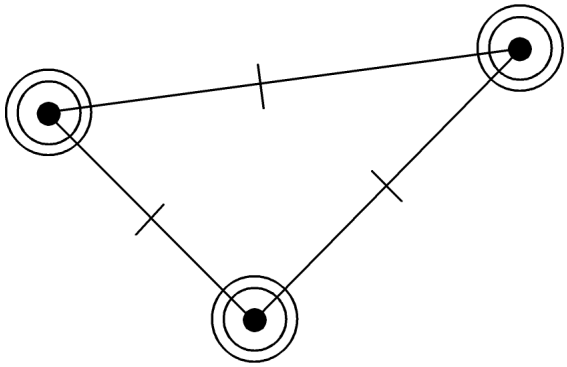
Legyen most

$$T_k \text{ háromszög}, \quad u_{|T_k} \text{ ötödfokú polinom},$$

$(\forall k = 1, \ldots, M)$, azaz

$$P(T_k) = P^5(T_k).$$

Ennek 21 együtthatója van. Csomópontoknak a háromszög csúcsait és élfelezőit választjuk, a csúcsokban a polinomok meghatározásához a függvényértékeket, a két első és a három második parciális deriváltat (tehát 6-6 adatot), az élfelezőkben a normális irányú deriváltat használjuk. Ez a 21 érték egyértelműen meghatározza az ötödfokú polinomot.



3.15. ábra. Argyris-elem.

A korábbiakhoz hasonlóan látható, hogy az élek mentén, és így az egész $\overline{\Omega}$-on is értelmes folytonos függvényt kapunk: két szomszédos háromszög közös élén a csúcsokban adott két függvényértékből és a (parciális deriváltak által meghatározott) két-két élirányú első és második deriváltból kapott hat adat egyértelműen meghatározza az élen vett egydimenziós ötödfokú polinomot. Most azonban az is igazolható, hogy a teljes élek mentén a normális irányú deriváltak is megegyeznek a két szomszédos háromszögre nézve, így $C^1(\overline{\Omega})$-beli függvényt kapunk. (A definíció ezt csak az élfelezőkben garantálja. Az állítást itt nem bizonyítjuk, a számolást lásd [8]-ban.) Az Argyris-elem tehát 2-dimenziós C^1-elem, az altér pedig

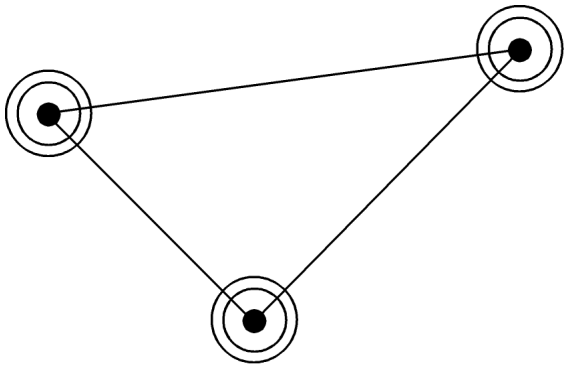
$$V_h = \{u \in C^1(\overline{\Omega}) : \ u_{|T_k} \in P^5(T_k) \ \forall k = 1, \ldots, M\}.$$

Bázisát az előzővel analóg módon hozhatjuk létre.

Az Argyris-elemet a 3.15. ábra szemlélteti, rendje 5.

12. 2D eset, $\mathbf{H_{18}}$-elem (ún. Bell-elem).

A Bell-elem az Argyris-elem módosítása úgy, hogy $u_{|T_k}$ nem tetszőleges ötödfokú polinom lehet, csak olyan, melynek az élekre való leszűkítése harmadfokú polinom. Igazolható, hogy ez hárommal csökkenti a szabadsági fokok számát, így csomópontoknak elég a háromszög csúcsait választani az Argyris-elemnél vett 18 adattal, valamint az is, ezzel nem vész el a $C^1(\overline{\Omega})$-beliség, vagyis a Bell-elem is 2-dimenziós C^1-elem (lásd szintén [8]-ban). A Bell-elemet a 3.16. ábra szemlélteti, rendje 4.



3.16. ábra. Bell-elem.

Téglalapon, ill. 3D-ben is értelmezhetők alkalmas Hermite-elemek, ezekről szintén [8]-ban olvashatunk.

3.13. Megjegyzés.

(i) A felsorolt példákban a T_k résztartományok azonos típusúak, vagyis egy adott triangulációcsupa háromszögből/tetraéderből vagy csupa négyszögből/téglatestből áll (rendre a 2D/3D esetben). Ilyenkor célszerű a résztartományokat és a rajtuk értelmezett megfelelő polinomokat egyetlen ún. referenciaelem és az azon értelmezett megfelelő polinomok transzformáltjaként reprezentálni. Ez affin (tehát inhomogén lineáris) transzformáció, így kezelése egyszerű. Erre a 3.4 (b) szakaszban mutatunk konkrét példát.

(ii) A végeselem-módszer a fentieknél általánosabb, görbe határú elemekkel is értelmezhető. Ennek legegyszerűbb módja, ha az előbb említett módon referenciaelemet használunk, és a transzformációt vesszük affin helyett általánosabban. Például, referenciaháromszög másodfokú transzformációival parabolaívekkel határolt elemek nyerhetők, melyekkel a tartomány határa szükség esetén jobban közelíthető, mint töröttvonallal (lásd pl. [27, 15.7.9. szakasz]).

(iii) A végeselem-módszer másik általánosítása az ún. nem konform vagy DG („discontinuous Galerkin") típusú módszer, amikor nem követeljük meg a folytonosságot az egész tartományon. Ennek elméletéhez a bevezetőbeli Galjorkin-módszert is általánosítani kell (lásd pl. [5]). ♢

3.14. Megjegyzés. Konkrét példaként érdemes megemlíteni az egyenletes (azaz négyzetrács egyirányú átlós felezéseiből kapott) háromszögrácshoz tartozó Courant-elemek merevségi mátrixát a Poisson-egyenlet esetén Dirichlet-peremfeltétellel. Könnyen látható (lásd 9.18. feladat), hogy $n \times m$ belső csomópont esetén

$$A_h = \begin{pmatrix} B & -I & & & & \\ -I & B & -I & & & \\ & -I & B & -I & & \\ & & \ddots & & \ddots & \\ & & & -I & B & -I \\ & & & & -I & B \end{pmatrix} \in \mathbb{R}^{nm \times nm}$$

blokk-tridiagonális mátrix, ahol I az $n \times n$-es identitásmátrix és

$$B = \begin{pmatrix} 4 & -1 & & & & \\ -1 & 4 & -1 & & & \\ & -1 & 4 & & & \\ & & \ddots & & \ddots & \\ & & & -1 & 4 & -1 \\ & & & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

tridiagonális mátrix, melyből m db szerepel. Ez a 2.2. szakasszal összevetve azt jelenti, hogy a fenti végeselemes $A_h = A_h^{FEM}$ és az ugyanezen csomópontokhoz tartozó FDM-es A_h^{FDM} mátrixokra

$$A_h^{FEM} = h^2 A_h^{FDM}.$$ ♢

3.3. Az FEM stabilitása

A végeselem-módszerrel konstruált u_h közelítő megoldásra a V_h altértől független (ún. rácsfüggetlen) stabilitási becslés igazolható. A rácsfüggetlenség abból adódik, hogy a pontos gyenge megoldásra felírt stabilitási becslés egy az egyben átvihető a közelítő megoldásra. Itt rögtön csak az utóbbival foglalkozunk, először az absztrakt Galjorkin-módszer esetén.

3.15. Tétel. *Legyen H valós Hilbert-tér, $a : H \times H \to \mathbb{R}$ korlátos, koercív bilineáris forma m alsó határral (azaz (3.2) teljesül), valamint legyen $\ell : H \to \mathbb{R}$ korlátos lineáris funkcionál. Legyen $V_h \subset H$ adott véges dimenziós altér, és $u_h \in V_h$ a (3.4) feladat megoldása:*

$$a(u_h, v_h) = \ell v_h \qquad (\forall v_h \in V_h). \tag{3.22}$$

Ekkor

$$\|u_h\| \leq \frac{1}{m} \|\ell\|. \tag{3.23}$$

Bizonyítás. Helyettesítsük a $v_h := u_h$ elemet (3.22)-be:

$$a(u_h, u_h) = \ell u_h.$$

A koercivitás miatt

$$m\|u_h\|^2 \le \|\ell\|\,\|u_h\|.$$

Ebből rögtön következik a kívánt becslés. □

A stabilitás szokásos következménye (a 2.19. állításhoz hasonlóan), hogy a jobboldal hibája korlátos mértékben öröklődik a megoldás hibájára: ha a (3.22) feladat ℓ_1 és ℓ_2 jobboldalakhoz tartozó megoldása rendre u_1^h és u_2^h, akkor

$$\|u_1^h - u_2^h\| \le \frac{1}{m}\|\ell_1 - \ell_2\|.$$

Alkalmazzuk a kapott becslést először a (3.11) Dirichlet-feladatra! Legyen tehát

$$\begin{cases} -\operatorname{div}(p\,\nabla u) = f, \\ u_{|\partial\Omega} = 0, \end{cases}$$

ahol $\Omega \subset \mathbb{R}^N$ korlátos tartomány, $p \in L^\infty(\Omega)$ és alkalmas $m > 0$ esetén $p(x) \ge m > 0$ (m-m. $\forall x \in \Omega$), valamint $f \in L^2(\Omega)$. Itt $a(u,v)$ és ℓv az (1.14) gyenge alak bal és jobb oldala, azaz

$$a(u,v) := \int_\Omega p\,\nabla u \cdot \nabla v, \qquad \ell v := \int_\Omega f v \qquad (u, v \in H_0^1(\Omega)).$$

Ekkor a bilineáris forma $H_0^1(\Omega)$-on vett alsó határának becslésére a p alsó határára megadott m vehető (lásd (1.20)), így csak $\|\ell\|$-et kell becsülnünk. Emlékeztetünk az 1.2. szakaszban bevezetett függvénynorma-jelölésekre. Itt a Cauchy–Schwarz-egyenlőtlenség és az (1.8) Poincaré–Friedrichs-egyenlőtlenség révén

$$|\ell v| \le \|f\|_0\|v\|_0 \le C_\Omega\|f\|_0|v|_1 \qquad (\forall v \in H_0^1(\Omega)), \tag{3.24}$$

így

$$\|\ell\| \le C_\Omega\|f\|_0 \quad \text{és} \quad |u_h|_1 \le \frac{C_\Omega}{m}\|f\|_0. \tag{3.25}$$

A C_Ω konstans éles értéke $\frac{1}{\sqrt{\lambda_1}}$, ahol $\lambda_1 > 0$ a Laplace-operátor legkisebb sajátértéke az Ω tartományon homogén Dirichlet-peremfeltétellel (lásd 9.15. feladat), λ_1-re pedig érvényes a

$$\lambda_1 \ge \frac{n\pi^2}{diam(\Omega)^2}$$

becslés, ahol n a tér dimenziója és $diam(\Omega)$ a tartomány átmérője (lásd 9.16. feladat). Ezekből

$$|u_h|_1 \leq \frac{diam(\Omega)}{m\pi\sqrt{n}}\|f\|_0,$$

amely könnyen kiszámítható és rácsfüggetlen stabilitási becslés.

Vegyes peremfeltételű feladat esetén a (3.24) becslés értelemszerűen módosul. Ekkor az ℓ funkcionál az (1.22) gyenge alak jobboldala:

$$\ell v := \int_\Omega fv + \int_{\Gamma_N} \gamma v \qquad (\forall v \in H^1_D(\Omega)),$$

melyre a (3.24) becslést az (1.12) beágyazás révén egészítjük ki (az L^2-normákban az alaphalmaz feltüntetésével):

$$|\ell v| \leq \|f\|_{0,\Omega}\|v\|_{0,\Omega} + \|\gamma\|_{0,\Gamma_N}\|v\|_{0,\Gamma_N} \leq \left(C_\Omega\|f\|_{0,\Omega} + C_{\Gamma_N}\|\gamma\|_{0,\Gamma_N}\right)|v|_1 \qquad (\forall v \in H^1_D(\Omega)),$$

így

$$\|\ell\| \leq \left(C_\Omega\|f\|_{0,\Omega} + C_{\Gamma_N}\|\gamma\|_{0,\Gamma_N}\right)\|f\|_0.$$

(A C_{Γ_N} konstansra nincs a C_Ω-hoz hasonlóan egyszerű fenti általános becslés, egyes sokszögű tartományokra a [24] könyvben találhatók ilyen becslések.)

3.4. Az FEM konvergenciája

3.4.1. Bevezetés: konvergencia és interpoláció

Mint láttuk, a Galjorkin-módszer konvergenciavizsgálatának alapja a 3.7. Céa-lemma. Tekintsük először a (3.11) Dirichlet-feladatot, és a 3.2. szakaszban ismertetett véges-elemes megoldást. Itt a gyenge megoldást a $H^1_0(\Omega)$ térben kaptuk, ezért a Céa-lemmát az $|u|_1 := \left(\int_\Omega |\nabla u|^2\right)^{1/2}$ H^1_0-normában használjuk. A 3.8. megjegyzés alapján infimum helyett rögtön minimumot írunk, így

$$|u^* - u_h|_1 \leq \frac{M}{m}\min_{v_h \in V_h}|u^* - v_h|_1,$$

ahol $m = \operatorname{ess\,inf} p$ és $M = \operatorname{ess\,sup} p = \|p\|_{L^\infty}$. A konvergenciához tehát az alterekre vonatkozóan H^1_0-normabeli approximációs tulajdonság kell:

$$\forall u \in H \quad dist(u, V_h) := \min_{v_h \in V_h}|u - v_h|_1 \to 0 \qquad (\text{ha} \quad h \to 0).$$

A gyakorlatban általában ennél több, éspedig a $dist(u, V_h) \to 0$ konvergencia rendje lesz a kérdés. Vegyes peremfeltétel esetén a Céa-lemmát a fentivel azonos formában kapjuk, mivel $|u|_1 := \left(\int_\Omega |\nabla u|^2\right)^{1/2}$ a $H^1_D(\Omega)$ téren is norma. Mivel ez az általánosabb feladatosztály, így a továbbiakban erre hivatkozunk:

3.16. Következmény. *A (3.11) feladat $u_h \in V_h$ végeselemes megoldására*

$$|u^* - u_h|_1 \leq \frac{M}{m} \min_{v_h \in V_h} |u^* - v_h|_1, \tag{3.26}$$

ahol M és m a (3.18) bilineáris forma határai.

(A határok értékére nézve lásd a 9.14. feladatot).

Hogyan becsülhető felülről a $\min_{v_h \in V_h} |u^* - v_h|_1$ érték, ha sem u^*-t, sem az optimális v_h függvényt nem ismerjük?

A becslés alapgondolata az, hogy becsüljük felülről ezt a minimumot az optimális v_h helyett vett másik alkalmas függvénnyel. Erre célszerű választás u^* megfelelő interpolációja a V_h altérben. Jelölje

$$\Pi_h u^* \in V_h$$

ezt az interpolációt; ekkor

$$|u^* - u_h|_1 \leq \frac{M}{m} |u^* - \Pi_h u^*|_1 . \tag{3.27}$$

Így tehát a jobb oldalon álló interpolációs hibát kell megbecsülnünk.

A becsléseket tetszőleges $u \in H_D^1(\Omega)$ függvényre végezzük el. Itt a peremfeltételt sem használjuk, azaz u tetszőleges $H^1(\Omega)$-beli is lehet; a peremfeltétel abban számít, hogy a becsléseknél megmaradunk $|u|_1$-nél, és ez a végeredményben vizsgált $u \in H_D^1(\Omega)$ esetben normát ad. Azaz, a továbbiakban a feladat:

$$|u - \Pi_h u|_1 \leq ? \qquad (u \in H_D^1(\Omega)). \tag{3.28}$$

Ebben a szakaszban a lineáris (azaz elsőrendű) interpolációt vizsgáljuk meg, a magasabbrendű esettel utána foglalkozunk. A számolásokat kétdimenziós esetre végezzük el.

3.4.2. Az FEM elsőrendű konvergenciabecslése Courant-elemekre

Részletesen a Courant-elemek, azaz háromszögeken lineáris polinomokkal értelmezett végeselemek esetén vizsgáljuk a konvergenciát, amelyre ekkor elsőrendű becslést fogunk kapni.

Elsőrendű interpolációs becslés. Az $|u - \Pi_h u|_1$ normát elemenként vizsgáljuk: mivel

$$\begin{aligned} |u - \Pi_h u|_1^2 :=& |u - \Pi_h u|_{H^1(\Omega)}^2 := \int_\Omega |\nabla(u - \Pi_h u)|^2 \\ =& \sum_{k=1}^{M} \int_{T_k} |\nabla(u - \Pi_h u)|^2 =: \sum_{k=1}^{M} |u - \Pi_h u|_{H^1(T_k)}^2, \end{aligned} \tag{3.29}$$

így a

$$|u - \Pi_h u|_{H^1(T_k)}$$

normákra kapott becslésekből összerakható a teljes tartományra való becslés.

Előzetes: az 1D eset. A becslések áttekintését segíti az 1D analógia vizsgálata. Tekintsük az $I = [0,1]$ intervallumot, és ennek ekvidisztáns felosztását. A részintervallumok vizsgálatát elég az elsőn, az $I_h := [0,h]$ intervallumon elvégezni, ahol $h > 0$ állandó. Ezen egy $u \in H^1(I_h)$ függvény lineáris interpoláltja:

$$(\Pi_h u)(x) = u(0) + \frac{u(h) - u(0)}{h}\, x.$$

3.17. Állítás. *Legyen $u \in H^2(I_h)$. Ekkor*

$$|u - \Pi_h u|_{H^1(I_h)} \le h\, \|u''\|_{L^2(I_h)}.$$

Bizonyítás. Itt

$$(\Pi_h u)'(x) = \frac{u(h) - u(0)}{h} = \frac{1}{h}\int_0^h u'(t)dt,$$

így

$$u'(x) - (\Pi_h u)'(x) = \frac{1}{h}\int_0^h (u'(x) - u'(t))\, dt = \frac{1}{h}\int_0^h \int_t^x u''(s)\, ds\, dt.$$

Ebből a Cauchy–Schwarz-egyenlőtlenség és konstansintegrálás többszöri alkalmazásával

$$\begin{aligned}\big|u'(x) - (\Pi_h u)'(x)\big|^2 &\le \frac{1}{h^2}\int_0^h \Big|\int_t^x u''(s)\, ds\Big|^2 dt \cdot \int_0^h 1^2\, dt = \frac{1}{h}\int_0^h \Big|\int_t^x u''(s)\, ds\Big|^2 dt \\ &\le \frac{1}{h}\int_0^h \Big(\int_t^x |u''(s)|^2 ds \int_t^x 1^2\, ds\Big)\, dt \le \frac{1}{h}\int_0^h \Big(\int_0^h |u''(s)|^2 ds \cdot \int_0^h 1^2\, ds\Big)\, dt \\ &= \int_0^h \int_0^h |u''(s)|^2 ds\, dt = \int_0^h \|u''\|^2_{L^2(I_h)}\, dt = h \cdot \|u''\|^2_{L^2(I_h)}\end{aligned}$$

és

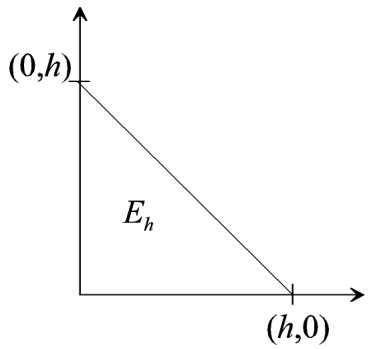
$$|u - \Pi_h u|^2_{H^1(I_h)} = \int_0^h \big|u'(x) - (\Pi_h u)'(x)\big|^2 dx \le h^2 \cdot \|u''\|^2_{L^2(I_h)},$$

ami a kívánt becslés négyzete. □

3.18. Megjegyzés. A kapott eredmény analóg a Taylor-sor hibájának nagyságrendjével, ami nem meglepő, hiszen u és a lineáris közelítés különbsége a 2. deriválttal állt elő. A fenti számolás ugyanígy folytatható magasabbrendű esetre: ha $u \in H^{k+1}(I_h)$, azaz m-m. létezik $u^{(k+1)}$ és $L^2(I_h)$-beli, valamint $\Pi_h u$ k-adfokú interpoláns, akkor a fenti integrálás k-szor folytatható és becslése mindig egy újabb h-s nagyságrendet produkál, amiből alkalmas $c > 0$ konstans mellett

$$|u - \Pi_h u|_{H^1(I_h)} \le c\, h^k\, \|u^{(k+1)}\|_{L^2(I_h)} = c\, h^k |u|_{H^{k+1}(I_h)}.$$ ♢

A 2-dimenziós eset háromszög-elemekkel. Tekintsük először az E_h speciális elemet (h-egységháromszög-elemet), lásd 3.17. ábra.



3.17. ábra. Az E_h speciális háromszög-elem.

Ekkor a 3.17. állításhoz hasonlóan igaz, hogy ha $u \in H^2(E_h)$, akkor van olyan $c > 0$, hogy

$$|u - \Pi_h u|_{H^1(E_h)} \leq ch \, \|D^2 u\|_{L^2(E_h)}, \tag{3.30}$$

ahol $D^2 u$ az u Hesse-mátrixa, és

$$\|D^2 u\|^2_{L^2(E_h)} := \int_{E_h} (|\partial_{11} u|^2 + |\partial_{12} u|^2 + |\partial_{22} u|^2).$$

(A hosszabb számolás a [7] könyvben található.)

A továbbiakban áttérünk az $E := E_1$ háromszögre. Ez a *referenciaháromszög*, melynek affin transzformációjával bármely háromszög előállítható (ezt a merevségi mátrix összeállításánál is ki szokás használni). A referenciaháromszögön használjuk fel a fenti becslést a $h = 1$ esetre:

3.19. Következmény. *Ha $u \in H^2(E)$, akkor van olyan $c > 0$, hogy*

$$|u - \Pi_h u|_{H^1(E)} \leq c \, \|D^2 u\|_{L^2(E)}.$$

Erről fogjuk átvinni a becslést általános háromszögre.

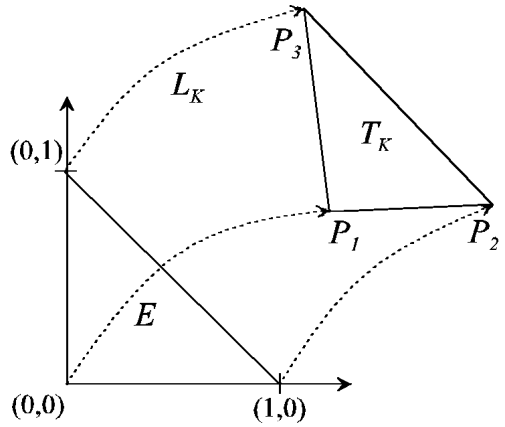
Legyen T_k tetszőleges háromszög egy adott triangulációban, és csúcsait jelöljük P_1, P_2, P_3-mal. Jelölje $L_k : E \to T_k$ az affin transzformációt, lásd 3.18. ábra.

Legyen

$$P_1 = (x_1, y_1), \quad P_2 = (x_2, y_2), \quad P_3 = (x_3, y_3).$$

Ekkor

$$L_k \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} x_1 + (x_2 - x_1)\xi + (x_3 - x_1)\eta \\ y_1 + (y_2 - y_1)\xi + (y_3 - y_1)\eta \end{pmatrix},$$



3.18. ábra. A referenciaelem leképezése.

$$J_k := L_k' \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} =: \begin{pmatrix} u_2 & u_3 \\ v_2 & v_3 \end{pmatrix}.$$

A további számolásokhoz felhasználjuk az alábbi elemi tulajdonságokat:

(i) Az $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$ vektorok által kifeszített háromszög területe: $\frac{1}{2}\, det \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}$. Ebből

$$det(J_k) = det(J_k^T) = det \begin{pmatrix} \overrightarrow{P_1P_2} \\ \overrightarrow{P_1P_3} \end{pmatrix} = 2|T_k|,$$

ahol $|T_k|$ jelöli T_k területét.

(ii) Az inverz mátrix képletéből

$$(L_k^{-1})' = (L_k')^{-1} = \frac{1}{det(J_k)} \begin{pmatrix} v_3 & -u_3 \\ -v_2 & u_2 \end{pmatrix} = \frac{1}{det(J_k)}\, \widetilde{J}_k,$$

ahol $\widetilde{J}_k$ jelöli a fenti számlálóbeli mátrixot.

(iii) Bármely $n \times n$-es A mátrixnak az euklideszi vektorhossz által indukált operátor-normájára

$$\|A\| \le \|A\|_F, \qquad \text{ahol} \qquad \|A\|_F^2 = \sum_{i,j=1}^{n} a_{ij}^2$$

az A mátrix Frobenius-normája.

3.20. Állítás. *Legyen $v \in H^1(T_k)$ és $\widetilde{v} := v \circ L_k \in H^1(E)$. Ekkor*

$$\int_{T_k} |\nabla v|^2 \le \frac{2h_k^2}{|T_k|} \int_E |\nabla \widetilde{v}|^2,$$

ahol $h_k := diam(T_k)$ a T_k háromszög átmérője és $|T_k|$ a T_k területe.

Bizonyítás. Deriválva a $v := \widetilde{v} \circ L_k^{-1}$ egyenlőséget

$$\nabla v = (\nabla \widetilde{v} \circ L_k^{-1}) \cdot (L_k^{-1})' = \frac{1}{det(J_k)} (\nabla \widetilde{v} \circ L_k^{-1}) \cdot \widetilde{J}_k,$$

így

$$\int_{T_k} |\nabla v|^2 = \frac{1}{det(J_k)^2} \int_{T_k} \left|(\nabla \widetilde{v} \circ L_k^{-1}) \cdot \widetilde{J}_k\right|^2.$$

Az integráltranszformáció képlete szerint

$$\int_{T_k} z = \int_E (z \circ L_k)\, det(L_k') = det(J_k) \int_E (z \circ L_k) \qquad (\forall z \in L^1(T_k).$$

Ezt alkalmazva $z := \left|(\nabla \widetilde{v} \circ L_k^{-1}) \cdot \widetilde{J}_k\right|^2$ esetén, és mivel $\widetilde{J}_k \circ L_k = \widetilde{J}_k$ (hiszen ez konstans mátrix),

$$\int_{T_k} |\nabla v|^2 = \frac{1}{det(J_k)} \int_E \left|\nabla \widetilde{v} \cdot \widetilde{J}_k\right|^2.$$

Itt

$$\left|\nabla \widetilde{v} \cdot \widetilde{J}_k\right|^2 \le |\nabla \widetilde{v}|^2 \, \|\widetilde{J}_k\|^2 \le |\nabla \widetilde{v}|^2 \, \|\widetilde{J}_k\|_F^2 \le 4 |\nabla \widetilde{v}|^2 \, h_k^2,$$

mivel $\widetilde{J}_k$ mind a 4 eleme legfeljebb h_k. Másrészt $det(J_k) = 2|T_k|$. Ezekből

$$\int_{T_k} |\nabla v|^2 \le \frac{4h_k^2}{2|T_k|} \int_E |\nabla \widetilde{v}|^2,$$

vagyis az állítást beláttuk. □

3.21. Megjegyzés. Hasonló számolással igazolhatók az alábbi becslések:

(i) az állításbeli becslés ugyanúgy érvényes visszafelé is, azaz

$$\int_E |\nabla \widetilde{v}|^2 \le \frac{2h_k^2}{|T_k|} \int_{T_k} |\nabla v|^2.$$

(A két becslés szimmetrikus viszonyáról lásd a 3.27. megjegyzést.)

(ii) A második deriváltmátrix négyzetintegráljának becslésében h_k kettővel nagyobb rendben szerepel:

$$\int_E |D^2 \widetilde{v}|^2 \le c \frac{h_k^4}{|T_k|} \int_{T_k} |D^2 v|^2, \tag{3.31}$$

♢

ahol $c > 0$ független T_k-tól. (A megfelelő normákra ez a gyökvonás nyomán h_k eggyel nagyobb rendjét jelenti.)

Ezek alapján már közel járunk (3.30) megfelelőjéhez T_k-ra. Míg az eddigi becsléseknek rögzített rács (triangulació) esetén is volt tartalma, a továbbiakban triangulációk családjai esetén vizsgáljuk az interpoláció rendjét.

3.22. Definíció. Triangulációk *családjának* nevezzük triangulációk olyan $\mathcal{F}$ halmazát, melyre minden $h_0 > 0$ esetén létezik $\mathcal{T}_h \in \mathcal{F}$, hogy $\mathcal{T}_h$ finomsága kisebb h_0-nál. ♢

3.23. Állítás. *Legyen $\mathcal{F}$ háromszögű triangulációk egy családja, melyre*

$$\frac{h_k^2}{|T_k|} \quad (T_k \in \mathcal{T}_h,\ \mathcal{T}_h \in \mathcal{F}) \quad \textit{korlátos.}$$

Ha T_k adott háromszög és $u \in H^2(T_k)$, akkor

$$|u - \Pi_h u|_{H^1(T_k)} \le \tilde{c}\, h_k\, \|D^2 u\|_{L^2(T_k)}, \tag{3.32}$$

ahol $\tilde{c} > 0$ T_k-tól független.

Bizonyítás. Alkalmazzuk a 3.20. állítást a $v := u - \Pi_h u$ függvényre, majd a 3.19. következményt u helyett $\widetilde{u}$-ra, végül a 3.31 becslést $v := u$ esetre:

$$|u - \Pi_h u|^2_{H^1(T_k)} = \int_{T_k} |\nabla(u - \Pi_h u)|^2 \le \frac{2h_k^2}{|T_k|} \int_E |\nabla(\widetilde{u} - \Pi_h \widetilde{u})|^2$$

$$= \frac{2h_k^2}{|T_k|}\, |\widetilde{u} - \Pi_h \widetilde{u}|^2_{H^1(E)} \le const. \cdot \frac{h_k^2}{|T_k|}\, \|D^2 \widetilde{u}\|^2_{L^2(E)},$$

$$\le const. \cdot \frac{h_k^6}{|T_k|^2}\, \|D^2 u\|^2_{L^2(T_k)} = \left(const. \cdot \frac{h_k^2}{|T_k|}\right)^2 h_k^2\, \|D^2 u\|^2_{L^2(T_k)} \le const. \cdot h_k^2\, \|D^2 u\|^2_{L^2(T_k)},$$

ahol a konstans T_k-tól független. □

Meg kell még vizsgálnunk, mikor teljesül $\frac{h_k^2}{|T_k|}$ korlátossága.

3.24. Állítás. *Tetszőleges háromszögben*

$$\frac{1}{4}\, h^2 \sin\theta \le A \le \frac{1}{2}\, h^2 \sin\theta,$$

ahol A a háromszög területe, h a háromszög leghosszabb oldala és θ a legkisebb szöge.

Bizonyítás. A legkisebb θ szöget a két leghosszabb oldal zárja be, a leghosszabb h, a második leghosszabbat jelölje b. Ekkor $h \ge b \ge \frac{h}{2}$. A h-ra merőleges magasság $m = b \sin\theta$, így $A = \frac{1}{2} hb \sin\theta$. Ebbe beírjuk b előbbi kétoldali becslését. □

3.25. Következmény. *Legyen $\mathcal{F}$ háromszögű triangulációk egy családja és ebben T_k adott háromszög. Ekkor*

$$\frac{2}{\sin\theta_k} \le \frac{h_k^2}{|T_k|} \le \frac{4}{\sin\theta_k}, \qquad \textit{azaz} \quad \frac{h_k^2}{|T_k|} = O\Big(\frac{1}{\sin\theta_k}\Big),$$

ahol θ_k a T_k legkisebb szöge.

Speciálisan, $\frac{h_k^2}{|T_k|}$ $(T_k \in \mathcal{T}_h,\ \mathcal{T}_h \in \mathcal{F})$ pontosan akkor felülről korlátos, ha $\sin\theta_k$ pozitív korlát fölött marad, azaz ha θ_k pozitív korlát fölött marad.

3.26. Definíció. Háromszögű triangulációk egy $\mathcal{F}$ családja *reguláris*, ha teljesíti az ún. minimumszög-feltételt:

$$\exists\theta_0 > 0: \quad \theta_k \ge \theta_0 \quad (\forall T_k \in \mathcal{T}_h,\ \mathcal{T}_h \in \mathcal{F}).$$

Röviden gyakran úgy mondjuk, hogy „a trianguláció reguláris”. ♢

3.27. Megjegyzés. A konvergenciabecslés folytatása előtt kitérőként érdemes a reguláris trianguláció fogalma alapján értelmezni a 3.21. megjegyzés eredményeit. E megjegyzés (i) pontja a 3.20. állítással együtt ekkor azt mutatja, hogy $\int_E |\nabla\widetilde{v}|^2$ és $\int_{T_k} |\nabla v|^2$ azonos nagyságrendűek. Ez szemléletesen azért van így, mert utóbbiban az alapterület nagyságrendileg h_k^2-szerese, az integrandus viszont h_k^{-2}-szerese az előbbiben szereplőnek. Reguláris triangulációban tehát

$$|\widetilde{v}|_{H^1(E)} = O\big(|v|_{H^1(T_k)}\big) \qquad (v \in H^1(T_k)). \tag{3.33}$$

A megjegyzés (ii) pontjának megfelelően

$$|\widetilde{v}|_{H^2(E)} = h_k \cdot O\big(|v|_{H^2(T_k)}\big) \qquad (v \in H^2(T_k)),$$

és hasonlóan adódik bármely $r \in \mathbb{N}$ esetén, hogy

$$|\widetilde{v}|_{H^{r+1}(E)} = h_k^r \cdot O\big(|v|_{H^{r+1}(T_k)}\big) \qquad (v \in H^{r+1}(T_k)). \tag{3.34}$$

A magasabbrendű becslések hátterében az áll, hogy a $\widetilde{v} := v \circ L_k$ egyenlőség deriválásaikor mindig eggyel több h_k rendű tényezővel jelennek meg a szorzatok: először a 3.20. állítás bizonyításához hasonlóan $\widetilde{\nabla} v = (\nabla v \circ L_k) \cdot J_k$, azaz ($\eta_{ij}$-vel jelölve L_k elemeit)

$$\partial_\ell(\widetilde{\nabla} v) = \sum_{i=1}^{2} (\partial_i v \circ L_k)\, \eta_{i\ell} \qquad (\ell = 1, 2),$$

majd

$$\partial_m \partial_\ell(\widetilde{\nabla} v) = \sum_{i,j=1}^{2} (\partial_j \partial_i v \circ L_k)\, \eta_{i\ell}\ \eta_{jm} \qquad (\ell, m = 1, 2),$$

és hasonlóan $r \ge 2$ esetén. Mivel J_k elemei h_k nagyságrendűek, azaz $\eta_{ij} = O(h_k)$ $(i, j = 1, 2)$, így r növelésével a nagyságrendben h kitevője is mindig eggyel nő. ♢

A 3.26. definícióval a 3.23. állításból és a 3.25. következményből adódik tehát, hogy ha a triangulációt reguláris és $u \in H^2(T_k)$, akkor teljesül a kívánt (3.32) becslés a T_k háromszögön. Utolsó lépésként összegezzük a háromszögeken kapott becsléseket, ebből adódik:

3.28. Állítás. *Ha $u \in H^2(\Omega)$ és a háromszögű triangulációt reguláris, akkor*

$$|u - \Pi_h u|_1 \leq c_1 h \, |u|_2,$$

ahol $c_1 > 0$ független a triangulációtól.

Bizonyítás. A szakasz kiindulási egyenlőségéből

$$|u - \Pi_h u|_1^2 := |u - \Pi_h u|_{H^1(\Omega)}^2 = \sum_{k=1}^{M} |u - \Pi_h u|_{H^1(T_k)}^2.$$

Itt minden háromszögön $u_{|T_k} \in H^2(T_k)$, így teljesül (3.32). Ebből és a $h_k \leq h := \max diam(T_k)$ egyenlőtlenségből

$$|u - \Pi_h u|_1^2 \leq \tilde{c}^2 \sum_{k=1}^{M} h_k^2 \, \|D^2 u\|_{L^2(T_k)}^2 \leq \tilde{c}^2 h^2 \sum_{k=1}^{M} \|D^2 u\|_{L^2(T_k)}^2$$

$$= \tilde{c}^2 h^2 \, \|D^2 u\|_{L^2(\Omega)}^2 = \tilde{c}^2 h^2 \, |u|_2^2,$$

ami a kívánt becslés négyzete. □

Elsőrendű konvergenciabecslés. A 3.28. állításból már rögtön adódik a konvergenciatétel:

3.29. Tétel. *Ha a (3.17) feladat megoldására $u^* \in H^2(\Omega)$, és a Courant-elemekben használt háromszögű triangulációt reguláris, akkor*

$$|u^* - u_h|_1 \leq ch \, |u^*|_2,$$

ahol $c > 0$ független a triangulációtól.

Bizonyítás. A 3.28. állítás és (3.27) révén

$$|u^* - u_h|_1 \leq \frac{M}{m} \, |u^* - \Pi_h u^*|_1 \leq ch \, |u^*|_2 \,,$$

ahol $c = \frac{c_1 M}{m}$. □

3.30. Megjegyzés. Speciálisan, az 1.5. tétel feltételei mellett fennáll $u^* \in H^2(\Omega)$, így tehát ha az Ω tartomány C^2-diffeomorf egy konvex tartománnyal és $p \in Lip(\overline{\Omega})$, akkor a (3.11) Dirichlet-feladat reguláris triangulációjú végeselemes megoldására teljesül, hogy $|u^* - u_h|_1 \leq ch\,|u|_2$. ♢

Egy modellfeladat végeselemes megoldását Poisson-egyenlet és homogén Dirichlet-peremfeltétel esetén a 8.2.5. és a 8.2.6. *animációk* mutatják be. A közelítő megoldások grafikonját, ill. felülnézetét (függvényértékek szerinti színezését) is szemléltetjük. Érdemes egybevetni az eredményt a 8.2.4. animációban látható véges differenciás megoldással.

3.31. Megjegyzés. Hasonló tétel érvényes más elemek esetén a trianguláció regularitásának értelemszerű átfogalmazásával:

- téglalap-elemek esetén a regularitáson azt értjük, hogy

$$\frac{h_k^+}{h_k^-} \quad (T_k \in \mathcal{T}_h,\ \mathcal{T}_h \in \mathcal{F}) \quad \text{korlátos},$$

  ahol h_k^+ és h_k^- a hosszabb ill. rövidebb oldalt jelöli T_k-ban. Mivel $\frac{h_k^+}{h_k^-} = \frac{{h_k^+}^2}{h_k^- h_k^+} = \frac{{h_k^+}^2}{|T_k|}$, teljesül a 3.23. állítás megfelelője. Átfogalmazva, van olyan $\beta > 0$, hogy $\frac{h_k^-}{h_k^+} \geq \beta$ $(\forall T_k \in \mathcal{T}_h,\ \mathcal{T}_h \in \mathcal{F})$, ezért az ilyeneket „nem-keskeny" téglalapoknak hívjuk.

- 3D-ben tetraéderek esetén a 3.26. definíció analógiájára az él- és lapszögeknek is pozitív alsó korlátja kell legyen a regularitáshoz.

- 3D-ben téglatestek esetén az egyes elemeken a leghosszabb és az egyéb oldalak hányadosai legyenek korlátosak. ♢

3.4.3. Konvergencia regularitás nélkül

Az előbbi becslésben feltettük, hogy $u \in H^2(\Omega)$. Megmutatjuk, hogy enélkül is igaz a konvergencia a $H_D^1(\Omega)$ téren, de ekkor nem kapunk nagyságrendet a konvergenciára.

3.32. Tétel. *Legyen $u^* \in H_D^1(\Omega)$ a (3.17) feladat megoldása. Ha a triangulációk családja reguláris és háromszög/tetraéder vagy téglalap/téglatest elemekből áll, akkor*

$$|u^* - u_h|_1 \to 0 \quad (ha \quad h \to 0).$$

Bizonyítás. Mivel $H^2(\Omega) \cap H_D^1(\Omega)$ sűrű $H_D^1(\Omega)$-ben, így bármely $\varepsilon > 0$ esetén van olyan $w \in H^2(\Omega) \cap H_D^1(\Omega)$, melyre $|u - w|_1 < \frac{\varepsilon}{2}$. A 3.28. állítás (ill. nem háromszögű triangulációk esetén a 3.31. megjegyzés) szerint

$$|w - \Pi_h w|_1 \leq c_1 h\,|w|_2,$$

ahol $c_1 > 0$ független a triangulációtól. Ekkor

$$|u - \Pi_h w|_1 \le |u - w|_1 + |w - \Pi_h w|_1 < \frac{\varepsilon}{2} + c_1 h\,|w|_2 < \varepsilon,$$

ha $h \le h_0 := \frac{\varepsilon}{2c_1|w|_2}$, azaz $dist(u^*, V_h) \le \varepsilon$. Így minden $\varepsilon > 0$ esetén találtunk olyan $h_0 > 0$ számot, hogy bármely $h \le h_0$ esetén $dist(u^*, V_h) \le \varepsilon$, vagyis

$$dist(u^*, V_h) \to 0 \quad (\text{ha} \quad h \to 0).$$

Ebből a 3.16. következmény szerint következik a kívánt konvergencia. □

3.5. Magasabbrendű interpoláció és konvergencia, Bramble–Hilbert-lemma

3.5.1. Interpolációs becslések H^1-normában

Ennek a pontnak a fő eredménye a 3.18. megjegyzésben felírt egydimenziós interpolációs tulajdonság magasabbrendű megfelelője. Ezt először a 3.4 (b) szakaszban bevezetett E_h h-egységháromszög-elemen (3.17. ábra) igazoljuk:

3.33. Tétel. *Ha $u \in H^{k+1}(E_h)$ és $\Pi_h u$ k-adfokú interpoláns, akkor alkalmas $c > 0$ konstans mellett*

$$|u - \Pi_h u|_{H^1(E_h)} \le c\,h^k\,|u|_{H^{k+1}(E_h)}.$$

A tételt több részben igazoljuk, ui. alapja egy általános, más esetekben is hasznos becslés, az ún. Bramble–Hilbert-lemma, valamint ennek változata. Legyen mindvégig $\Omega \subset \mathbb{R}^N$ korlátos, szakaszonként sima peremű tartomány, és jelölje P^k a legfeljebb k-adfokú n-változós polinomok halmazát.

3.34. Állítás. ***(„Bramble–Hilbert-lemma”)*** *Legyen $\phi : H^{k+1}(\Omega) \to \mathbb{R}$ korlátos lineáris funkcionál, melyre $\phi p = 0$ $(\forall p \in P^k)$. Ekkor van olyan $c > 0$ konstans, hogy*

$$|\phi u| \le c\,|u|_{H^{k+1}(\Omega)} \qquad (\forall u \in H^{k+1}(\Omega)). \tag{3.35}$$

Bizonyítás. Legyen $u \in H^{k+1}(\Omega)$ adott függvény. Könnyen látható, hogy egyértelműen létezik olyan $p \in P^k$ n-változós polinom, melyre

$$\int_\Omega \partial^\alpha p = \int_\Omega \partial^\alpha u \qquad (\forall |\alpha| \le k), \tag{3.36}$$

ahol α a lehetséges multiindexeket és $|\alpha|$ ezek rendjét jelöli. (A p polinom együtthatóit az ezek számával, $|\alpha|$-val megegyező számú egyenlőségből rekurzívan kifejezhetjük.)

Alkalmazzuk a $(k+1)$-edrendű Poincaré–Neumann-egyenlőtlenséget (lásd (1.10)) az $u-p$ függvényre: ekkor (3.36) miatt

$$\|u-p\|_{k+1}^2 \le C_{k+1}\Big(|u-p\,|_{k+1}^2 + \sum_{|\alpha|\le k}\Big|\int_\Omega \partial^\alpha(u-p)\Big|^2\Big) = C_{k+1}|u-p\,|_{k+1}^2.$$

Mivel $\phi : H^{k+1}(\Omega) \to \mathbb{R}$ korlátos lineáris funkcionál, így $|\phi v| \le M\|v\|_{k+1}$ $(\forall v \in H^{k+1}(\Omega))$ alkalmas $M > 0$ mellett. Ezt, a $\phi p = 0$ feltételt és a fenti becslést felhasználva

$$|\phi u| = |\phi(u-p)| \le M\|u-p\|_{k+1} \le MC_{k+1}^{1/2}|u-p\,|_{k+1} =: c\,|u-p\,|_{k+1}.$$

Itt a p k-adfokú polinom $(k+1)$-edrendű deriváltjai nullák, így

$$|u-p\,|_{k+1}^2 := \sum_{|\alpha|=k+1}\int_\Omega(\partial^\alpha u - \partial^\alpha p\,)^2 = \sum_{|\alpha|=k+1}\int_\Omega(\partial^\alpha u)^2 = |u|_{k+1}^2,$$

ezekből

$$|\phi u| \le c\,|u|_{k+1},$$

ahol $c > 0$ nem függ u-tól. □

3.35. Megjegyzés. A Bramble–Hilbert-lemma bizonyos értelemben a Taylor-maradéktagos becslés helyét veszi át. Ha p a lemmabeli k-adfokú polinom és $r := u - p$ jelöli a maradéktagot, akkor az u függvényt $u = p + r$ alakban írtuk fel. Ha e felbontásra alkalmazzuk ϕ-t, akkor a bal oldal ϕu, a jobb oldal pedig a ϕp tag eltűnése és ϕ korlátossága miatt az r maradéktag nagyságrendjével arányos, amely a lemma szerint $|u|_{k+1} = \|D^{k+1}u\|_{L^2}$ nagyságrendű. ♢

3.36. Állítás. ***(„Bilineáris Bramble–Hilbert-lemma”)*** *Legyen*

$$\Psi : H^{k+1}(\Omega) \times H^{k+1}(\Omega) \to \mathbb{R}$$

korlátos bilineáris forma, melyre $\Psi(r,s) = 0$, *ha* $r \in P^k$ *vagy* $s \in P^k$. *Ekkor van olyan* $c > 0$ *konstans, hogy*

$$|\Psi(u,v)| \le c\,|u|_{H^{k+1}(\Omega)}|v|_{H^{k+1}(\Omega)} \qquad (\forall u, v \in H^{k+1}(\Omega)). \tag{3.37}$$

Bizonyítás. Alkalmazzuk a Bramble–Hilbert-lemmát az első, majd a második változóban. □

A 3.33. tétel bizonyítása. Legyen először E a 3.4 (b) szakaszban bevezetett referenciaháromszög ($h = 1$ eset), és

$$\Psi(u, v) := \langle u - \Pi_h u,\ v - \Pi_h v \rangle_{H_0^1(E)} \qquad (u, v \in H^{k+1}(E)).$$

Ekkor $\Psi : H^{k+1}(E) \times H^{k+1}(E) \to \mathbb{R}$ korlátos bilineáris forma, és ha $p \in P^k$, akkor $p = \Pi_h p$, így $\Psi(p, v) = \Psi(v, p) = 0$, így alkalmazható a bilineáris Bramble–Hilbert-lemma. A (3.37) becslésből $u = v$ esetén

$$|u - \Pi_h u|^2_{H^1(E)} = \Psi(u, u) \le c\,|u|^2_{H^{k+1}(E)},$$

azaz

$$|u - \Pi_h u|_{H^1(E)} \le const. \cdot |u|_{H^{k+1}(E)}.$$

Alkalmazzuk a két oldalra a (3.33) és a (3.34) egyenlőségeket a reguláris elem szerepében E_h-ra! Ekkor az ottani h_k méretből most csak h lesz, a kitevő az ottani r helyett a most k-val jelölt érték. Így a jobb oldal h^k-os szorzót kap, azaz van olyan $c > 0$ konstans, hogy

$$|u - \Pi_h u|_{H^1(E_h)} \le c h^k\,|u|_{H^{k+1}(E_h)}. \qquad \square$$

Most már áttérhetünk az egyetlen E_h háromszögről általános triangulációk reguláris családjaira. A kapott eredmény értelemszerűen átvihető a triangulációk tetszőleges résztartományaira, majd ebből az egész tartományra, így megkapjuk a 3.28. állítás megfelelőjét:

3.37. Állítás. *Ha $u \in H^{k+1}(\Omega)$, a háromszögű triangulációs reguláris és $\Pi_h u$ elemenként k-adfokú interpoláns, akkor*

$$|u - \Pi_h u|_1 \le c_k h^k\,|u|_{k+1},$$

ahol $c_k > 0$ független a triangulációtól és $|u|_{k+1} = |u|_{H^{k+1}(\Omega)}$.

Bizonyítás. A triangulációk résztartományaira megismételhetjük a 3.33. tétel fenti bizonyítását, az utolsó lépésben az E-ről való áttéréskor felhasznált a (3.33) és (3.34) egyenlőségek reguláris triangulációk résztartományaira is vonatkoztak. Végül a résztartományokról az egész Ω-ra a 3.28. állítás bizonyításával azonos módon, egyszerű összegzéssel és az elemátmérők rendjének h-val történő becslésével jutunk el. $\square$

3.38. Megjegyzés. A $\Pi_h u$ interpolánsok k-adfokú volta a (3.19) jelöléssel azt a feltételt jelenti, hogy $P(T) \supset P^k(T)$ $(\forall T \in \mathcal{T}_h,\ \forall \mathcal{T}_h \in \mathcal{F})$. $\diamondsuit$

3.39. Megjegyzés. A 3.31. megjegyzés mintájára a fentihez hasonló tétel érvényes más elemek esetén is megfelelő reguláris triangulációk esetén. $\diamondsuit$

3.5.2. Az FEM magasabbrendű konvergenciája

A 3.37. állítás és a Céa-lemmából származtatott (3.27) becslés révén közvetlenül adódik a 3.29. tétel általánosítása:

3.40. Tétel. *Legyen $k \in \mathbb{N}^+$, és*

(i) a (3.17) feladat megoldására $u^* \in H^{k+1}(\Omega)$;

(ii) az FEM-ben használt háromszögű triangulációreguláris;

(iii) a polinomokra érvényes $P(T) \supset P^k(T)$ $(\forall T \in \mathcal{T}_h,\ \forall \mathcal{T}_h \in \mathcal{F})$.

Ekkor

$$|u^* - u_h|_1 \le c\, h^k\, |u^*|_{k+1},$$

ahol $c > 0$ független a triangulációtól.

Összefoglalva, k-adrendű (azaz a k-adfokú polinomokat tartalmazó) elemekre elég sima megoldás esetén a konvergencia rendje is k, azaz $|u^* - u_h|_1 \le O(h^k)$. A 3.2 (b) szakaszban ismertetett elemek esetén érvényes rendeket az alábbi táblázat írja le:

Elem	rend (elem és konv.)	feltétele u^*-ra
$\mathbf{T_3}$	1	H^2
$\mathbf{R_4}$	1	H^2
$\mathbf{T_6}$	2	H^3
$\mathbf{R_8}$	2	H^3
$\mathbf{H_{10}}$	3	H^4
Bell	4	H^5
Argyris	5	H^6
$\mathbf{T_4^3}$	1	H^2
$\mathbf{R_8^3}$	1	H^2
$\mathbf{T_{10}^3}$	2	H^3

3.5.3. Interpolációs becslések magasabbrendű H^ℓ-normákban

A Bramble–Hilbert-lemmából az (a) ponthoz hasonló módon levezethetők a 3.37. állításhoz hasonlóan az interpolációk rendjei akkor is, ha az $u - \Pi_h u$ eltérést nem H^1-, hanem általánosabban H^ℓ-normában $(1 \le \ell \le k)$ mérjük. Az $|u - \Pi_h u|_\ell$ norma értelmes voltához szükséges, hogy $\Pi_h u \in H^\ell(\Omega)$ legyen (u-ra ezt eleve tudjuk); mivel a T háromszögeken $\Pi_h u$ a $P(T)$ polinomhalmaz tagja, így ez a $P(T) \subset H^\ell(\Omega)$ tartalmazást jelenti az eredeti $P(T) \supset P^k(T)$ mellett.

3.41. Állítás. *Legyenek $k, \ell \in \mathbb{N}^+$ és $1 \le \ell \le k$. Tegyük fel, hogy*

(i) $u \in H^{k+1}(\Omega)$;

(ii) az FEM-ben használt háromszögű triangulációk reguláris;

(iii) a polinomokra érvényes $P^k(T) \subset P(T) \subset H^\ell(\Omega)$ $(\forall T \in \mathcal{T}_h,\ \forall \mathcal{T}_h \in \mathcal{F})$.

Ekkor

$$|u - \Pi_h u|_\ell \le c\, h^{k-\ell+1}\, |u|_{k+1},$$

ahol $c > 0$ független a triangulációtól.

Bizonyítás. Először követhetjük a 3.33. tétel bizonyítását: az E referenciaháromszögön a

$$\Psi(u, v) := \langle u - \Pi_h u,\ v - \Pi_h v \rangle_{H_0^\ell(E)} \qquad (u, v \in H^{k+1}(E))$$

bilineáris formára alkalmazzuk a bilineáris Bramble–Hilbert-lemmát, amiből szintén $u = v$ esetén

$$|u - \Pi_h u|_{H^\ell(E)} \le const. \cdot |u|_{H^{k+1}(E)}.$$

A két oldalra most a (3.34) egyenlőségeket $r := \ell - 1$ és $r := k$ mellett alkalmazzuk egy T reguláris elemen. Ekkor a bal oldal $h^{\ell-1}$-es, a jobb oldal h^k-os szorzót kap, így $h^{\ell-1}$-gyel egyszerűsítve megkapjuk a kívánt rendű becslést minden $T \in \mathcal{T}_h$ elemre. Végül ezeket a 3.37. állításhoz hasonlóan összegezzük. □

A H^ℓ-normákban való interpolációs becslésekből a konvergenciára nem kapunk új információt, mert a (3.27) alapegyenlőtlenséget csak H^1-normában tudjuk. Később, a 3.7. szakaszban fogjuk használni a H^2-normában érvényes interpolációs becslést.

3.6. További tudnivalók a konvergenciáról

3.6.1. Az FEM és FDM konvergenciájának összehasonlítása

Hasonlítsuk össze a véges differenciák módszerénél az 5-pontos sémával kapott konvergenciát (különböző simaságú megoldások esetén) a végeselem-módszer megfelelő eredményeivel! Az FDM-nél $C^k(\overline{\Omega})$-beli megoldás esetét értelemszerű az FEM-nél $H^k(\Omega)$-beli gyenge megoldással összemérni.

FDM		FEM	
feltétel u^*-ra	konv.	feltétel u^*-ra	konv.
C^1	nincs	H^1	$\to 0$
C^2	$\to 0$	H^2	$O(h)$
C^3	$O(h)$	H^3	$O(h^2)$
C^4	$O(h^2)$	H^4	$O(h^3)$

Látható, hogy adott $k = 1, \dots, 4$ rendű simaság esetén az FEM-nél elég kevesebbet (C^k helyett H^k-regularitást) megkövetelni, és így is eggyel nagyobb a konvergencia rendje.

Az viszont az FDM mellett szól, hogy az $O(h^2)$ becslést is a legegyszerűbb sémája nyújtja, míg az FEM-nél a legegyszerűbb Courant-elemekre csak $O(h)$-t kapunk, a táblázatban álló $O(h^2)$ és $O(h^3)$ csak magasabbrendű polinomok használatával érhető el. Felmerül tehát a kérdés, lehetséges-e $O(h^2)$ konvergenciát produkálni Courant-elemekkel.

Mivel a Courant-elemek $O(h)$ konvergenciabecslése H^1-normában értendő, célszerű az eggyel kisebb rendű, azaz L^2-normát vizsgálni. Ezzel valóban $O(h^2)$ konvergenciát érhetünk el, ezt a következő szakaszban mutatjuk meg.

3.6.2. Nitsche-trükk, L^2-konvergenciabecslés

Az L^2-konvergenciabecslésnél az ún. adjungált feladat segítségével lehet megmutatni, hogy egy rendet nyerhetünk a H^1-konvergenciához képest. Ezt hívják Nitsche-trükknek (vagy Aubin-Nitsche-trükknek).

Ez röviden leírható a megszokott absztakt formalizmus használatával. A jobboldali ℓ funkcionál helyett azonban a megfelelő függvényeket kell feltüntetnünk.

Tekintsünk egy

$$Lu = f$$

elliptikus feladatot adott $f \in L^2(\Omega)$ mellett, és legyen ennek gyenge megoldása $u^* \in H$, azaz

$$a(u^*, v) = \langle f, v \rangle_0 \qquad (\forall v \in H),$$

ahol $\langle ., . \rangle_0$ az L^2-skalárszorzatot jelöli, és $H \subset H^1(\Omega)$ a feladathoz tartozó Szoboljev-tér ($H_0^1(\Omega)$, $H_D^1(\Omega)$ vagy $\dot{H}^1(\Omega)$). Az a bilineáris forma teljesíti a szokásos korlátosságot és koercivitást, azaz igaz (3.1)–(3.2). Legyen $V_h \subset H$ végeselemes altér, és ebben $u_h \in V_h$ a végeselemes megoldás.

Tekintsük az

$$Lz = u^* - u_h$$

ún. adjungált feladatot. (Ennek jobboldalát természetesen nem ismerjük, a feladat csak a becslés elméleti célját szolgálja.)

3.42. Tétel. *Tegyük fel, hogy*

(i) $u^* \in H^2(\Omega)$

(ii) $z^* \in H^2(\Omega)$, *és* $\exists c_1 > 0$: $\quad |z^*|_2 \le c_1 \|u^* - u_h\|_0$.

Ekkor

$$\|u^* - u_h\|_0 \le c\, h^2\, |u^*|_2,$$

ahol $c > 0$ független a triangulációtól.

Bizonyítás. Az adjungált feladat gyenge megoldására

$$a(z^*, v) = \langle u^* - u_h, v\rangle_0 \qquad (\forall v \in H).$$

Legyen $v := u^* - u_h$ és használjuk fel az $a(u^* - u_h, z_h) = 0$ Galjorkin-ortogonalitást (3.5. állítás), akkor

$$\|u^* - u_h\|_0^2 = a(z^*, u^* - u_h) = a(z^* - z_h, u^* - u_h) \le M|z^* - z_h|_1|u^* - u_h|_1.$$

Mivel $u^*, z^* \in H^2(\Omega)$, mindkettőre érvényes az elsőrendű konvergencia, vagyis a 3.29. tétel. Ezt és a (ii) feltételt használva

$$M|z^* - z_h|_1|u^* - u_h|_1 \le M\, c^2 h^2\, |z^*|_2|u^*|_2 \le M\, c^2 c_1 h^2\, \|u^* - u_h\|_0|u^*|_2.$$

Ezekből a $c := M\, c^2 c_1$ jelöléssel

$$\|u^* - u_h\|_0^2 \le c\, h^2\, \|u^* - u_h\|_0|u^*|_2,$$

amiből következik a kívánt becslés. □

3.43. Megjegyzés. Az (i)-(ii) feltételek teljesülnek pl. Dirichlet-feladat esetén akkor, ha Ω C^2-diffeomorf egy konvex tartománnyal és $p \in Lip(\overline{\Omega})$, lásd 1.5. tétel. ♢

3.6.3. A numerikus integrálás hatása

A végeselemes megoldás együtthatóinak kiszámítása az $A_h c = b_h$ lineáris algebrai egyenletrendszer megoldását igényli, ahol A_h és b_h elemei integrálok, ahogy ezt a 3.2. szakaszban felírtuk. A gyakorlatban azonban ezeket az integrálokat csak közelítőleg, numerikusan tudjuk kiszámítani (néhány egészen speciális eset kivételével).

Ebben a pontban megvizsgáljuk, milyen feltételeket kell teljesítenie a használt numerikus integrálási kvadratúráknak ahhoz, hogy a módszer kövesse az elméletet, ami elsősorban azt jelenti, hogy szeretnénk megőrizni a konvergenciára elvileg kapott rendet.

Általában véve, egy T_j elemen vett kvadratúra egy

$$Q_j(f) \approx \int_{T_j} f$$

közelítést jelent, ahol

$$Q_j(f) := \sum_{i=1}^{s} w_i\, f(x_i)$$

alkalmas $x_1, \ldots, x_s \in T_j$ csomópontok és $w_1, \ldots, w_s$ súlyok mellett. Az egyszerűség kedvéért fel fogjuk tenni, hogy $w_i > 0 \ \forall i$. Az elemeken vett kvadratúrákból állítjuk össze a

$$Q(f) := \sum_{j=1}^{M} Q_j(f) \approx \int_\Omega f$$

globális kvadratúrát.

A kvadratúrákra vett első szükséges kritérium, hogy az ezzel való közelítés megőrizze a norma-tulajdonságot, amiből az a nem nyilvánvaló rész, hogy csak triviálisan lehessen 0, azaz

$$Q(|\nabla u_h|^2) \approx \int_\Omega |\nabla u_h|^2 = |u_h|_1^2 > 0$$

miatt legyen

$$Q(|\nabla u_h|^2) > 0 \qquad (\forall u_h \in V_h, \ u_h \neq 0). \tag{3.38}$$

Ezt az garantálja biztosan, ha V_h elemeire ez a közelítés elemenként és így az egész Ω-n is pontos, azaz ha

$$Q_j(|\nabla u_h|^2) = \int_{T_j} |\nabla u_h|^2 \qquad (\forall j = 1, \ldots, M). \tag{3.39}$$

Ha k-adfokú polinomokat használunk V_h-ban, akkor a fenti kifejezésben u_h k-adfokú polinom, ∇u_h koordinátái $(k-1)$-edfokú polinomok és végül $|\nabla u_h|^2$ $(2k-2)$-edfokú polinom minden T_j-n, azaz

$$|\nabla u_h|^2_{|T_j} \in P^{2k-2}(T_j).$$

Ebből adódik a továbbiakban feltett ún.

Egzaktsági feltétel: a kvadratúra elemenként legyen pontos minden $(2k-2)$-edfokú polinomon, azaz

$$Q_j(p) = \int_{T_j} p \qquad (\forall p \in P^{2k-2}(T_j), \ j = 1, \ldots, M).$$

A pontosabb tárgyaláshoz tekintsük a (3.11) homogén Dirichlet-feladatot. A végeselemes feladat az

$$a(u,v) := \int_\Omega p \ \nabla u \cdot \nabla v, \qquad \ell v := \int_\Omega f v$$

forma ill. funkcionál mellett

$$a(u_h, v_h) = \ell v_h \qquad (\forall v_h \in V_h), \tag{3.40}$$

Tekintsük most a kvadratúrákkal módosított feladatot, azaz legyen

$$a_h(u,v) := Q(p \ \nabla u \cdot \nabla v), \qquad \ell_h v := Q(fv),$$

és az ezekből kapott végeselemes megoldás legyen $\overline{u}_h$, azaz

$$a_h(\overline{u}_h, v_h) = \ell_h v_h \qquad (\forall v_h \in V_h), \tag{3.41}$$

Célunk, hogy erre is

$$|u^* - \overline{u}_h|_1 \le O(h^k)$$

legyen, ha ez u_h-ra igaz volt.

A módosított feladatra az első követelmény, hogy a_h is koercív legyen V_h-ban. (A korlátosság a véges dimenzió miatt nyilvánvaló.)

3.44. Állítás. *Ha $w_i > 0$ $(\forall i = 1, \dots, s)$ és fennáll az egzaktsági feltétel, akkor a_h koercív V_h-ban.*

Bizonyítás. A $w_i > 0$ és $p(x) \ge m > 0$ feltételekből, valamint az egzaktsági feltételből bármely $v_h \in V_h$ esetén

$$a_h(v_h, v_h) = Q(p\,|\nabla v_h|^2) = \sum_{j=1}^{M} Q_j(p\,|\nabla v_h|^2) = \sum_{j=1}^{M}\sum_{i=1}^{s} w_i\, p(x_i)\,|\nabla v_h(x_i)|^2$$

$$\ge m \sum_{j=1}^{M}\sum_{i=1}^{s} w_i\,|\nabla v_h(x_i)|^2 = m\,Q(|\nabla v_h|^2) = m\,|v_h|_1^2.$$

□

3.45. Következmény. *A (3.41) feladatnak egyértelműen létezik $\overline{u}_h \in V_h$ megoldása.*

A rend becslése az alábbi tulajdonságon múlik:

3.46. Állítás. *Ha $w_i > 0$ $(\forall i = 1, \dots, s)$ és fennáll az egzaktsági feltétel, akkor van olyan $c > 0$, hogy*

$$|u_h - \overline{u}_h|_1 \le c\,(\|a - a_h\| + \|\ell - \ell_h\|).$$

Bizonyítás. A feltételek miatt az előző állítás szerint a_h koercív V_h-ban. Így

$$m|u_h - \overline{u}_h|_1^2 \le a_h(u_h - \overline{u}_h, u_h - \overline{u}_h) = a_h(u_h, u_h - \overline{u}_h) - a_h(\overline{u}_h, u_h - \overline{u}_h).$$

Itt a második tagra (3.41) és $v_h := u_h - \overline{u}_h \in V_h$ miatt

$$a_h(\overline{u}_h, u_h - \overline{u}_h) = \ell_h(u_h - \overline{u}_h).$$

Az első tagra

$$a_h(u_h, u_h - \overline{u}_h) = (a_h - a)(u_h, u_h - \overline{u}_h) + a(u_h, u_h - \overline{u}_h),$$

ennek második tagjára (3.40) miatt

$$a(u_h, u_h - \overline{u}_h) = \ell(u_h - \overline{u}_h).$$

Együtt

$$\begin{aligned} m|u_h - \overline{u}_h|_1^2 &\le (a_h - a)(u_h, u_h - \overline{u}_h) + (\ell - \ell_h)(u_h - \overline{u}_h) \\ &\le \|a_h - a\|\,|u_h|_1\,|u_h - \overline{u}_h|_1 + \|\ell - \ell_h\|\,|u_h - \overline{u}_h|_1 \\ &= \big(\|a_h - a\|\,|u_h|_1 + \|\ell - \ell_h\|\big)\,|u_h - \overline{u}_h|_1. \end{aligned}$$

Itt (3.25) révén $|u_h|_1 \le \frac{C_\Omega}{m}\|f\|_0$. Így

$$m|u_h - \overline{u}_h|_1 \le \max\Big\{\frac{C_\Omega}{m}\|f\|_0,\ 1\Big\}\big(\|a_h - a\| + \|\ell - \ell_h\|\big),$$

tehát az állítás teljesül. □

3.47. Következmény. *Ha* $\|a - a_h\| + \|\ell - \ell_h\| \le O(h^k)$, *akkor érvényes* $|u^* - \overline{u}_h|_1 \le O(h^k)$.

Az $\|a - a_h\|$ és $\|\ell - \ell_h\|$ normák becslése a kvadratúrák pontosságától függ. A feladat együtthatóinak kellő simasága esetén a módszerre eddig tett feltételek elegendőek a kívánt rendhez. Az erre vonatkozó, hosszabb számolást igénylő eredményt a kvadratúrák elméletéből itt bizonyítás nélkül mondjuk ki:

3.48. Tétel. ***(Ciarlet, [8])*** *Tegyük fel, hogy*

(i) a (3.11) feladatban $p \in C^k(\overline{\Omega})$, $u \in H^{k+1}(\Omega)$, *és* $\exists q > \frac{n}{k}$: $f \in W^{k,q}(\Omega)$;

(ii) a háromszögű triangulációreguláris, és k*-adfokú Lagrange-elemeket használunk;*

(iii) $w_i > 0$ $(\forall i = 1, \dots, s)$ *és fennáll az egzaktsági feltétel.*

Ekkor van olyan $c > 0$, *hogy*

$$\|a - a_h\| + \|\ell - \ell_h\| \le c\,h^k\,(\|u\|_{k+1} + \|f\|_{W^{k,q}}).$$

3.49. Következmény. *A 3.48. tétel feltételei mellett*

$$|u^* - \overline{u}_h|_1 \le O(h^k).$$

A gyakorlatban a kvadratúra választásának fő szempontja tehát az egzaktsági feltétel teljesítése. Néhány egyszerű példa:

2D lineáris elemek ($\mathbf{T_3}$) esetén $k = 1$, azaz $2k - 2 = 0$, vagyis elég a konstansokat pontosan integrálni. Erre megfelelő az egypontos súlyponti kvadratúra.

2D kvadratikus elemek ($\mathbf{T_6}$) esetén $k = 2$, azaz $2k - 2 = 2$, azaz ekkor a legfeljebb másodfokú polinomokat kell pontosan integrálni. Erre megfelelő az oldalfelező pontokra illesztett 3-pontos kvadratúra.

3D lineáris elemek ($\mathbf{T_4^3}$) esetén $k = 1$, így $\mathbf{T_3}$-hoz hasonlóan megfelelő az egypontos súlyponti kvadratúra.

További alkalmas kvadratúrák pl. [19]-ben találhatók.

3.6.4. Rácsfinomítás, adaptív végeselem-módszer

A végeselem-módszer megvalósításában fontos kérdés, milyen finomságú rácsot érdemes használni és honnan tudjuk, hogy elfogadható pontosságú-e a numerikus megoldásunk, ill. milyen finomítással javítható a pontosság.

Utóbbi megállapításának eszközei az ún. *a posteriori hibabecslések*, amelyek egy adott rácson vett u_h numerikus megoldás pontosságát becslik felülről. Egyszerűség kedvéért ebben a pontban a (3.11) alakú feladatra szorítkozunk:

$$\begin{cases} Lu := -\operatorname{div}(p\,\nabla u) = f, \\ u_{|\partial\Omega} = 0, \end{cases}$$

ahol $\Omega \subset \mathbb{R}^2$ korlátos tartomány.

Az alapprobléma természetesen az, hogy a valódi

$$|u_h - u^*|_1$$

hiba nem számítható ki, hiszen nem ismerjük u^*-ot. Ha $u^* \in H^2(\Omega)$ és $u_h \in H^2(\Omega)$ (amit igen ritkán konstruálunk így, hiszen ehhez, mint láttuk, 5-ödfokú elemek kellenek), akkor egészen egyszerű becslés adható, mert értelmes $Lu_h - f$. Ekkor a Green-formulából

$$m\,|u_h - u^*|_1^2 \le a(u_h - u^*, u_h - u^*) = \int_\Omega p\,|\nabla(u_h - u^*)|^2 = \int_\Omega (Lu_h - Lu^*)(u_h - u^*)$$

$$= \int_\Omega (Lu_h - f)(u_h - u^*) \le \|Lu_h - f\|_{L^2(\Omega)} \|u_h - u^*\|_{L^2(\Omega)} \le C_\Omega \|Lu_h - f\|_{L^2(\Omega)} |u_h - u^*|_1$$

(ahol a Poincaré–Friedrichs-egyenlőtlenséget használtuk), így

$$|u_h - u^*|_1 \le \frac{C_\Omega}{m} \|Lu_h - f\|_{L^2(\Omega)},$$

ami kiszámítható becslés.

Az általános $u_h \in H_0^1(\Omega)$, $u_h \notin H^2(\Omega)$ esetben a Green-formulát csak résztartományonként alkalmazhatjuk és peremtagok is bejönnek. Ilyenkor az $|u_h - u^*|_1$ hiba helyett az

$$R(u_h)v := \ell v - a(u_h, v) \qquad (v \in H_0^1(\Omega))$$

reziduális hibafunkcionál normáját szokás vizsgálni. Könnyen látható, hogy a koercivitás miatt

$$|u_h - u^*|_1 \le \frac{1}{m}\,\|R(u_h)\|, \tag{3.42}$$

lásd 9.21. feladat. Ekkor igazolható [4], hogy reguláris trianguláció esetén alkalmas $c > 0$ állandó mellett

$$\|R(u_h)\|^2 \;\le\; c\left(\sum_T h_T^2 \|Lu_h - f\|_{L^2(T)}^2 + \sum_e h_e^2 \|[p\,\partial_\nu u_h]\|_{L^2(e)}^2\right),$$

ahol T a trianguláció résztartományait, e az éleket, $[p\partial_\nu u_h]$ pedig a $p\partial_\nu u_h$ függvény élen vett ugrását jelöli.

A kapott becslést elsősorban arra használják, hogy megmérje, a tartomány mely részén nagyobb a hiba. Ahol a szummában nagyobb tagok szerepelnek (a „nagy” pl. azt jelenti, hogy a legnagyobb tag q-szorosánál nagyobb, ahol $0 < q < 1$ általunk választott küszöb), az a rész okozza inkább a hibát.

Erre alapszik az ún. *adaptív végeselem-módszer*: a rácsot nem egyenletesen finomítjuk, hanem csak ott, ahol a hiba nagy a fenti értelemben, és ezt ismételjük. Az adaptív végeselem-módszer egy ciklusának lépései tehát: az aktuális rácson megoldjuk a feladatot, majd a fenti hibabecslés alapján megjelöljük azokat az elemeket, ahol a hiba nagy, végül ezeken finomítunk. Tömören:

$$\text{megoldás} \to \text{hibabecslés} \to \text{megjelölés} \to \text{finomítás}.$$

Egy modellfeladat végeselemes megoldását Laplace-egyenlet és Dirichlet-peremfeltétel esetén a 8.2.7. és a 8.2.8. *animációk* mutatják be. Jól látható, hogy a konkáv sarok környezetében jobban kell finomítani az elemeket; ez megfelel annak, hogy (amint az 1.2.2. szakaszban említettük) a konkáv sarok csökkenti a megoldás regularitását.

A finomítás során fontos, hogy a rácsok kapott családja reguláris legyen, mert a becslések erre érvényesek. Ennek egy elegáns módja a „leghosszabb él felezésének módszere”, azaz ha az új csomópontok a háromszögek leghosszabb élének felezőpontjai: ez reguláris családot hoz létre [17], és ez a módszer 3D-ben is működik.

3.7. Nem szimmetrikus és negyedrendű egyenletek

3.7.1. Nem szimmetrikus másodrendű egyenletek

Az eddig vizsgált peremérték-feladatok szimmetrikusak voltak abban az értelemben, hogy a másod- és nulladrendű deriváltakat tartalmazó operátor szimmetrikus, és ennek megfelelően a hozzá tartozó bilineáris forma is szimmetrikus. Ha az egyenlet elsőrendű tagot is tartalmaz, ami a gyakorlatban konvekció típusú mennyiségeket (pl. szél) fejez ki, akkor ez a szimmetria már nem érvényes. Az eddigi elmélet a megoldhatóság és a végeselem-módszer konstrukciója szempontjából enyhe módosításokkal hasonlóan alkalmazható lesz.

Dirichlet-feladat. Legyen $\Omega \subset \mathbb{R}^n$ korlátos tartomány szakaszonként sima peremmel, és tekintsük az alábbi feladatot:

$$\begin{cases} -\operatorname{div}(p\,\nabla u) + \mathbf{w}\cdot\nabla u = f, \\ u_{|\partial\Omega} = 0. \end{cases} \tag{3.43}$$

3.50. Feltevés.

(i) $p \in L^\infty(\Omega)$, $p(x) \geq m > 0$ (m. m. $x \in \Omega$);

(ii) $\mathbf{w} \in C^1(\overline{\Omega}, \mathbb{R}^n)$, $\operatorname{div} \mathbf{w} = 0$ (azaz $\mathbf{w}$ divergenciamentes vektormező). ♢

A gyenge megoldás fogalmát az előző szakaszhoz hasonlóan értelmezzük: olyan $u \in H_0^1(\Omega)$ függvényt keresünk, melyre

$$\int_\Omega \Big(p\, \nabla u \cdot \nabla v + (\mathbf{w} \cdot \nabla u) v \Big) = \int_\Omega f v \qquad (\forall v \in H_0^1(\Omega)). \tag{3.44}$$

A Lax–Milgram-lemmát szeretnénk használni. Legyen $B : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ az alábbi bilineáris forma:

$$B(u,v) := \int_\Omega \Big(p\, \nabla u \cdot \nabla v + (\mathbf{w} \cdot \nabla u) v \Big).$$

3.51. Állítás. *A 3.50. feltételek mellett a B bilineáris forma korlátos és koercív.*

Bizonyítás. Egyrészt

$$|B(u,v)| \leq \|p\|_{L^\infty} |u|_1 \, |v|_1 + \|\mathbf{w}\|_{L^\infty} |u|_1 \, \|v\|_0 \leq \big(\|p\|_{L^\infty} + C_\Omega \|\mathbf{w}\|_{L^\infty} \big) \, |u|_1 \, |v|_1,$$

ahol az (1.8) Poincaré–Friedrichs-egyenlőtlenséget használtuk, így B korlátos. A koercivitáshoz felhasználjuk az alábbi azonosságokat:

$$\operatorname{div}(\mathbf{w} u^2) = (\operatorname{div} \mathbf{w})\, u^2 \, + \mathbf{w} \cdot \nabla(u^2) = (\operatorname{div} \mathbf{w})\, u^2 \, + 2(\mathbf{w} \cdot \nabla u) u = 2(\mathbf{w} \cdot \nabla u) u$$

(a $\operatorname{div} \mathbf{w} = 0$ feltevésből), így a Gauss–Osztrogradszkij-tételből és $u_{|\partial\Omega} = 0$ révén

$$\int_\Omega 2(\mathbf{w} \cdot \nabla u) u = \int_\Omega \operatorname{div}(\mathbf{w} u^2) = \int_{\partial\Omega} (\mathbf{w} u^2) \cdot \nu \, = 0, \tag{3.45}$$

tehát $\int_\Omega (\mathbf{w} \cdot \nabla u) u = 0$. Ebből

$$\langle Bu, u \rangle_{H_0^1} = \int_\Omega \Big(p\, |\nabla u|^2 + (\mathbf{w} \cdot \nabla u) u \Big) = \int_\Omega p\, |\nabla u|^2 \geq m |u|_1^2 \qquad (\forall u \in H_0^1(\Omega)), \tag{3.46}$$

így B koercív is. Összességében B határai

$$M := \|p\|_{L^\infty} + C_\Omega \|\mathbf{w}\|_{L^\infty}, \qquad m := \operatorname{ess\,inf} p. \tag{3.47}$$

□

Másrészt $\phi v := \int_\Omega f v$ korlátos lineáris funkcionál a $H_0^1(\Omega)$ téren (ugyanúgy, mint az eddigi szimmetrikus feladatokban, hisz ez nem függ az operátortól). Így a Lax–Milgram-lemma alapján teljesül a megoldhatósági eredmény:

3.52. Következmény. *Ha teljesülnek a 3.50. feltételek, akkor a (3.43) peremértékfeladatnak bármely $f \in L^2(\Omega)$ esetén egyértelműen létezik $u^* \in H_0^1(\Omega)$ gyenge megoldása.*

3.53. Megjegyzés.

(i) A tétel nulladrendű taggal együtt is igazolható a bizonyítás értelemszerű módosításával: egyrészt, ha $Lu := -\operatorname{div}(p\,\nabla u) + \mathbf{w}\cdot\nabla u + cu$, ahol $c \in L^\infty(\Omega)$ és $c \geq 0$; általánosabban pedig, ha a $\operatorname{div}\mathbf{w} = 0$ és $c \geq 0$ feltételek helyett a $c - \frac{1}{2}\operatorname{div}\mathbf{w} \geq 0$ egyenlőtlenség teljesül.

(ii) Ha az Ω tartomány C^2-diffeomorf egy konvex tartománnyal, és $p \in Lip(\overline{\Omega})$ (pl. $p \in C^1(\overline{\Omega})$), akkor $u \in H^2(\Omega)$. Ez következik abból, ha az 1.5. tételt az $\tilde{f} := f - \mathbf{w}\cdot\nabla u$ jobboldallal alkalmazzuk. ♢

Most már hozzáfoghatunk a végeselem-módszer alkalmazásához. Általánosságban a Galjorkin-módszer szerint tekintsünk egy alkalmas

$$V_h \subset H_0^1(\Omega)$$

véges dimenziós alteret, és ebben keressük a vetületi egyenlet megoldását: $u_h \in V_h$, melyre

$$\int_\Omega \Big(p\,\nabla u_h \cdot \nabla v_h + (\mathbf{w}\cdot\nabla u_h)v_h \Big) = \int_\Omega f v_h \qquad (\forall v_h \in V_h). \tag{3.48}$$

Ha $\varphi_1, \ldots, \varphi_n$ bázis V_h-ban és a közelítő megoldást a szokott $u_h = \sum_{j=1}^n c_j\varphi_j$ alakban keressük, akkor a megfelelő

$$A_h c = b_h$$

lineáris algebrai egyenletrendszerben

$$a_{ij} = \int_\Omega \Big(p\,\nabla\varphi_j \cdot \nabla\varphi_i + (\mathbf{w}\cdot\nabla\varphi_j)\varphi_i \Big) \qquad \text{és} \qquad b_i = \int_\Omega f\varphi_i \qquad (i,j = 1,\ldots,n).$$

(A lineáris algebrai egyenletrendszer most is egyértelműen megoldható a koercivitás és a 3.2. következmény miatt.) Most A_h nem szimmetrikus, így a_{ij} második tagjában fontos φ_j és φ_i sorrendje.

A V_h altér konkrét konstrukciójához ugyanazokat a végeselem-típusokat használhatjuk, mint a szimmetrikus feladatoknál, lásd a 3.2. szakasz (b) pontjában.

A konvergencia alapja most is (a 3.4. szakasz (a) pontjához hasonlóan) a B bilineáris forma korlátos és koercív voltából következő Céa-lemma:

3.54. Következmény. *A (3.43) feladat $u_h \in V_h$ végeselemes megoldására*

$$|u^* - u_h|_1 \leq \frac{M}{m} \min_{v_h \in V_h} |u^* - v_h|_1, \tag{3.49}$$

ahol M és m a (3.18) bilineáris forma határai (lásd (3.47)).

Ennek pedig döntő következménye, hogy a konvergencia vizsgálata innentől függetleníthető a (3.43) feladattól: a (3.27) mintájára az

$$|u^* - u_h|_1 \leq \frac{M}{m} |u^* - \Pi_h u^*|_1 \,. \tag{3.50}$$

jobb oldalán álló interpolációs hiba becslésére van szükségünk. Ez viszont pontosan ugyanaz, mint amit már ismerünk a 3.4. szakaszból! Azaz, változatlan formában igazak az ottani konvergenciabecslések.

3.55. Következmény. *Tekintsük a (3.17) feladatot a 3.50. feltételekkel, és a (3.48) FEM-es megoldását reguláris triangulációmellett.*

(1) Ha a megoldásra $u^* \in H^2(\Omega)$, akkor

$$|u^* - u_h|_1 \leq ch\, |u^*|_2,$$

ahol $c > 0$ független a triangulációtól. (Ez pl. igaz a 3.53. megjegyzés (ii) pontjának helyzetében.)

(2) Ha a megoldásra $u^* \in H^{k+1}(\Omega)$ (ahol $k \in \mathbb{N}^+$), és a polinomokra érvényes $P(T) \supset P^k(T) \quad (\forall T \in \mathcal{T}_h, \forall \mathcal{T}_h \in \mathcal{F})$, akkor

$$|u^* - u_h|_1 \leq c\, h^k\, |u^*|_{k+1},$$

ahol $c > 0$ független a triangulációtól.

Vegyes peremfeltétel. Tekintsük most az alábbi feladatot:

$$\begin{cases} -\operatorname{div}(p\,\nabla u) + \mathbf{w} \cdot \nabla u = f, \\ u_{|\Gamma_D} = 0, \quad (p\,\partial_\nu u + su)_{|\Gamma_N} = \gamma, \end{cases} \tag{3.51}$$

ahol továbbra is $\Omega \subset \mathbb{R}^n$ korlátos tartomány szakaszonként sima peremmel, és teljesül:

3.56. Feltevés.

(i) Teljesülnek a 3.50. feltételek.

(ii) $s \in L^\infty(\Gamma_N)$, $s \geq 0$ m. m., $\gamma \in L^2(\Gamma_N)$.

(iii) $\mathbf{w} \cdot \nu \geq 0$ a Γ_N felületen, ahol ν a külső normális. ♢

A gyenge alak ekkor

$$\int_\Omega \Big(p\, \nabla u \cdot \nabla v + (\mathbf{w} \cdot \nabla u) v \Big) + \int_{\Gamma_N} suv = \int_\Omega fv + \int_{\Gamma_N} \gamma v \qquad (\forall v \in H^1_D(\Omega)). \tag{3.52}$$

A bal oldal által meghatározott bilineáris forma korlátos és koercív volta az előbbi Dirichlet-feladat és a (3.17) szimmetrikus vegyes feladat mintájára következik, az egyetlen új tag a $\mathbf{w}$ függvény peremen való viselkedése miatt lép fel: most ui. (3.45) nem lesz 0, hiszen a peremnek u csak a Γ_D részfelületén tűnik el, azaz

$$\int_\Omega 2(\mathbf{w}\cdot\nabla u)u = \int_\Omega \operatorname{div}(\mathbf{w}u^2) = \int_{\partial\Omega}(\mathbf{w}u^2)\cdot\nu = \int_{\Gamma_N}(\mathbf{w}\cdot\nu)u^2. \tag{3.53}$$

A 3.56. (iii) feltétel alapján viszont ebből következik az, hogy

$$\int_\Omega (\mathbf{w}\cdot\nabla u)u \geq 0.$$

Mivel $B(u,u)$ maradék része megegyezik a szimmetrikus vegyes feladatéval, melyre már tudjuk a koercivitást, így a nemnegatív új tag ezt nem rontja el, azaz

$$B(u,u) = \int_\Omega p\,|\nabla u|^2 + \int_{\Gamma_N} su^2 + \int_\Omega (\mathbf{w}\cdot\nabla u)u \geq \int_\Omega p\,|\nabla u|^2 + \int_{\Gamma_N} su^2 \geq m\,|u|_1^2$$

(ha $u \in H^1_D(\Omega)$).

Innen már ugyanúgy haladunk tovább, azaz az értelemszerűen konstruált végeselemes megoldásra ugyanúgy érvényes a Céa-lemma és ennek révén a 3.55. következménybeli konvergenciabecslések.

3.57. Megjegyzés. Mit jelent és természetellenes megszorítás-e a 3.56. feltétel (iii) pontjában szereplő egyenlőtlenség, hogy

$$\mathbf{w}\cdot\nu \geq 0 \quad \text{a } \Gamma_N \text{ felületen?} \tag{3.54}$$

Ennek jelentéséhez idézzük fel az elsőrendű lineáris parciális differenciálegyenletekről ismert tulajdonságokat. (Elsőrendű lineáris parciális differenciálegyenletekről időfüggő kontextusban a könyv második részében lesz szó részletesebben, most stacionárius feladatként vizsgáljuk.) A szemléletesség kedvéért legyen $\Omega \subset \mathbb{R}^2$. Tekintsük a (3.51) egyenlet diffúziós tag nélküli részét Ω-ban, először peremfeltétel nélkül:

$$\mathbf{w}\cdot\nabla u = f.$$

Ismeretes (lásd pl. [9]), hogy ha bevezetjük az ehhez tartozó karakterisztikus görbéket (vagy más szóval áramvonalakat), azaz a

$$\dot{\xi}(t) = \mathbf{w}(\xi(t))$$

közönséges differenciálegyenlet-rendszer megoldásait, akkor

$$(u\circ\xi)^{\cdot}(t) = \nabla u(\xi(t))\cdot\mathbf{w}(\xi(t)) = f(\xi(t))$$

bármely karakterisztikus görbe bármely értékére, és így a karakterisztikus görbék mentén meghatározhatók u értékei a peremmel vett metszéspontokon vett értékekből. Éspedig, az $f \equiv 0$ homogén esetben u állandó a karakterisztikus görbék mentén, az $f \equiv 0$ inhomogén esetben pedig integrálni kell az $f \circ \xi$ függvényt ξ mentén a peremmel vett metszésponttól a kiválasztott pontig.

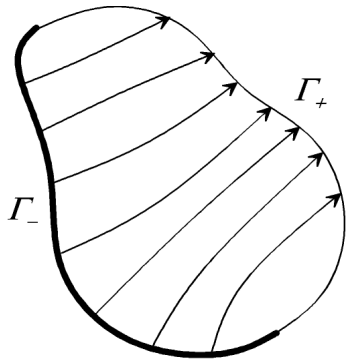
Ez azt jelenti, hogy u értékeit Ω-ban meghatározzák az Ω-ba belépő karakterisztikus görbék perempontjain vett értékei. Legyen

$$\Gamma_- := \{x \in \partial\Omega : \ \mathbf{w}(x) \cdot \nu(x) < 0\}, \qquad \Gamma_+ := \{x \in \partial\Omega : \ \mathbf{w}(x) \cdot \nu(x) \geq 0\}.$$

Ha a Γ_--on áthaladó karakterisztikus görbék egyrétűen lefedik Ω-t, akkor u_{Γ_-} meghatározza u értékeit Ω-ban, azaz egyértelműen megoldható a

$$\begin{cases} \mathbf{w} \cdot \nabla u = f, \\ u_{|\Gamma_-} = 0 \end{cases}$$

peremérték-feladat, lásd 3.19. ábra.



3.19. ábra. Konvekciós peremfeltételek.

Tekintsük most a (3.51) feladatot! Ekkor (3.54) azt jelenti, hogy

$$\Gamma_- \subset \Gamma_D.$$

A másodrendű elliptikus feladatban (azaz a $-\operatorname{div}(p\,\nabla u)$ diffúziós tag hozzávétele után) az egész peremen meg kellett adni feltételt az egyértelmű megoldhatósághoz. A fenti tartalmazás azt jelenti, hogy ahol az elsőrendű feladatban megadtunk függvényértéket a peremen (azaz a Γ_- halmazon), ott az elliptikus esetben is függvényértéket adunk meg, azaz Dirichlet-peremfeltételt. (A fennmaradó Γ_+ halmazon lehet Dirichlet-féle vagy vegyes is.) A (3.54) feltétel ebben az értelemben természetes.

Végül megjegyezzük, hogy ha a diffúziós tag kicsi (ún. konvekció-dominált feladat), akkor a karakterisztikus görbék mentén egy darabig közel lesz az első- és másodrendű

feladat megoldása, viszont ha a Γ_+ halmazon is Dirichlet- peremfeltételt adunk meg (amely független a karakterisztikus görbék mentén vett értéekektől), akkor ott nagy eltérés mutatkozhat. Ez hirtelen, nagy deriváltértékekkel is jelentkezhet, azaz itt torzítja el jelentősen a diffúziós tag hozzávétele az elsőrendű feladat megoldását. Erről a következő pontban lesz szó. ♢

Konvekció-dominált feladatok. Tekintsük a (3.43) feladat speciális esetét:

$$\begin{cases} -\varepsilon\Delta u + \mathbf{w}\cdot\nabla u = f, \\ u_{|\partial\Omega} = 0, \end{cases} \tag{3.55}$$

ahol $\varepsilon > 0$ állandó. Itt a $-\varepsilon\Delta u$ tag a diffúziót, a $\mathbf{w}\cdot\nabla u$ tag a konvekciót írja le. Gyakran a diffúziós tag igen kicsi ($\varepsilon \approx 0$), ekkor a feladatot konvekció-dominántnak hívjuk.

Az $\varepsilon \approx 0$ tulajdonság a megoldás jellegzetes viselkedéséhez vezet, amelynek kezelése nem nyilvánvaló. Amint az előző pont végén említettük, ha a diffúziós tag kicsi, akkor a karakterisztikus görbék mentén egy darabig közel lesz az első- és másodrendű feladat megoldása, viszont a Γ_+ halmaz közelében (ahol az előírt Dirichlet-peremfeltétel független a karakterisztikus görbék mentén vett értékektől) nagy eltérés mutatkozhat. A tartomány ezen részét *határrétegnek* hívjuk, ahol nehezen közelíthető nagy deriváltértékek adódnak.

Ezt jól szemlélteti az alábbi speciális 1D eset:

$$-\varepsilon u'' + u' = 1, \qquad u(0) = u(1) = 0.$$

Ennek megoldása

$$u(x) = x - \frac{e^{x/\varepsilon} - 1}{e^{1/\varepsilon} - 1} \qquad (x \in [0,1]).$$

A megfelelő elsőrendű feladat

$$z' = 1, \qquad z(0) = 0$$

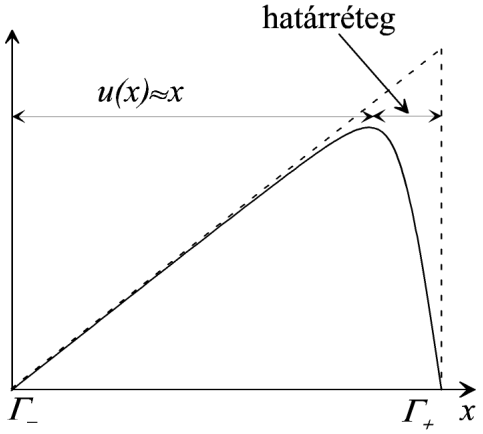
(itt Γ_- a 0 pont), melynek megoldása $z(x) = x$. Látható (3.20. ábra), hogy 0-ból indulva sokáig $u(x) \approx z(x)$, majd az 1 pont körül az $u(1) = 0$ peremfeltétel az u megoldást hirtelen 0-ba vonzza. A feladat 2D analógiája [10, 3.1.1. példa].

Az $\varepsilon \approx 0$ tulajdonság további hátránya, hogy most ε a bilineáris forma alsó határa: (3.46) alapján, most $p \equiv m := \varepsilon$ mellett

$$a(u,u) = \int_\Omega \Big(\varepsilon\,|\nabla u|^2 + (\mathbf{w}\cdot\nabla u)u\Big) \geq \varepsilon \int_\Omega |\nabla u|^2 \qquad (\forall u \in H_0^1(\Omega)), \tag{3.56}$$

így a koercivitás majdnem elvész. Emiatt pl. a (3.23) stabilitási becslés most a gyakorlatilag használhatatlan

$$\|u_h\| \leq \frac{1}{\varepsilon}\|\ell\| \tag{3.57}$$



3.20. ábra. Konvekció-dominált 1D feladat megoldása.

alakú. (Hasonlóan, a (3.50) becslésben az M/ε konstans lesz nagyon nagy.)

A fenti problémák kezelése nem nyilvánvaló. A határrétegen a megoldás közelítésének egy módja a rács lokális finomítása. A (3.56) becslés problémája miatt azonban inkább magát az alkalmazott végeselem-módszert szokás módosítani. Röviden összefoglaljuk az ún. *áramvonal-menti diffúziós végeselem-módszert* (SDFEM, az angol „streamline diffusion FEM" névből), amely stabilizálja a koercivitási határt a bilineáris forma módosításával, alkalmas új tag hozzávételével.

A módosított bilineáris formához a 3.6. megjegyzésben említett Petrov–Galjorkin-módszeren keresztül és a forma elemenkénti tagokra bontásával lehet eljutni. Tegyük fel, hogy a V_h alteret szakaszonként lineáris Courant-elemekkel értelmeztük, emellett azt is, hogy $\mathbf{w}$ szakaszonként konstans. Míg a standard végeselemes megoldás alakja $u_h \in V_h$, ahol

$$\int_\Omega \Big(\varepsilon\,\nabla u_h\cdot\nabla v_h + (\mathbf{w}\cdot\nabla u_h)v_h\Big) = \int_\Omega f v_h \qquad (\forall v\in V_h), \tag{3.58}$$

ezt most helyettesítsük a

$$\sum_{T_k\in\mathcal{T}_h}\int_{T_k}\Big(\varepsilon\,\nabla u_h\cdot\nabla w_h + (\mathbf{w}\cdot\nabla u_h)w_h\Big) = \int_\Omega f w_h \qquad (\forall v_h\in W_h) \tag{3.59}$$

alakkal (ahol $\mathcal{T}_h$ a V_h altérhez tartozó trianguláció), majd a tesztfüggvényeket válasszuk $w_h := v_h + \delta\mathbf{w}\cdot\nabla v_h$ alakúaknak, ahol $v_h\in V_h$ és $\delta>0$ adott paraméter. Ekkor

$$\sum_{T_k\in\mathcal{T}_h}\int_{T_k}\Big(\varepsilon\,\nabla u_h\cdot\nabla\big(v_h+\delta\mathbf{w}\cdot\nabla v_h\big)+(\mathbf{w}\cdot\nabla u_h)(v_h+\delta\mathbf{w}\cdot\nabla v_h)\Big) = \int_\Omega f(v_h+\delta\mathbf{w}\cdot\nabla v_h) \tag{3.60}$$

$(\forall v_h\in V_h)$. Itt a feltevések alapján $\mathbf{w}$ és ∇v_h szakaszonként konstans, így $\delta\mathbf{w}\cdot\nabla v_h$ is

konstans minden T_k elemen, ezért az integrálokban

$$\int_{T_k} \nabla\big(v_h + \delta \mathbf{w}\cdot\nabla v_h\big) = \int_{T_k} \nabla v_h \qquad (\forall T_k \in \mathcal{T}_h).$$

Emiatt, ill. a középső tag szétbontásával (3.60) végeredményben a következő:

$$\sum_{T_k\in\mathcal{T}_h} \int_{T_k} \Big(\varepsilon\,\nabla u_h\cdot\nabla v_h + (\mathbf{w}\cdot\nabla u_h)v_h + \delta(\mathbf{w}\cdot\nabla u_h)(\mathbf{w}\cdot\nabla v_h)\Big) = \int_\Omega f(v_h + \delta\mathbf{w}\cdot\nabla v_h) \quad (\forall v_h \in V_h).$$

Ebben a bal oldalt már visszaírhatjuk Ω-n vett integrállá, ez az lesz új bilineáris forma:

$$a_{SD}(u_h, v_h) := \int_\Omega \Big(\varepsilon\,\nabla u_h\cdot\nabla v_h + (\mathbf{w}\cdot\nabla u_h)v_h + \delta(\mathbf{w}\cdot\nabla u_h)(\mathbf{w}\cdot\nabla v_h)\Big) \qquad (u_h,\, v_h \in V_h).$$

A jobb oldalt jelölje

$$\ell_{SD} v := \int_\Omega f(v_h + \delta\mathbf{w}\cdot\nabla v_h) \qquad (v_h \in V_h),$$

ekkor a feladat

$$a_{SD}(u_h, v_h) = \ell_{SD} v \qquad (\forall v_h \in V_h).$$

Látható, hogy végeredményben megmaradtunk a V_h altérben, viszont az eredeti $a(u_h, v_h)$ bilineáris forma (azaz (3.58) bal oldala) kiegészült a $\delta(\mathbf{w}\cdot\nabla u_h)(\mathbf{w}\cdot\nabla v_h)$ taggal, ami a $\mathbf{w}$ vektormezőhöz (az áramvonalakhoz) tartozó másodrendű kifejezés gyenge alakja, ezért hívják áramvonal-menti diffúziós tagnak.

Az új tag lényege, hogy a szimmetria miatt a skalárszorzatot is módosíthatjuk vele, és erre nézve $a_{SD}(u, v)$ alsó határa már nem függ ε-tól. Éspedig, legyen

$$\langle u_h, v_h\rangle_{SD} := \int_\Omega \Big(\varepsilon\,\nabla u_h\cdot\nabla v_h + \delta(\mathbf{w}\cdot\nabla u_h)(\mathbf{w}\cdot\nabla v_h)\Big) \qquad (u_h,\, v_h \in V_h).$$

Felhasználva a (3.46) becslést (az $u = u_h \in V_h \subset H_0^1(\Omega)$ esetre), amelyben most $p \equiv m := \varepsilon$:

$$\int_\Omega \Big(\varepsilon\,|\nabla u_h|^2 + (\mathbf{w}\cdot\nabla u_h)u_h\Big) \geq \varepsilon \int_\Omega |\nabla u_h|^2 \qquad (\forall u_h \in V_h),$$

adódik, hogy

$$a_{SD}(u_h, u_h) = \int_\Omega \Big(\varepsilon\,|\nabla u_h|^2 + (\mathbf{w}\cdot\nabla u_h)u_h + \delta(\mathbf{w}\cdot\nabla u_h)^2\Big)$$

$$\geq \int_\Omega \Big(\varepsilon\,|\nabla u_h|^2 + \delta(\mathbf{w}\cdot\nabla u_h)^2\Big) = \|u_h\|_{SD}^2,$$

tehát az új alsó határ 1.

Ennek alapján a stabilitási becslés most (3.57) helyett

$$\|u_h\|_{SD} \leq \|\ell_{SD}\|.$$

Hasonlóan, a (3.50) becslésben sem lesz ε nevező. Összességében az új tag hozzáadásával egy ún. stabilizált végeselem-módszert kaptunk, amely kiküszöböli a konvekció-domináltságból adódó nehézséget.

3.7.2. Negyedrendű egyenletek

Tekintsük az ún. biharmonikus feladatot:

$$\begin{cases} \Delta^2 u \;=\; f, \\ u_{|\partial\Omega} = \partial_\nu u\big|_{\partial\Omega} = 0\,, \end{cases} \tag{3.61}$$

ahol $\Omega \subset \mathbb{R}^2$ korlátos tartomány. Ez egy Ω vékony lemez kis deformációját írja le, ahol f a terhelő erő. A peremfeltétel azt jelenti, hogy a lemezt mereven rögzítettük a szélén.

A feladat gyenge alakja: keresendő $u \in H_0^2(\Omega)$, melyre

$$\int_\Omega D^2u : D^2v \;=\; \int_\Omega fv \qquad (v \in H_0^2(\Omega)) \tag{3.62}$$

(lásd 9.19. feladat), ahol a Hesse-mátrixokra az

$$A : B \;:=\; \sum_{i,k=1}^{2} A_{ik}B_{ik} \qquad (A, B \in \mathbb{R}^{2\times 2}) \tag{3.63}$$

Frobenius-skalárszorzatot használjuk.

Könnyen látható (lásd 9.20. feladat), hogy az

$$a(u,v) := \int_\Omega D^2u : D^2v \tag{3.64}$$

bilineáris forma korlátos és koercív a $H_0^2(\Omega)$ téren a szokásos

$$\langle u, v\rangle_{H_0^2} := \int_\Omega \sum_{|\alpha|=2} (\partial^\alpha u)(\partial^\alpha v) \tag{3.65}$$

skalárszorzatra nézve $M = 2$ és $m = 1$ határokkal.

A végeselem-módszer alkalmazásához tekintsünk egy alkalmas

$$V_h \subset H_0^2(\Omega)$$

véges dimenziós alteret, és ebben keressük a vetületi egyenlet megoldását: $u_h \in V_h$, melyre

$$\int_\Omega D^2 u_h : D^2 v_h \;=\; \int_\Omega f v_h \qquad (v \in V_h). \tag{3.66}$$

Ha $\varphi_1, \dots, \varphi_n$ bázis V_h-ban és a közelítő megoldást a szokott $u_h = \sum_{j=1}^n c_j \varphi_j$ alakban keressük, akkor a megfelelő $A_h c = b_h$ lineáris algebrai egyenletrendszerben

$$a_{ij} = \int_\Omega D^2\varphi_i : D^2\varphi_j \qquad \text{és} \qquad b_i = \int_\Omega f\varphi_i \qquad (i,j = 1,\dots,n).$$

A lineáris algebrai egyenletrendszer egyértelműen megoldható a koercivitás és a 3.2. következmény miatt.

A V_h altér konkrét konstrukciójánál a $V_h \subset H_0^2(\Omega)$ feltétel szakaszonként polinomok esetén a

$$V_h \subset C^1(\overline{\Omega})$$

feltételt jelenti, azaz C^1-elemeket kell használni. A 3.2. szakasz (b) pontjában bevezetett Argyris- vagy Bell-elemek megfelelnek e célra.

A konvergencia alapja a Céa-lemma és interpolációs következménye, melyben most $M/m = 2$:

$$|u^* - u_h|_2 \le 2 \min_{v_h \in V_h} |u^* - v_h|_2 \le 2|u^* - \Pi_h u^*|_2. \tag{3.67}$$

Az $|u^* - \Pi_h u^*|_2$ interpolációs hiba becsléséhez a 3.41. állítást használhatjuk $\ell = 2$ mellett a k-adfokú polinomok esetére, ha u^* elég sima. Itt Argyris-elemek esetén $k \le 5$ lehet: ha $u^* \in H^{k+1}(\Omega)$ és az FEM-ben használt háromszögű trianguláció reguláris, akkor

$$|u^* - \Pi_h u^*|_2 \le c\, h^{k-1}\, |u^*|_{k+1},$$

így

$$|u^* - u_h|_2 \le c\, h^{k-1}\, |u^*|_{k+1} \qquad (2 \le k \le 5),$$

ahol $c > 0$ független a triangulációtól. A minimális eredmény $k = 2$ mellett

$$|u^* - u_h|_2 \le c\, h\, |u^*|_3, \quad \text{ha } u^* \in H^3(\Omega), \tag{3.68}$$

de $u^* \in H^6(\Omega)$ regularitású megoldás esetén $O(h^4)$ is elérhető. Bell-elemek esetén csak $k \le 4$, mert utóbbiak nem minden 5-ödfokú polinomot tartalmaznak; a (3.68) becslés ekkor is igaz, az elérhető legnagyobb rend $O(h^3)$.

4. fejezet

A diszkretizált elliptikus feladatok iterációs megoldása

E szakasz témája a véges differenciák módszere vagy a végeselem-módszer nyomán keletkező

$$A_h c = f_h$$

lineáris algebrai egyenletrendszerek néhány alapvető megoldási módszere. Ezekben az A_h mátrixok alapvető jellemzője, hogy sávos szerkezetűek, ezért érdemes iterációs módszert alkalmazni, hiszen az itt fellépő mátrix-vektor-szorzások művelet- és tárolásigénye a sávos szerkezet miatt nem nagy.

Itt röviden kimondunk néhány fő eredményt. Iterációs módszerekről részletesebben pl. a [3, 16, 27] könyvekben olvashatunk.

A továbbiakban tekintsünk egy

$$Ax = b \tag{4.1}$$

lineáris algebrai egyenletrendszert. Csak ott jelezzük, hogy $A_h c = f_h$ alakú diszkretizált elliptikus feladatról van szó, ahol ennek van jelentősége.

Ebben a szakaszban jelölje $\langle ., . \rangle$ az euklideszi skalárszorzatot. (Skalárszorzatot az első két pontban használunk. A jelöléssel tükrözzük, hogy az említett módszerek nem használják ki, hogy véges dimenziós feladatról van szó [16].)

4.1. Egyszerű iterációk

Az egyszerű iterációk alapgondolata az, hogy a (4.1) lineáris algebrai egyenletrendszer ekvivalens az $x = x - \alpha(Ax - b)$ rendszerrel, ha $\alpha \neq 0$ paraméter; ez pedig egy fixpont-típusú feladat, melyben az $\alpha > 0$ választás és az A-ra tett egyes feltételek mellett kontrakció szerepel.

4.1. Tétel. *Legyen A szimmetrikus, pozitív definit mátrix, melynek sajátértékei az $M \geq m > 0$ számok közé esnek, azaz*

$$m|x|^2 \leq \langle Ax, x\rangle \leq M|x|^2 \qquad (\forall x \in \mathbb{R}^n). \tag{4.2}$$

Ha $0 < \alpha < \frac{2}{M}$, akkor tetszőleges $x_0 \in H$ esetén az

$$x_{n+1} := x_n - \alpha\,(Ax_n - b) \qquad (n \in \mathbb{N})$$

iteráció lineárisan konvergál. Az optimális eset:

$$x_{n+1} := x_n - \frac{2}{m+M}\,(Ax_n - b) \qquad (n \in \mathbb{N}),$$

melyre

$$|x_n - x^*| \leq \frac{1}{m}\,|Ax_0 - f|\left(\frac{M-m}{M+m}\right)^n.$$

A bizonyítást l. pl.: [16, 16.1. tétel]. Ez az egyszerű iteráció az A mátrix által meghatározott $\Phi(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ kvadratikus funkcionálhoz tartozó gradiens-módszer.

A fenti iteráció akkor is jó, ha a mátrix nem szimmetrikus, de ekkor a konvergencia lassabb lesz.

4.2. Tétel. *Legyen A pozitív definit mátrix, melyre léteznek olyan $M \geq m > 0$ számok, hogy (4.2) teljesül. Ha $0 < \alpha < \frac{2m}{M^2}$, akkor tetszőleges $x_0 \in H$ esetén az*

$$x_{n+1} := x_n - \alpha\,(Ax_n - b) \qquad (n \in \mathbb{N})$$

iteráció lineárisan konvergál. Az optimális eset:

$$x_{n+1} := x_n - \frac{m}{M^2}\,(Ax_n - b) \qquad (n \in \mathbb{N})$$

melyre

$$|x_n - x^*| \leq \frac{1}{m}\,|Ax_0 - f|\Big(1 - \frac{m^2}{M^2}\Big)^{n/2}.$$

A bizonyítás pl. a [16] könyv 18.2. tételéből következik.

4.2. A konjugált gradiens-módszer

Legyen A szimmetrikus, pozitív definit mátrix. Az egyszerű iteráció (gradiens-módszer) általános lépése a következő volt:

$$x_{n+1} := x_n - \alpha_n r_n,$$

ahol $r_n = Ax_n - b$ a reziduális hibavektor. Rekurzióval látható, hogy ekkor x_{n+1}-et az x_0 és a $r_0, r_1, \ldots, r_n$ vektorok feszítik ki, utóbbiak negatív együtthatókkal. Szemléletesen: annál nagyobb halmazon (kúpon) kereshetjük az újabb közelítést, minél „függetlenebbek" az r_i vektorok, pontosabban, minél nagyobb szöget zárnak be páronként. Ebből adódik az ún. konjugált gradiens-módszer (KGM) alapgondolata: az r_n helyett a p_n, úgynevezett konjugált irányokban keresünk, ahol a p_n vektorok merőlegesek az A-skalárszorzatban:

$$\langle Ap_i, p_j \rangle = 0 \qquad (\forall i \neq j). \tag{KONJ}$$

Ezután a sorozat a gradiens-módszerhez hasonló: legyen $x_0 \in H$ tetszőleges, és ha x_n megvan, akkor

$$x_{n+1} = x_n - \alpha_n p_n,$$

ahol $\alpha_n > 0$ állandó. Utóbbit optimálisan akarjuk megválasztani abban az értelemben, hogy a hibavektor legyen merőleges az előző irányokra:

$$\langle r_{n+1}, p_i \rangle = 0 \qquad (i = 1, 2, \ldots, n), \tag{ORT}$$

ez analóg a Galjorkin-ortogonalitással. A p_n irányokat az ún. Krylov-alterekkel konstruáljuk. A részletekért lásd a 16.2. szakaszt a [16] könyvben, itt csak megadjuk a KGM konstrukcióját és a fő konvergenciatételt.

A konjugált gradiens-módszer (KGM) algoritmusa:

- Legyen $x_0 \in H$ tetszőleges, $p_0 := r_0 (= Ax_0 - f)$;
- ha $n \in \mathbb{N}$ és x_n, p_n ismert, akkor

$$x_{n+1} := x_n - \alpha_n p_n, \text{ ahol } \alpha_n = \frac{\langle r_n, p_n \rangle}{\langle Ap_n, p_n \rangle},$$

$$p_{n+1} := r_{n+1} - \beta_n p_n, \text{ ahol } \beta_n = \frac{\langle Ar_{n+1}, p_n \rangle}{\langle Ap_n, p_n \rangle}.$$

Megj.: Az α_n és β_n értékeket gyakran másik alakban használják, ezzel az algoritmus az

$$x_{n+1} := x_n + \alpha_n p_n \quad \text{és} \quad r_{n+1} := r_n + \alpha_n A p_n, \qquad \text{ahol } \alpha_n = -\frac{|r_n|^2}{\langle Ap_n, p_n \rangle},$$

$$p_{n+1} := r_{n+1} + \beta_n p_n, \qquad \text{ahol } \beta_n = \frac{|r_{n+1}|^2}{|r_n|^2}$$

alakban írható.

A konvergenciára az $|x|_A^2 = \langle Ax, x \rangle$ energianormában kapjuk a fő becslést:

4.3. Tétel. *Ha A szimmetrikus, pozitív definit mátrix, melynek sajátértékei az $M \geq m > 0$ számok közé esnek (azaz (4.2) teljesül), akkor a KGM által létrehozott $e_n := x_n - x^*$ hibavektorokra*

$$\frac{|e_n|_A}{|e_0|_A} \leq 2 \left(\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} - \sqrt{m}} \right)^n \qquad (\forall n \in \mathbb{N}).$$

(Lásd pl. [16, 16.10. tétel].)

A KGM tehát gyorsabb, mint az egyszerű iteráció, és ez lényeges pl. akkor, ha az M/m kondíciószám nagy. Ha A sajátértékei 1 körül torlódnak, akkor igazolható a szuperlineáris konvergencia is (vö. [16, 16.12. tétel]).

4.4. Megjegyzés. Nem szimmetrikus lineáris algebrai egyenletrendszerre a fenti KGM többféleképp általánosítható. Egy kézenfekvő lehetőség az eredeti nem szimmetrikus $Ax = b$ rendszer helyett a szimmetrizált $A^T Ax = A^T b$ rendszert megoldani a fenti algoritmussal, lásd pl. [16, 16. fejezet]. Más típusú algoritmusok nyerhetők a reziduális ortogonalitásának (ORT) vagy minimalitásának megkövetelésével, lásd pl. [3]. ♢

4.3. Prekondicionálás

Tekintsünk egy

$$x_{n+1} := x_n - \alpha (Ax_n - b) \qquad (n \in \mathbb{N})$$

egyszerű iterációt. Mint láttuk, ennek konvergenciasebessége végső soron az M/m számtól függ, sőt ez igaz a KGM esetén is. Ha A szimmetrikus, akkor M/m nem más, mint A kondíciószáma, jelölése:

$$\kappa(A) := \frac{M}{m}.$$

Ha ez nagy, akkor a lineáris algebrai egyenletrendszert rosszul kondicionáltnak nevezzük, ekkor a fenti iterációk lassan konvergálnak. Tipikusan ilyen egy diszkretizált elliptikus feladat, ekkor

$$\kappa(A_h) := O(h^{-2})$$

nagyságrendű (lásd 9.22.–9.23. feladatok). Fontos kérdés, hogyan javítható ilyenkor az iteráció konvergenciasebessége.

A prekondicionálás alapgondolata, hogy transzformáljuk a lineáris algebrai egyenletrendszert olyan rendszerré, melyben a mátrix kondíciószáma lényegesen kisebb, mint A kondíciószáma. Legyen B alkalmasan választott invertálható mátrix. Az eredeti

$$Ax = b$$

lineáris algebrai egyenletrendszert formailag átírjuk a vele ekvivalens

$$B^{-1}Ax = B^{-1}b$$

rendszerré. Célunk a B mátrixot úgy választani, hogy

$$\kappa(B^{-1}A) \ll \kappa(A)$$

legyen. Erre az egyszerű iteráció

$$x_{n+1} := x_n - B^{-1}(Ax_n - b) \qquad (n \in \mathbb{N})$$

alakú lesz, ahol B-be beépítettük az iterációs paramétert. Látható, hogy az iteráció végrehajtása ekkor B-re vonatkozó segédfeladatok megoldását igényli, azaz a fenti iteráció n. lépése átírható a

$$\begin{aligned} Bz_n = r_n &:= Ax_n - b, \\ x_{n+1} &:= x_n - z_n \end{aligned}$$

alakba. Ezt azt is jelenti, hogy a segédfeladatok megoldásának jóval egyszerűbbnek kell lennie az eredetiénél. A két szempont tehát ellentmondó: B jól közelítse A-t, de a rá vonatkozó lineáris algebrai egyenletrendszereket jóval egyszerűbben lehessen megoldani.

A fenti szempontokat gyakran azzal próbáljuk teljesíteni, hogy B az A egy adott (lehetőleg egyszerűbb) részét foglalja magába. Ezt úgy írjuk le, hogy A-t felbontjuk két részre, és az első lesz a prekondicionáló mátrix:

$$A = B - R.$$

Ezzel az

$$x_{n+1} := x_n - B^{-1}(Ax_n - b) \qquad (n \in \mathbb{N})$$

egyszerű iteráció a következőképp írható át:

$$Bx_{n+1} = Bx_n - (Ax_n - b) \quad \Leftrightarrow \quad Bx_{n+1} = Rx_n + b \qquad (n \in \mathbb{N}).$$

Ez úgy fogható fel, hogy az

$$Ax = b \quad \Leftrightarrow \quad Bx = Rx + b$$

rendszerre az iterációban a bal oldalon az új, a jobb oldalon a régi iteráltat írjuk, azaz csak az egyszerűbb részt frissítjük.

Idézzük fel a két leggyakoribb ilyen prekondicionált iterációt!

Jacobi-iteráció. Ekkor $B := D$ az A mátrix főátlója:

$$x_{n+1} := x_n - D^{-1}(Ax_n - b) \qquad (n \in \mathbb{N}),$$

azaz lépésenként

$$Dz_n = r_n,$$
$$x_{n+1} := x_n - z_n.$$

A módszer definíciójához elég azt feltenni, hogy $d_i := a_{ii} > 0$ $(i = 1, ..., n)$. A Jacobi-módszerre a segédfeladatok megoldása a lehető legegyszerűbb: $(z_n)_i = (r_n)_i/d_i$ $(i = 1, ..., n)$.

Gauss–Seidel-iteráció. Ekkor $B := L + D$ az A mátrix alsóháromszög-része a főátlót is beleértve. Az iterációban lépésenként

$$(L + D)z_n = r_n$$
$$x_{n+1} := x_n - z_n.$$

A segédfeladatok rekurzív visszahelyettesítéssel megoldhatók.

4.5. Tétel. *Ha A diagonálisan domináns vagy M-mátrix, akkor a Jacobi- és Gauss–Seidel-iteráció is konvergens.*

A bizonyítást lásd pl. a [27] könyvben.

4.6. Megjegyzés. Nehány további alapvető prekondicionálási módszer:

(i) *Csillapított Jacobi- vagy Gauss–Seidel-iteráció.* A fenti két iteráció gyorsasága egyes esetekben javítható a lépéshosszt módosító paraméter bevezetésével:

$$x_{n+1} := x_n - \omega D^{-1}(Ax_n - b), \qquad \text{ill.} \qquad x_{n+1} := x_n - \omega(L + D)^{-1}(Ax_n - b).$$

(ii) *Inkomplett LU- vagy inkomplett Cholesky-felbontás.* Prekondicionáló mátrix gyanánt az A mátrix jó közelítését kaphatjuk, ha az A mátrix LU-felbontását vagy szimmetrikus esetben Cholesky-felbontását csak közelítőleg hajtjuk végre úgy, hogy a felbontásba 0 elem kerül ott, ahol az A-ban 0 volt, vagy általánosabban ott is, ahol a felbontásban adott küszöbnél kisebb érték állna. Ezzel pontosan vagy közelítőleg megtartjuk az eredeti mátrix ritkasági mintázatát, azaz megakadályozzuk vagy csökkentjük a feltöltődést, és a felbontás munkaigénye jóval kisebb, mint egy pontos LU- vagy Cholesky-felbontásé.

(iii) *Prekondicionálás segédoperátorral.* Ha A egy általános függvényegyütthatós (akár nem szimmetrikus) elliptikus operátor végeselemes megoldásának merevségi mátrixa, akkor prekondicionáló mátrixként célszerű egy állandó együtthatós szimmetrikus operátor, lényegében a Laplace-operátor merevségi mátrixát alkalmazni. Ekkor a prekondicionált mátrix kondíciószáma a rácsmérettől független korláttal becsülhető, azaz rácsfüggetlen, lásd [16, 19.4.1. szakasz]. ♢

5. fejezet

A többrácsos (multigrid-)módszer

Elliptikus feladatok egyik leghatékonyabban bevált megoldási módszere a többrácsos vagy multigrid-módszer, rövidítve MG. Ez optimálisan ötvözi az iterációk tulajdonságait a rácsméret variálásának lehetőségeivel abban az értelemben, hogy a szükséges műveletigény rendje arányos a változók számával, vagyis a lehető legkisebb. A multigrid-módszer mind az FDM-rel, mind az FEM-rel felírható. Először ismertetjük a módszer alapgondolatát két rácsra egy egyszerű helyzetben, részletesebben foglalkozunk a kétrácsos módszer konvergenciájával az FEM esetén, majd kitérünk a többrácsos esetre és annak műveletigényere.

5.1. A kétrácsos módszer alapelve és konstrukciója

A defekt-korrekciós elv két ráccsal. Tekintsünk először egy általános

$$Au = b$$

feladatot, ahol A lehet mátrix vagy általánosabban lineáris operátor is, és tegyük fel, hogy ismerjük ennek egy w kiindulási közelítő megoldását. Hogyan javítsuk w-t, hogy közelebb kerüljünk az igazi megoldáshoz? A defekt-korrekciós elv azt jelenti, hogy a pontos megoldás eléréséhez szükséges korrekciót egy olyan egyenletből kapjuk, ahol a jobboldal a w-hez tartozó reziduális hiba negáltja (w „defektje"). Éspedig, ha meg akarjuk határozni azt a p-t, melyre

$$u = w + p$$

a pontos megoldás, akkor meg kellene oldanunk az

$$Ap = -r, \qquad \text{ahol} \quad r := Aw - b$$

egyenletet, hiszen ekkor

$$Au = Aw + Ap = Aw - Aw + b = b.$$

A közelítő megoldásokra nézve a defekt-korrekciós elv azt jelenti, hogy a segédfeladatban A helyett annak alkalmas közelítését használjuk. Ez volt a prekondicionált egyszerű iterációk lényege is, ahol $B \approx A$ révén a fenti egyszeri pontos $w \mapsto u = w - A^{-1}r = w - A^{-1}(Aw - b)$ lépés közelítését a $w \mapsto w - B^{-1}(Aw - b)$ lépés iterálásával adtuk meg.

A kétrácsos módszer esetén egy

$$A_h u = b_h$$

alakú feladatot vizsgálunk, ahol A_h egy elliptikus feladat diszkretizációs mátrixa egy adott h finomságú (FDM-es vagy FEM-es) rácson. Legyen w_h kiindulási közelítő megoldás. Ekkor az $A_h p = -r_h$ korrekciós lépésben közelítésként célszerűnek tűnik az A_h mátrixot egy durvább, H finomságú rácson vett A_H mátrixszal helyettesíteni, s megoldani az

$$\text{„ } A_H p_H = r_h\text{”} \quad \text{egyenletet,} \quad \text{majd} \quad \text{„ } u_h = w_h + p_H\text{”}$$

alakban továbblépni. Itt azonban értelmezési probléma, hogy r_h, u_h és w_h a finom rácson, p_H és $A_H p_H$ a durva rácson értelmezett vektor, így ebben a formában nem állhat fent egyik egyenlőség sem. Be kell tehát vezetni alkalmas áttérési leképezéseket a V_h finom és V_H durva rácsú altér között:

$$R : V_h \to V_H \;\; \text{restrikció (megszorítás)}, \qquad P : V_H \to V_h \;\; \text{prolongáció (kiterjesztés)},$$

a fentieket pedig $A_H p_H = R r_h$ és $u_h = w_h + P p_H$ alakkal helyettesíteni. Ez már értelmes, és iterált formában az

$$u_{i+1} = u_i - P A_H^{-1} R(A u_i - b_h)$$

iterációhoz vezet.

Ez azonban meg mindig nem elég jó. Ha az $r_h := A u_i - b_h \in V_h$ vektornak nincs V_H-beli („durva”) komponense, akkor $R r_h = R(A u_i - b_h) = 0$, igy $u_{i+1} = u_i$, azaz az iterációs lépés nem javít, az iteráció leáll u_i-ben. A durva rácson vett segédfeladat ötletét tehát nem elég ebben a formában iterálni, hanem egy másik alapgondolattal kell ötvözni, amit a következő pontban nézünk meg egy egyszerű példán.

Egyszerű iterációk simító hatása. A vizsgálandó tulajdonsághoz az előző szakaszban látott jelenségből kiindulva jutunk el. A diszkretizált elliptikus feladatoknál láttuk, hogy a kapott mátrixok kondíciószáma nagy és $O(h^{-2})$ rendben romlik a rács finomításával. Szeretnénk először megtalálni, mi okozza ezt a hátrányos jelenséget. Tekintsük egyszerű példaként a

$$-u'' = f, \qquad u(0) = u(1) = 0$$

egydimenziós feladat FDM-es megoldását.

Ennek mátrixa a tridiagonális

$$A_h = \frac{1}{h^2} \;\; \texttt{tridiag}[-1, 2, -1]$$

mátrix. Alkalmazzuk a csillapított Jacobi-módszert! Ekkor

$$D = \frac{2}{h^2}\, I,$$

ahol I az identitásmátrix. Jelöljük most x^k-val $(k = 1, 2, ...)$ az iteráció tagjait. Ekkor

$$x^{k+1} := x^k - \omega \frac{h^2}{2}(A_h x^k - b) \qquad (k \in \mathbb{N}).$$

Az $e^k := x^k - x^*$ hibára

$$e^{k+1} = \left(I - \omega \frac{h^2}{2} A_h\right) e^k \qquad (k \in \mathbb{N}).$$

Így a hiba euklideszi normája lépésenként az

$$\left\| I - \omega \frac{h^2}{2} A_h \right\| = \max_{j=1,...,n} \left| 1 - \omega \frac{h^2}{2} \lambda_j(A_h) \right| = \max_{j=1,...,n} \left| 1 - 2\omega \, \sin^2\left(\frac{j\pi h}{2}\right) \right|$$

euklideszi norma, ahol

$$\lambda_j(A_h) = \frac{4}{h^2} \sin^2\left(\frac{j\pi h}{2}\right) \qquad (j = 1, ..., n)$$

az A_h sajátértékei (lásd (2.11)).

Látható, hogy a

$$\mu_j := 1 - 2\omega \, \sin^2\left(\frac{j\pi h}{2}\right) \qquad (j = 1, ..., n)$$

értékek az

$$f_\omega(x) := 1 - 2\omega \, \sin^2\left(\frac{\pi x}{2}\right) \qquad (x \in [0, 1]) \tag{5.1}$$

függvénynek az $x = jh$ pontokban vett értékei. Ha j kicsi $(j \approx 0)$, akkor

$$f_\omega(x) \approx f_\omega(0) = 1,$$

ha viszont j nagy $(j \approx n)$, akkor

$$f_\omega(x) \approx f_\omega(1) = 1 - 2\omega.$$

Így ω választásával a kis j indexű értékeket nem tudjuk érdemben befolyásolni, a nagy j indexű értékeket viszont igen.

Tekintsük ezért az $\frac{n}{2} \le j \le n$ indexek esetét. Ekkor könnyen látható (lásd 9.32. feladat):

$$\max_{\frac{n}{2}\le j\le n} |\mu_j| \le \max_{x\in[\frac{1}{2},1]} |f_\omega(x)| = \max\{|1-\omega|,\ |2\omega-1|\},$$

és a jobboldal akkor a legkisebb, ha

$$\omega = \frac{2}{3},$$

ekkor

$$|\mu_j| \le \frac{1}{3} \qquad (\frac{n}{2} \le j \le n)$$

n-től függetlenül.

Végül írjuk fel a hibavektort a normált sajátvektorok szerint kifejtve, és bontsuk fel kis indexű ($1 \le j < \frac{n}{2}$) és nagy indexű ($\frac{n}{2} \le j \le n$) komponensekre:

$$e^k = \sum_{j<\frac{n}{2}} c_j v_j \ + \sum_{j\ge\frac{n}{2}} c_j v_j =: e^k_- + e^k_+ \ .$$

Ekkor

$$e^{k+1} = \Big(I - \omega\frac{h^2}{2}A\Big)\, e^k = \sum_{j<\frac{n}{2}} \mu_j c_j v_j \ + \sum_{j\ge\frac{n}{2}} \mu_j c_j v_j =: e^{k+1}_- + e^{k+1}_+ \ .$$

A fentiek alapján a második tagra, mely a nagy indexű komponensekből jön,

$$|e^{k+1}_+|^2 = \Big|\sum_{j\ge\frac{n}{2}} \mu_j c_j v_j\Big|^2 = \sum_{j\ge\frac{n}{2}} \mu_j^2 c_j^2 \le \frac{1}{9}\sum_{j\ge\frac{n}{2}} c_j^2 = \frac{1}{9}|e^k_+|^2,$$

azaz az iteráció a nagy komponensű tag hosszát legfeljebb $\frac{1}{3}$-ára csökkenti (az n rácspontszámtól függetlenül), míg az egész vektor hossza alig csökken, hisz kis j-re $\mu_j \approx 1$.

Így kimondható, hogy a lassú konvergenciát a kis indexű komponensek okozzák, míg a nagy indexű resz rácsfüggetlen mértékben, viszonylag gyorsan csökken.

Ez azt is jelenti, hogy elég sok iterációs lépés után a hibának szinte csak a kis indexű komponensei maradnak meg, a nagy indexűek eltörpülnek. Mivel láttuk, hogy a sajátvektorok az első n pontos sajátfüggvény rácspontbeli értékeiből származnak, így a kis indexű ($j \ge \frac{n}{2}$) komponensek az első $n/2$ pontos sajátfüggvényhez tartoznak. Ezek kevésbé oszcillálnak, ezért szokás „simább” sajátfüggvénynek nevezni őket, és hasonlóan a hiba kis indexű komponenseit „simább” komponenseknek. Másrészt az említett sajátfüggvények a kétszeres lépésközű durvább rácshoz tartozó sajátfüggvények, így a hiba érdemben megmaradt komponense a durva rácsból származó.

A fenti szóhasználattal azt mondhatjuk, hogy bár az iteráció a hibavektor hosszát csak kicsit csökkenti, a hibavektor komponenseit egyre inkább a kis indexű ($j \ge \frac{n}{2}$) altérbe szorítja vissza, azaz a hibavektor „simább” lesz. Az iteráció tehát „simítja” a hibavektort.

A kétrácsos módszer algoritmusa. A fenti két alapgondolatot a következőképp ötvözzük. Induljunk ki egy adott rácsból, ezt nevezzük finom rácsnak. A durva rács FDM esetén álljon a finom rács csomópontjainak alkalmas részhalmazából (a koordinátairányokban minden második csomópontot vesszük bele), FEM esetén a durva altér legyen altere a finom altérnek (egyenletes rács esetén kézenfekvő itt is kétszer akkora rácsparamétert tekinteni). Alkalmazzunk először egyszerű iterációt a finom rácsú feladatra adott számú lépésben. (Ekkor a hiba simább lesz, azaz érdemben megmaradt komponense a durva rácsból származó.) A kapott közelítő megoldásra alkalmazzuk a kétrácsos defektkorrekciót, azaz a reziduális restrikciójával kapott jobboldalra megoldjuk a feladatot a durva rácson, ezt prolongáljuk a finom rácsra és hozzáadjuk az előző közelítéshez. (Mindez FDM és FEM esetén is értelmes.) Ebből felírható a kétrácsos módszer algoritmusa:

Legyen u_h^0 tetszőleges kiinduló közelítés a finom rácson. Ha $i = 0, 1, 2, \ldots$:

- Tegyük fel, hogy megkonstruáltuk az u_h^i közelítést.
- *Simítás* (k darab belső iteráció): legyen B alkalmas prekondicionáló mátrix (pl. Jacobi vagy Gauss–Seidel), $v_0 := u_h^i$, majd

$$v_{j+1} := v_j - B^{-1}(A_h v_j - b) \qquad (j = 0, ..., k-1).$$

  Legyen

$$v := v_k$$

  és a reziduális

$$r_h := A_h v - b_h.$$

- A reziduális *restrikciója:* $R : r_h \mapsto r_H$.
- Megoldás *durva rácson:*

$$A_H e_H = -r_H.$$

- *Prolongáció:* $P : e_H \mapsto e_h$.
- *Korrekció* (az új közelítés):

$$u_h^{i+1} := v + e_h.$$

A módszer további vizsgálatához célszerű először felírni, hogyan kapjuk u_h^i-bol u_h^{i+1}-et. A fenti algoritmus v utáni lépéseit összegezve, és u_h^*-gal jelölve a finom rácson való megoldást,

$$u_h^{i+1} = v - PA_H^{-1}RA_h(v - u_h^*).$$

Hasonlóan, az

$$e_h^i := u_h^i - u_h^*$$

hibák vizsgálatához felírjuk, hogyan kapjuk e_h^i-bol e_h^{i+1}-et. Itt azt is felhasználjuk, hogy a simítás révén

$$v - u_h^* := v_k - u_h^* = (I - B^{-1}A_h)^k(v_0 - u_h^*) = (I - B^{-1}A_h)^k(u_h^i - u_h^*) = (I - B^{-1}A_h)^k e_h^i.$$

Így

$$e_h^{i+1} := u_h^{i+1} - u_h^* = v - u_h^* - PA_H^{-1}RA_h(v - u_h^*)$$
$$= (I - PA_H^{-1}RA_h)(v - u_h^*) = (I - PA_H^{-1}RA_h)(I - B^{-1}A_h)^k e_h^i.$$

Az első tényezőből kiemelünk A_h-t, ezzel

$$e_h^{i+1} = (A_h^{-1} - PA_H^{-1}R)A_h(I - B^{-1}A_h)^k e_h^i. \tag{5.2}$$

A régiből az új hibát tehát a

$$K := (A_h^{-1} - PA_H^{-1}R)A_h(I - B^{-1}A_h)^k$$

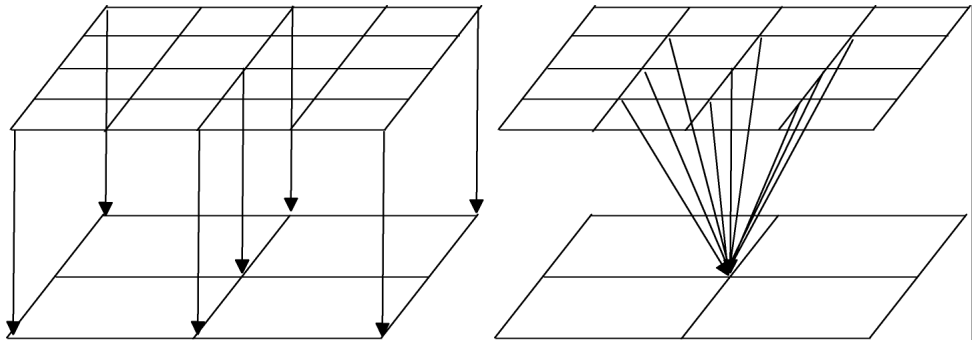
mátrix hozza létre. A konvergenciához azt kell majd igazolni, hogy ennek normája valamely 1-nél kisebb σ konstanssal becsülhető.

5.1. Megjegyzés. A kapott K mátrix nem szimmetrikus. Ha módosítjuk az algoritmust úgy, hogy a végén újabb k darab simítást, ún. utósimítást végzünk, és feltesszük, hogy $R = P^T$, akkor könnyen látható, hogy a kapott

$$\tilde{K} := (I - B^{-1}A_h)^k(A_h^{-1} - PA_H^{-1}R)A_h(I - B^{-1}A_h)^k$$

mátrix mar szimmetrikus az A_h-skalárszorzatra nézve. ♢

5.2. Megjegyzés. A módszer konstrukciójának fontos része a restrikció és prolongáció felírása. A legegyszerűbb prolongáció az FDM esetén a lineáris kiterjesztés, míg a legegyszerűbb restrikció a megszorítás. Ezek mátrixa (lásd 9.24.–9.25. feladat) azonban nem egymás transzponáltja. Ezért a restrikciót úgy szokás felírni, hogy mátrixa a lineáris kiterjesztés prolongációs mátrixának transzponáltja legyen. Ez olyan súlyozott megszorítás, amely a távolabbi pontokat is figyelembe veszi (lásd 9.26. feladat). FEM esetén természetes prolongációt ad, hogy a durva altér elemei egyben a finom altér elemei is, így a beágyazás meghatározza az együtthatókra adódó mátrixot (lásd 9.28.–9.29. feladat). A restrikciót itt is úgy választjuk, hogy mátrixa a prolongáció mátrixának transzponáltja legyen. ♢



5.1. ábra. Restrikció téglalaprácson.

5.2. A kétrácsos módszer konvergenciája

A konvergencia alapja. A konvergencia vizsgálatánál az (5.2) egyenlőségből és abból indulunk ki, hogy az 5.2. megjegyzés alapján a restrikciót a prolongációs mátrix transzponáltja alapján választjuk. Ekkor az új hiba a régiből az

$$e_h^{i+1} = Ke_h^i := (A_h^{-1} - PA_H^{-1}P^T)A_h(I - B^{-1}A_h)^k e_h^i \tag{5.3}$$

összefüggéssel adódik. Célunk, hogy K az $|x|_{A_h}^2 = \langle Ax, x\rangle$ energianormában (azaz diszkrét súlyozott Szoboljev-normában) kontraktív legyen, azaz létezzen olyan $\sigma < 1$ konstans, hogy $|e_h^{i+1}|_{A_h} \le \sigma\,|e_h^i|_{A_h}$ legyen.

Részletesen végigmegyünk 2D FEM mellett a szakaszonként lineáris/bilineáris eseten, majd a végén kitérünk az egyéb esetekre is.

A kiindulás, hogy a kívánt becslést két részre daraboljuk:

5.3. Állítás. *Ha teljesülnek az alábbiak:*

(i) az approximációs tulajdonság: *van olyan $C > 0$ konstans, hogy*

$$|(A_h^{-1} - PA_H^{-1}P^T)b|_{A_h} \le C|b| \qquad (\forall b \in \mathbb{R}^n),$$

(ii) a simító tulajdonság: van olyan $\varepsilon_k \to 0$ h-tól független sorozat, hogy

$$|A_h(I - B^{-1}A_h)^k x| \le \varepsilon_k |x|_{A_h} \qquad (\forall x \in \mathbb{R}^n,\ k \in \mathbb{N}),$$

akkor van olyan $k \in \mathbb{N}^+$, hogy a k darab simítást tartalmazó kétrácsos algoritmus konvergens, azaz van olyan $\sigma < 1$ konstans, hogy

$$|e_h^{i+1}|_{A_h} \le \sigma\,|e_h^i|_{A_h} \qquad (\forall i \in \mathbb{N}). \tag{5.4}$$

Bizonyítás. A két tulajdonságot kombinálva

$$|e_h^{i+1}|_{A_h} = |(A_h^{-1} - PA_H^{-1}P^T)A_h(I - B^{-1}A_h)^k e_h^i|_{A_h}$$

$$\leq C|A_h(I - B^{-1}A_h)^k e_h^i| \leq C\varepsilon_k |e_h^i|_{A_h}.$$

Válasszuk k-t akkorára, hogy

$$\sigma := C\varepsilon_k < 1$$

legyen, ekkor a megfelelő algoritmusra (5.4) teljesül. □

Az approximációs tulajdonság. Tekintsünk egy elliptikus feladatot, végeselemes diszkretizációját és a kétrácsos konstrukciót az alábbi tulajdonságokkal:

5.4. Feltevés.

(i) A feladat H^2-reguláris, azaz bármely $f \in L^2(\Omega)$ jobboldal esetén a megoldásra $u^* \in H^2(\Omega)$ és létezik $c_0 > 0$, hogy $|u^*|_2 \leq c_0\|f\|_0$. (Ez fennáll pl. az 1.5. tétel feltételeivel.)

(ii) $V_h \subset H_D^1(\Omega)$ adott végeselemes altér $\mathbf{T_1}$- vagy $\mathbf{R_1}$-elemekkel, és legyen a trianguláció reguláris. Jelölje a bázist $\varphi_1, \ldots, \varphi_n$.

(iii) A durva rács H és a finom rács h paraméterére $\frac{H}{h} \leq \varrho$, ahol $\varrho > 0$ nem függ h-tól (tipikusan $\varrho = 2$), és a restrikciót $R = P^T$ alapján választjuk. ♢

Az approximációs tulajdonság bizonyításához szükségünk van V_h-beli függvények normái közti összefüggésre.

5.5. Állítás. *Legyen $f \in V_h$, és $f_h \in \mathbb{R}^n$ az a vektor, melyre $(f_h)_i = \int_\Omega f\varphi_i$ $(i = 1, ..., n)$. Ekkor $|f_h| = O(h)\|f\|_0$ (ahol $\|f\|_0 := \|f\|_{L^2}$).*

A bizonyításhoz lásd a 9.31. feladatot.

5.6. Állítás. *Ha teljesülnek az 5.4. feltételek, akkor igaz az approximációs tulajdonság, azaz van olyan $C > 0$ konstans, hogy*

$$|(A_h^{-1} - PA_H^{-1}P^T)b|_{A_h} \leq C|b| \qquad (\forall b \in \mathbb{R}^n). \tag{5.5}$$

Bizonyítás. Legyen $b \in \mathbb{R}^n$ adott vektor, és legyen c_h az

$$A_h c_h = b \tag{5.6}$$

lineáris algebrai egyenletrendszer megoldása. Legyen $f \in V_h$ az a függvény, melynek b az f együtthatóvektora V_h-ban (azaz $b = f_h$), és legyen $u_h \in V_h$ az $Lu = f$ elliptikus feladat gyenge megoldása. Ekkor $u_h = \sum_{i=1}^{n} c_{hi}\varphi_i$, azaz u_h koordinátavektora V_h-ban épp a fenti c_h vektor. Emellett (3.6) szerint

$$a(u_h, u_h) = A_h c_h \cdot c_h = |c_h|^2_{A_h}. \tag{5.7}$$

Tekintsük most b restrikcióját V_H-ba, ami $R = P^T$, és oldjuk meg ezzel a jobboldallal az

$$A_H c_H = P^T b \tag{5.8}$$

lineáris algebrai egyenletrendszert. Legyen $u_H = \sum_{j=1}^{n_H} c_{Hj}\,\varphi_j^H$, ahol $\varphi_1^H, ..., \varphi_{n_H}^H$ a V_H altér bázisa. Mivel $V_H \subset V_h$, így $u_H \in V_h$. A prolongáció definíciója alapján u_H koordinátavektora V_h-ban a Pc_H vektor, így $u_h - u_H$ koordinátavektora V_h-ban a $c_h - Pc_H$ vektor, amiből (5.7)-hoz hasonlóan

$$a(u_h - u_H, u_h - u_H) = |c_h - Pc_H|^2_{A_h}. \tag{5.9}$$

Másrészt (5.8)-ből

$$Pc_H = PA_H^{-1}P^T b,$$

ebből és (5.6)-ből

$$a(u_h - u_H, u_h - u_H) = |(A_h^{-1} - PA_H^{-1}P^T)b|^2_{A_h},$$

ami épp (5.5) bal oldalának négyzete. A bilineáris forma korlátossága miatt

$$|(A_h^{-1} - PA_H^{-1}P^T)b|_{A_h} = a(u_h - u_H, u_h - u_H)^{1/2} \le M^{1/2}|u_h - u_H|_1,$$

illetve a 3.29. tétel, a feltételekbeli $H \le 2h$ becslés és H^2-regularitás, valamint az 5.5. állítás és a $b = f_h$ egyenlőség alapján van olyan $C > 0$, hogy

$$|u_h - u_H|_1 \le |u_h - u^*|_1 + |u_H - u^*|_1 \le c(h+H)|u^*|_2 \le (1+\varrho)ch|u^*|_2$$
$$\le (1+\varrho)cc_0 h\|f\|_0 = O(h)\|f\|_0 \le C|f_h| = C|b|.$$

Ezekből következik a kívánt tulajdonság. □

A simító tulajdonság. Ennek igazolása a felhasznált iteráció típusától függő egyenkénti vizsgálatot igényel. Itt olyan csillapított Jacobi-iteráció esetén mutatjuk meg, amelyben a motiváló példához hasonlóan A_h főátlója az identitás konstansszorosa, ami kellően egyenletes rács esetén áll fenn. Ez tehát valójában egyszerű iteráció megfelelő paraméterrel. Az alábbi állítás az iterációs mátrixra ad feltételt, ez a feltétel az egyes feladatokra külön számolásokkal igazolható.

5.7. Állítás. *Legyen a prekondicionáló mátrix $B := \theta I$, ahol $\theta > 0$ állandó és I az identitásmátrix. Ha van olyan $0 < Q < 1$ szám, hogy az $I - B^{-1}A_h = I - \frac{1}{\theta}A_h$ iterációs mátrix sajátértékeire $\mu_i \geq -Q$, akkor van olyan $\varepsilon_k \to 0$ h-tól független sorozat, hogy*

$$|A_h(I - B^{-1}A_h)^k x| \leq \varepsilon_k |x|_{A_h} \qquad (\forall x \in \mathbb{R}^n,\ k \in \mathbb{N}).$$

Bizonyítás. Legyenek A_h sajátértékei λ_i, megfelelő normált sajátvektorai v_i. Legyen $x \in \mathbb{R}^n$, és írjuk fel $x = \sum_{i=1}^n c_i v_i$ alakban. Ekkor $|x|^2_{A_h} = A_h x \cdot x = \sum_{i=1}^n \lambda_i c_i^2$, így

$$|A_h(I - B^{-1}A_h)^k x|^2 = \left|A_h\left(I - \frac{1}{\theta}A_h\right)^k x\right|^2 = \sum_{i=1}^n \lambda_i^2 \left(1 - \frac{\lambda_i}{\theta}\right)^{2k} c_i^2$$

$$\leq \max_{i=1,..,n} \lambda_i \left(1 - \frac{\lambda_i}{\theta}\right)^{2k} \sum_{i=1}^n \lambda_i c_i^2 = \max_{i=1,..,n} \lambda_i \left(1 - \frac{\lambda_i}{\theta}\right)^{2k} |x|^2_{A_h}.$$

Itt $I - \frac{1}{\theta}A_h$ sajátértékeire $\mu_i = 1 - \frac{\lambda_i}{\theta}$, így $\lambda_i = \theta(1 - \mu_i)$, és a $-Q \leq \mu_i \leq 1$ feltételt is használva

$$\max_{i=1,..,n} \lambda_i \left(1 - \frac{\lambda_i}{\theta}\right)^{2k} = \theta \max_{i=1,..,n} (1 - \mu_i)\mu_i^{2k} \leq \theta \max_{\mu \in [-Q,1]} (1 - \mu)\mu^{2k} =: \theta \max_{\mu \in [-Q,1]} r(\mu).$$

Elemi deriválás alapján az $r(\mu) = \mu^{2k} - \mu^{2k+1}$ $(\mu \in [-Q, 1])$ nemnegatív függvény maximuma vagy a $-Q$ végpontban, vagy abban a 0 és 1 közti belső pontban van, ahol deriváltja 0, azaz a $\mu_0 = \frac{2k}{2k+1}$ pontban. Itt $r(\mu_0) = (1 - \mu_0)\mu_0^{2k} \leq 1 - \mu_0 = \frac{1}{2k+1}$, ill. a $-Q$ végpontban $r(-Q) = (1 + Q)Q^{2k} \leq 2Q^{2k}$. Így

$$\theta \max_{\mu \in [-Q,1]} r(\mu) \leq \theta \max\{2Q^{2k}, \frac{1}{2k+1}\} =: r_k \to 0,$$

ha $k \to \infty$. Ebből $\varepsilon_k := r_k^{1/2}$ mellett következik a kívánt becslés. □

A fenti állításból az egyes feladatokra konkrét számolásokkal igazolható a simító tulajdonság.

5.8. Példa. 2D FEM egyenletes háromszögrácson.

1. A $[0,1]^2$ tartományon. Felhasználjuk (lásd 3.14. megjegyzés), hogy az FDM-es A_h^{FDM} és FEM-es A_h^{FEM} mátrixokra $A_h^{FEM} = h^2 A_h^{FDM}$. Itt (2.13) szerint

$$\lambda_{ij}(A_h^{FDM}) = \frac{4}{h^2}\left(\sin^2\left(\frac{i\pi h}{2}\right) + \sin^2\left(\frac{j\pi h}{2}\right)\right),$$

   így

$$\lambda_{ij}(A_h^{FEM}) = 4\left(\sin^2\left(\frac{i\pi h}{2}\right) + \sin^2\left(\frac{j\pi h}{2}\right)\right)$$

$(i, j = 1, ..., n)$. Alkalmazzuk a csillapított Jacobi-módszert! Most a főátló

$$D = 4I,$$

azaz az 5.7. állítást $\theta = \frac{4}{\omega}$ mellett használjuk, ahol $\omega > 0$ paraméter. Az $I - \frac{1}{\theta}A_h = I - \frac{\omega}{4}A_h$ iterációs mátrix sajátértékeire

$$\mu_{ij} = 1 - \frac{\omega}{4}\lambda_{ij}(A_h^{FEM}) = 1 - \omega\left(\sin^2\left(\frac{i\pi h}{2}\right) + \sin^2\left(\frac{j\pi h}{2}\right)\right).$$

A motiváló egydimenziós példához hasonlóan

$$\min_{i,j=1,..,n} \mu_{ij} \geq \min_{-1\leq x,y\leq 1}\left(1 - \omega\left(\sin^2\left(\frac{\pi x}{2}\right) + \sin^2\left(\frac{\pi y}{2}\right)\right)\right) =: \min_{-1\leq x,y\leq 1} f_\omega(x, y) = 1-2\omega,$$

így $\omega < 1$ esetén $1 - 2\omega =: -Q > -1$.

Az optimális simításhoz az „ $x \geq \frac{1}{2}$ vagy $y \geq \frac{1}{2}$ ” feltétel mellett keresendő f_ω minimuma, mert a durva rácshoz az „ $x < \frac{1}{2}$ és $y < \frac{1}{2}$ ” feltétel tartozik; lásd 9.33. feladat.

2. Általánosabb esetben, ha az $A_h := A_h^{FEM}$ mátrix nem pontosan azonos blokkokból áll, igazolható a regularitásból, hogy A_h korlátos h-tól függetlenül, lásd pl. [10]. Ezért ha $0 \leq Q < 1$ adott állandó és $\theta \geq \frac{\|A\|_2}{1+Q}$, akkor

$$\mu_{ij} = 1 - \frac{1}{\theta}\lambda_{ij}(A_h) \geq 1 - \frac{1}{\theta}\|A\|_2 \geq -Q,$$

így teljesül az 5.7. állítás feltétele. ♢

Más diszkretizációk. A fentiekben részletesen megvizsgáltuk 2D FEM esetén a szakaszonként lineáris/bilineáris esetet. Más diszkretizációk esetén alkalmas módosításokkal értelmezett approximációs és simító tulajdonságból vezethető le a konvergencia: ezek általában abban a formában igazak, hogy megfelelő $\alpha > 0$ mellett

$$|(A_h^{-1} - PA_H^{-1}P^T)b|_{A_h} \leq Ch^\alpha|b| \qquad (\forall b \in \mathbb{R}^n),$$

$$|A_h(I - B^{-1}A_h)^k x| \leq \varepsilon_k h^{-\alpha}|x|_{A_h} \qquad (\forall x \in \mathbb{R}^n,\ k \in \mathbb{N})$$

(lásd pl. [10, 27]). Ekkor a két h-hatvány kiejti egymást, így a végkövetkeztetés érvényben marad. Például:

- 2D FDM esetén: $\alpha = 2$;
- 3D FEM esetén: $\alpha = 1/2$ (lineáris/bilineáris elemek).

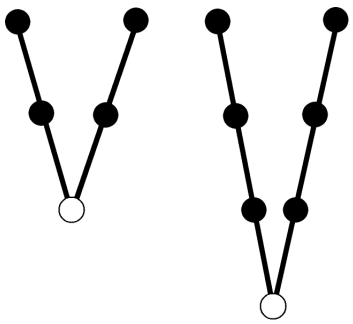
5.3. Többrácsos algoritmusok

Konstrukció és konvergencia. A fenti kétrácsos módszerben a simítás után restrikcióval „leszállunk” a durvább rácsra, ott valahogy megoldjuk a feladatot, majd prolongációval visszatérünk a finom rácsra és ott (a gyakorlatban használt szimmetrizált esetben előbb utósimítást is alkalmazva) javítjuk az előző közelítést. Felmerül a kérdés, hogyan oldjuk meg a feladatot a durva rácson. Ha itt direkt megoldást használunk, akkor is nyerünk a kétrácsos módszerrel, pl. 2D esetben $H = 2h$ esetén a durva rácshoz kb. 1/4-szer annyi ismeretlen tartozik, amire a megoldás költsége (a mátrix ritkasága révén $O(n^2)$-tel becsülve) kb. 1/16-ára csökken. A durva rácson sem muszáj azonban direkt megoldást használni, hanem erre is alkalmazható újabb kétrácsos modszer (így összességében már háromrácsos a módszer), ill. mindez még tovább folytatható.

A legegyszerűbb L-rácsos módszerben így egyre durvább rácsokra szállunk le restrikciókkal, a legdurvább rácson direkt megoldást használunk, majd prolongációkkal visszalépegetünk a legfinomabb rácsig, miközben a le- és felszálláskor az egyes szinteken simítunk. Ez az ún. *V-ciklus*, mely nevét a sematikus diagramról kapta (5.2. ábra). Formálisan, jelöljön K_2 egy, a fentiekben leirt kétrácsos ciklust, P a prolongációt és R a restrikciót. Ekkor az L-rácsos V-ciklust a

$$K_L := PK_{L-1}R \qquad (L = 3, 4, \dots)$$

rekurzió definiálja. Az L-rácsos módszerben egy ilyen ciklust ismételünk iterációként.



5.2. ábra. A V-ciklus szemléltetése három és négy rács esetén.

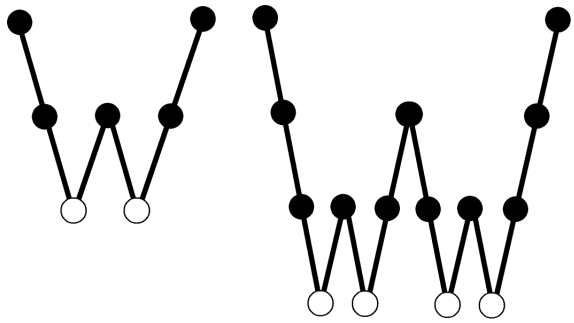
Másik változatot kapunk, ha több munkát fordítunk a durva komponensekre (több javítás), és 2-szer alkalmazzuk az alacsonyabb szintet. Ezzel az L-rácsos ún. *W-ciklust* a

$$K_L := PK_{L-1}^2R \qquad (L = 3, 4, \dots)$$

rekurzió definiálja (5.3. ábra). Most is egy ilyen ciklust ismételünk iterációként.

Elvileg lehet meg többször alkalmazni az alacsonyabb szintet:

$$K_L := PK_{L-1}^\gamma R \qquad (L = 3, 4, \dots),$$



5.3. ábra. A W-ciklus szemléltetése három és négy rács esetén.

ahol $\gamma \in \mathbb{N}^+$ adott egész. Azonban, mint látni fogjuk, 2D feladat esetén az optimális műveletigény miatt csak $\gamma \leq 3$ felel meg. A gyakorlatban elég a $\gamma = 1$ vagy 2 választás, azaz a V- vagy a W-ciklust használják.

Az L-rácsos módszerek konvergenciáját itt nem bizonyítjuk, a levezetéseket lásd pl. a [10, 27] könyvekben. Lényege ugyanaz, mint a kétrácsosnál: az algoritmus összetevői a kétrácsos módszerhez hasonlóan

- simítás $\longrightarrow$ kontraktív;
- restrikció $\longrightarrow$ korlátos;
- megoldás durva rácson $\longrightarrow$ korlátos;
- prolongáció $\longrightarrow$ korlátos.

(A durva rácson való megoldásnál a restrikcióval és prolongációval együtt a korlátosság az approximációs tulajdonság érvényesülését jelenti.) Ezért az e lépésekből összetevődő teljes eljárás is kontraktív, és így iterációként ismételve konvergens.

Műveletigény. A MG-módszer legfontosabb előnye az optimális műveletigény abban az értelemben, hogy arányos a változók számával, vagyis a lehető legkisebb. Először egy adott ℓ-edik szint műveletigényére adunk becslést. Legyen itt a mátrix mérete N_ℓ. Tegyük fel egyszerűség kedvéért, hogy olyan 2D feladatban vagyunk, ahol az egyes szintek közt kétszeres rácstávolsághoz pontosan négyszeres mátrixméret tartozik, azaz $N_k = 4N_{k-1}$ $(k = 2, ..., \ell)$.

A k-adik szinten jelölje $\overline{Q}_k$ az alábbiak műveletigényét: simítás, restrikció, prolongáció, reziduális kiszámítása. A mátrix sávos szerkezete miatt ezekre érvényes az $O(N_k)$ nagyságrend, azaz

$$\overline{Q}_k \leq c\, N_k \qquad (k = 1, ..., \ell).$$

Az ℓ-edik szinten jelölje Q_ℓ az összes műveletigényt. Ehhez rekurzívan juthatunk el. Minden k-adik szinten szükséges egyrészt a fenti négy lépés, ami $\overline{Q}_k$ műveletet igényel, ill. γ-szor ismételjük a $(k-1)$-edik szintet, így ennyiszer hozzáadódik annak teljes Q_{k-1} műveletigénye:

$$Q_k = \overline{Q}_k + \gamma Q_{k-1} \qquad (k = 2, \dots, \ell - 1).$$

Itt legyen $1 \le \gamma \le 3$ állandó. Az első szinten a durva megoldás műveletigénye lép fel. Mivel ez fix rács, így a műveletigény adott, erre írható $Q_1 \le cN_1$.

5.9. Állítás. *Van olyan (ℓ-től független) $c_0 > 0$ konstans, hogy*

$$Q_\ell \le c_0 N_\ell.$$

Bizonyítás. Rekurzióval

$$Q_\ell = \overline{Q}_\ell + \gamma Q_{\ell-1} = \overline{Q}_\ell + \gamma\overline{Q}_{\ell-1} + \gamma^2 Q_{\ell-2} = \dots = \sum_{k=0}^{\ell-2} \gamma^k \overline{Q}_{\ell-k} + \gamma^{\ell-1}\overline{Q}_1$$

$$\le c\left(\sum_{k=0}^{\ell-2} \gamma^k N_{\ell-k} + \gamma^{\ell-1} N_1\right) = c\sum_{k=0}^{\ell-1} \gamma^k N_{\ell-k} = c\, N_\ell \sum_{k=0}^{\ell-1} \Big(\frac{\gamma}{4}\Big)^k < \frac{4c}{4-\gamma} N_\ell = c_0 N_\ell\,,$$

ahol a mértani sor $\gamma \le 3$ miatt konvergens és $c_0 := \frac{4c}{4-\gamma}$. □

A teljes ciklus műveletigénye az alábbi: az $\ell = 1, \dots, L$ szinteket összegezve

$$\sum_{\ell=1}^{L} Q_\ell \le c_0 \sum_{\ell=1}^{L} N_\ell = c_0 \sum_{\ell=1}^{L} \frac{N_L}{4^{L-\ell}} = c_0 \sum_{k=0}^{L-1} \frac{N_L}{4^k} < \frac{4c_0}{3} N_L.$$

Végül az iteráció során adott pontossághoz korlátos számú lépés kell, mert a $\sigma < 1$ konvergenciahányados nem függ a rácsoktól. Így a teljes eljárás műveletigénye is

$$O(N_L),$$

ahol N_L a kiindulási rácshoz tartozó ismeretlenek száma, azaz a műveletigény optimális.

6. fejezet

Nyeregpont-feladatok numerikus megoldása

Ebben a fejezetben egy speciális szerkezetű elliptikus PDE-rendszerrel foglalkozunk az ennek hátterét alkotó absztrakt feladattípus elmélete alapján. Ez a rendszer az ún. Stokes-feladat. Alkalmas transzformációval bizonyos értelemben visszavezetjük a korábbi koercivitási megközelítésre, de a végeselemes megoldásnál a skalár elliptikus egyenletekhez képest megszorítás jelenik meg, ami LBB-feltétel néven vált ismertté. Megemlítjük, hogy más feladatokban is megjelenik a nyeregpont-szerkezet (pl. skalár elliptikus egyenletek vegyes végeselemes megoldásánál), ezeket azonban itt nem tárgyaljuk. Mindezekről pl. a [6] cikk és a [27, III. 17.5. fejezet] könyv nyújt részletes összefoglalást. Az elméleti háttér tömör leírásában a [16] könyvet követjük.

Áramlási feladatokban lép fel az alábbi PDE-rendszer, melyet Stokes-feladatnak hívunk:

$$\begin{cases} -\Delta \mathbf{u} + \nabla p = \mathbf{f} \\ \operatorname{div} \mathbf{u} = 0 \\ \mathbf{u}_{|\partial\Omega} = 0\,, \end{cases} \tag{6.1}$$

ahol $\Omega \subset \mathbb{R}^N$ ($N = 2$ vagy 3) korlátos tartomány szakaszonként sima peremmel, $\mathbf{u} : \Omega \to \mathbb{R}^N$ az áramlás sebességvektora és $p : \Omega \to \mathbb{R}$ a nyomás. Az $\mathbf{f} : \Omega \to \mathbb{R}^N$ adott függvény a külső erőkből származtatható, a $-\Delta \mathbf{u}$ kifejezés és az $\mathbf{u}_{|\partial\Omega} = 0$ peremfeltétel koordinátánként értendő. A (6.1) rendszer időben stacionárius lassú áramlást ír le. (Időben változó áramlás esetén a megfelelő első egyenletet időben diszkretizálva a fentihez hasonló feladatot kapunk, de az első egyenlet kiegészül egy $\frac{1}{\tau}\mathbf{u}$ taggal, ahol $\tau > 0$; az alább elmondottak erre az esetre is értelemszerűen átvihetők.)

6.1. Megoldhatóság, inf-sup-feltétel

Kiindulásunk az, hogy a (6.1) feladat speciális szerkezetű az alábbi miatt:

6.1. Állítás. *A* $\nabla : H_0^1(\Omega) \to L^2(\Omega)^N$ *operátor adjungáltjára*

$$\nabla^* \mathbf{u} = -\operatorname{div} \mathbf{u} \qquad (\forall \mathbf{u} \in H^1(\Omega)^N).$$

(A bizonyításhoz lásd a 9.34. feladatot.)

Emiatt a Stokes-feladat egyenleteit átírhatjuk

$$\left\{ \begin{array}{rcl} -\Delta \mathbf{u} + \nabla p & = & \mathbf{f} \\ \nabla^* \mathbf{u} & = & 0 \end{array} \right. \tag{6.2}$$

alakba (a 2. egyenletben a mínusz szorzót elhagyhattuk), az $\mathbf{u}_{|\partial\Omega} = 0$ peremfeltételt pedig majd az alaptérbe építjük be. A kapott alak miatt először vizsgáljunk meg ilyen szerkezetű operátoregyenlet-rendszereket. A felépítés egyszerűbb, ha előbb korlátos operátorra, majd ebből bilineáris formákra írjuk ezt fel, mivel az utóbbit a Stokes-feladat gyenge alakjára alkalmazhatjuk majd.

Operátorokkal megadott nyeregpont-feladatok. Tekintsük az

$$\left\{ \begin{array}{rcl} Au + Bp & = & f \\ B^* u & = & g \end{array} \right. \tag{6.3}$$

feladatot, ahol H, K valós Hilbert-terek, $f \in H$, $g \in K$, $A : H \to H$ és $B : K \to H$ korlátos lineáris operátorok, valamint A önadjungált és egyenletesen pozitív is, azaz van olyan $m > 0$, hogy

$$\langle Au, u \rangle \geq m \|u\|^2 \qquad (\forall u \in H). \tag{6.4}$$

6.2. Megjegyzés.

(i) Ez lényegében egy operátormátrixra vonatkozó egyenlet a $H \times K$ szorzattéren. A „nyeregpont-feladat” elnevezés abból származik, hogy a fenti rendszer megoldása egy alkalmas kvadratikus típusú funkcionál nyeregpontjaként áll elő. Éspedig [16, 14.4. áll.], a $\Psi : H \times K \to \mathbb{R}$,

$$\Psi(u, p) = \langle Au, u \rangle + 2\langle Bp, u \rangle - 2\langle f, u \rangle - 2\langle g, p \rangle$$

funkcionálra fennáll, hogy ha $(u^*, p^*) \in H \times K$ a (6.3) feladat megoldása, akkor bármely $(u, p) \in H \times K$ esetén

$$\Psi(u^*, p) \leq \Psi(u^*, p^*) \leq \Psi(u, p^*).$$

(ii) A $\|.\|$ jelölést a H és K terekben két különböző normára fogjuk használni. Bár lehetne a precízség kedvéért $\|.\|_H$ és $\|.\|_K$ jelöléseket is írni, ezt – a kevesebb index érdekében – nem tesszük, mivel a környezetből egyértelmű, melyik normáról van szó. ♢

A (6.3) rendszer megoldásának alapja, hogy A bijekció voltát kihasználva átrendezzük az egyenletet. Fejezzük ki az első egyenlőségből u-t:

$$u = A^{-1}(f - Bp), \tag{6.5}$$

és helyettesítsük a másodikba:

$$B^*A^{-1}(f - Bp) = g, \quad \text{azaz} \quad B^*A^{-1}Bp = B^*A^{-1}f - g =: \tilde{g}. \tag{6.6}$$

Legyen

$$S := B^*A^{-1}B, \tag{6.7}$$

ez az ún. *Schur-féle komplementer-operátor*. Itt $B^* : H \to K$, így $S : K \to K$. Ha meg tudjuk oldani a (6.6)-ban kapott

$$Sp = \tilde{g} \tag{6.8}$$

egyenletet a K téren, akkor (6.5)-ből u-t is megkapjuk, így kész vagyunk.

A (6.8) egyenlet megoldhatósága nem nyilvánvaló, mivel B, ill. B^* általában nem bijekciók. Mégis szeretnénk, hogy S (korlátos lineáris) bijekció legyen a K téren. Triviálisan szükséges feltétel ehhez, hogy B injektív legyen. Ha B inverzének korlátosságát felírjuk B-vel, akkor ki fog derülni, hogy ez már elég is lesz, vagyis a megoldhatóság kulcsa az alábbi feltétel lesz:

$$\text{van olyan } \gamma > 0, \text{ hogy} \quad \|Bp\| \geq \gamma\|p\| \quad (\forall p \in K). \tag{6.9}$$

Ezt az alábbi alakban szokás felírni:

6.3. Definíció. A (6.3) feladathoz tartozó **inf-sup-feltétel**:

$$\inf_{p\in K\setminus\{0\}} \sup_{u\in H\setminus\{0\}} \frac{\langle Bp, u\rangle}{\|p\|\|u\|} =: \gamma > 0. \tag{6.10}$$

♢

Könnyen látható, hogy (6.9) és (6.10) ekvivalens, lásd 9.35. feladat.

6.4. Tétel. *Legyenek H, K valós Hilbert-terek, $A \in B(H)$ és $B \in B(K, H)$, ahol A önadjungált és egyenletesen pozitív. Ha fennáll a (6.10) inf-sup-feltétel, akkor bármely $(f, g) \in H \times K$ esetén a (6.3) feladatnak létezik egyetlen $(u^*, p^*) \in H \times K$ megoldása.*

A bizonyítás lényege, hogy az inf-sup-feltételből levezethető az $S : K \to K$ Schur-féle komplementer-operátor egyenletes pozitivitása, így bijekció a K téren; lásd [16, 7.23. tétel].

Bilineáris formákkal megadott nyeregpont-feladatok. A korlátos bilineáris formák és lineáris funkcionálok Riesz-reprezentációja segítségével a fentiek közvetlenül átvihetők bilineáris formákkal megadott hasonló szerkezetű feladatokra. Legyenek most a fenti H, K valós Hilbert-tereken $a : H \times H \to \mathbb{R}$ és $b : K \times H \to \mathbb{R}$ korlátos bilineáris formák, ahol a koercív: van olyan $m > 0$, hogy

$$a(u,u) \geq m \|u\|^2 \qquad (\forall u \in H), \tag{6.11}$$

legyenek továbbá $\phi : H \to \mathbb{R}$ és $\psi : K \to \mathbb{R}$ korlátos lineáris funkcionálok. Tekintsük az alábbi feladatot: keresendő $(u,p) \in H \times K$, melyre

$$\begin{aligned} a(u,v) + b(p,v) &= \phi v \qquad (\forall v \in H), \\ b(q,u) &= \psi q \qquad (\forall q \in K). \end{aligned} \tag{6.12}$$

A fő tulajdonság most is a megfelelő inf-sup-feltétel:

$$\inf_{p \in K \setminus \{0\}} \sup_{u \in H \setminus \{0\}} \frac{b(p,u)}{\|p\| \|u\|} =: \gamma > 0. \tag{6.13}$$

6.5. Tétel. *(Lásd [16, 7.29. tétel]). Legyenek H, K valós Hilbert-terek, $a : H \times H \to \mathbb{R}$ és $b : K \times H \to \mathbb{R}$ korlátos bilineáris formák, ahol az a forma koercív is. Ha fennáll a (6.13) inf-sup-feltétel, akkor bármely $\phi : H \to \mathbb{R}$ és $\psi : K \to \mathbb{R}$ korlátos lineáris funkcionálok esetén a (6.12) feladatnak létezik egyetlen $(u,p) \in H \times K$ megoldása.*

A Stokes-feladat gyenge megoldása. Először értelmezzük a 6.1 rendszer gyenge alakját. A gyenge megoldásnál az $\mathbf{u}$ függvényt (a $-\Delta \mathbf{u}$ kifejezés és az $\mathbf{u}_{|\partial\Omega} = 0$ peremfeltétel miatt) a $H_0^1(\Omega)^N$ szorzattérben keressük, melyet most is valós értékű függvényekkel definiálunk, így valós Hilbert-tér. A p nyomásnál is szeretnénk a deriválttól megszabadulni és csak $L^2(\Omega)$-ban keresni. Mivel a (6.1) egyenletek a p függvényt csupán additív konstans erejéig határozzák meg, így az egyértelműség érdekében bevezetjük az alábbi teret:

$$\dot{L}^2(\Omega) := \{p \in L^2(\Omega) : \int_\Omega p = 0\} \tag{6.14}$$

a szokásos L^2-skalárszorzattal és $\|.\|_0$ indukált normával. A gyenge alak felírásához a szokott módon a két egyenletet rendre beszorozzuk $\mathbf{v} = (v_1, v_2, ..., v_N) \in H_0^1(\Omega)^N$ és $q \in \dot{L}^2(\Omega)$ tesztfüggvényekkel, majd alkalmazzuk a Green-formulát, ill. Gauss–Osztrogradszkij-tételt. A deriváltmátrixokra használjuk a (3.63) Frobenius-skalárszorzatot, amelyre $\nabla \mathbf{u} : \nabla \mathbf{v} := \sum_{i=1}^N \nabla u_i \cdot \nabla v_i$ teljesül, ill. az

$$\langle \mathbf{u}, \mathbf{v} \rangle_{H_0^1} := \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v}, \qquad |\mathbf{u}|_1^2 = \langle \mathbf{u}, \mathbf{u} \rangle_{H_0^1}$$

skalárszorzatot és indukált normát a $H_0^1(\Omega)^N$ téren.

6.6. Definíció. Az $(\mathbf{u}, p) \in H_0^1(\Omega)^N \times \dot{L}^2(\Omega)$ függvénypárt a (6.1) feladat *gyenge megoldásának* nevezzük, ha

$$\begin{cases} \displaystyle\int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v} - \int_\Omega p\,(\operatorname{div}\mathbf{v}) = \int_\Omega \mathbf{f}\cdot\mathbf{v} & (\forall \mathbf{v} \in H_0^1(\Omega)^N), \\ \displaystyle\int_\Omega q\,(\operatorname{div}\mathbf{u}) = 0 & (\forall q \in \dot{L}^2(\Omega)). \end{cases} \tag{6.15}$$

♢

A (6.1) feladat gyenge megoldhatóságához a 6.5. tételt szeretnénk felhasználni. Vezessük be az alábbi bilineáris formákat:

$$\begin{aligned} &a : H_0^1(\Omega)^N \times H_0^1(\Omega)^N \to \mathbb{R}, \quad a(\mathbf{u},\mathbf{v}) := \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}, \\ &b : \dot{L}^2(\Omega) \times H_0^1(\Omega)^N \to \mathbb{R}, \qquad b(p,\mathbf{v}) := -\int_\Omega p\,(\operatorname{div}\mathbf{v}). \end{aligned} \tag{6.16}$$

Ekkor a (6.15) rendszer éppen (6.12) alakú.

6.7. Állítás. *A fenti b formára teljesül az inf-sup-feltétel:*

$$\inf_{\substack{p\in\dot{L}^2(\Omega)\\ p\neq 0}} \sup_{\substack{\mathbf{u}\in H_0^1(\Omega)^N\\ \mathbf{u}\neq\mathbf{0}}} \frac{b(p,\mathbf{u})}{\|p\|_0|\mathbf{u}|_1} =: \gamma > 0. \tag{6.17}$$

Bizonyítás. Ez a divergencia-operátor szuperjektivitásának köszönhető, pontosabban, hogy [22] alapján bármely $p \in \dot{L}^2(\Omega)$ esetén van olyan $\mathbf{u} \in H_0^1(\Omega)^N$, melyre $p = -\operatorname{div}\mathbf{u}$ és $\|p\|_0 \geq \gamma|\mathbf{u}|_1$ alkalmas $\gamma > 0$ állandóval, így ezzel az $\mathbf{u}$-val

$$b(p,\mathbf{u}) = -\int_\Omega p\,(\operatorname{div}\mathbf{u}) = \int_\Omega p^2 = \|p\|_0^2.$$

Így tehát minden $p \in L^2(\Omega)$ függvényhez létezik $\mathbf{u} \in H_0^1(\Omega)^N$, hogy

$$\frac{b(p,\mathbf{u})}{\|p\|_0|\mathbf{u}|_1} = \frac{\|p\|_0}{|\mathbf{u}|_1} \geq \gamma,$$

ez pedig ekvivalens a (6.17) egyenlőtlenséggel. □

6.8. Megjegyzés. A γ konstans értéke Ω-tól függ, és nem adható rá általános képlet. Vannak azonban ismert explicit értékek speciális tartományokra és jól használható általános becslések is, lásd pl. [20, 26]. ♢

6.9. Tétel. *Bármely* $\mathbf{f} \in L^2(\Omega)^N$ *függvény esetén a (6.1) feladatnak létezik egyetlen* $(\mathbf{u}, p) \in H_0^1(\Omega)^N \times \dot{L}^2(\Omega)$ *gyenge megoldása.*

Bizonyítás. Itt $a : H_0^1(\Omega)^N \times H_0^1(\Omega)^N \to \mathbb{R}$ korlátos és koercív bilineáris forma, hiszen ez épp a $H_0^1(\Omega)^N$ tér skalárszorzata. Másrészt $b : \dot{L}^2(\Omega) \times H_0^1(\Omega)^N \to \mathbb{R}$ korlátos bilineáris forma, hiszen

$$|b(p, \mathbf{v})| \leq \|p\|_0 \| \operatorname{div} \mathbf{v}\|_0 \leq \sqrt{N}\|p\|_0 |\mathbf{v}|_1,$$

felhasználva, hogy

$$\| \operatorname{div} \mathbf{v}\|_0^2 = \int_\Omega \Big(\sum_{i=1}^N \partial_i v_i\Big)^2 \leq N \int_\Omega \sum_{i=1}^N (\partial_i v_i)^2 \leq N \int_\Omega \sum_{i,j=1}^N (\partial_i v_j)^2 = N\|\nabla \mathbf{v}\|_0^2 = N|\mathbf{v}|_1^2.$$

Mivel a 6.7. állítás alapján fennáll a (6.17) inf-sup-feltétel, a 6.5. tételből nyerjük a kívánt megoldhatóságot. □

6.10. Megjegyzés. A fenti becslésben az N-es szorzó valójában elhagyható: igazolható, hogy

$$\| \operatorname{div} \mathbf{v}\|_0 \leq |\mathbf{v}|_1 \qquad (\forall \mathbf{v} \in H_0^1(\Omega)^N). \tag{6.18}$$

Könnyen látható, hogy a becslés éles is (lásd 9.36. feladat). ♢

6.2. Az Uzawa-algoritmus

Ez az algoritmus egy olyan iterációs módszer, amit itt először elvi szinten magára a diszkretizáció nélküli PDE-re írunk fel, ez ui. a PDE szerkezete miatt segíti az érthetőséget. A következő szakaszban visszük át a módszert a végeselemes esetre.

Absztrakt Uzawa-iteráció. Tekintsük először ismét a (6.3) feladatot Hilbert-térben:

$$\left\{ \begin{array}{lcl} Au + Bp & = & f \\ B^* u & = & g, \end{array} \right. \tag{6.19}$$

az ott tett feltételekkel, ezen belül teljesüljön a (6.10) inf-sup-feltétel. A (6.19) feladat iterációsan megoldható ugyanazon elven, mint a megoldhatóság igazolása történt az előző szakaszban: a feladat visszavezethető az $S := B^* A^{-1} B$ Schur-féle komplementer-operátorra vonatkozó

$$Sp = \tilde{g} \tag{6.20}$$

egyenletre, ahol $\tilde{g} := B^* A^{-1} f - g$. Itt ugyanis S egyenletesen pozitív, így alkalmazhatjuk a 4. szakasz módszereit (pontosabban azok Hilbert-térbeli megfelelőjét, lásd [16]).

Az Uzawa-algoritmus nem más, mint az állandó lépésközű egyszerű iteráció a (6.20) egyenletre.

6.11. Tétel. *A (6.19) feladat feltételei mellett tekintsük az alábbi iterációt. Legyenek $u_0 \in H$, $p_0 \in K$ tetszőlegesek és $\alpha > 0$ adott szám, ha pedig $n \in \mathbb{N}$ és megvan $u_n \in H$ és $p_n \in K$, akkor*

$$\begin{cases} Au_{n+1} + Bp_n = f & \text{(azaz } u_{n+1} \text{ ennek megoldása)}, \\ p_{n+1} := p_n + \alpha(B^* u_{n+1} - g). \end{cases} \tag{6.21}$$

Ekkor van olyan $\alpha_0 > 0$, hogy $0 < \alpha < \alpha_0$ esetén a fenti iteráció lineárisan konvergál, vagyis alkalmas $c_1, c_2 > 0$ és $q < 1$ mellett

$$\|u_n - u^*\| \le c_1 q^n, \qquad \|p_n - p^*\| \le c_2 q^n \qquad (n \in \mathbb{N}).$$

Itt $\alpha_0 = \frac{2m}{\|B\|^2}$. Az optimális paraméter $\alpha_{opt} = \frac{2}{\Lambda+\lambda}$, ahol $\lambda := \frac{\gamma^2}{\|A\|^2}$ és $\Lambda := \frac{\|B\|^2}{m}$; ekkor $q = \frac{\Lambda-\lambda}{\Lambda+\lambda}$.

A bizonyítás lényege, hogy a $p_{n+1} := p_n - \alpha(Sp_n - \tilde{g})$ egyszerű iterációt kettédaraboljuk S definícióját felhasználva, lásd [16, 16.15. tétel].

Az egyszerű iteráció helyett a konjugált gradiens-módszert is használhatjuk a (6.20) egyenletre, lásd [16, 16.4. fejezet].

Térjünk most át bilineáris formákkal megadott nyeregpont-feladatra. Tekintsük ismét a (6.11) feladatot Hilbert-térben:

$$\begin{aligned} a(u,v) + b(p,v) &= \phi v \qquad (\forall v \in H), \\ b(q,u) &= \psi q \qquad (\forall q \in K) \end{aligned} \tag{6.22}$$

az ott tett feltételekkel, ezen belül teljesüljön a (6.13) inf-sup-feltétel. A 6.11. tétel a megoldhatóságnál látott módszerrel, azaz Riesz-reprezentáció segítségével közvetlenül átvihető a (6.22) feladatra. Ebből adódik:

6.12. Tétel. *A (6.22) feladat feltételei mellett tekintsük az alábbi iterációt. Legyenek $u_0 \in H$, $p_0 \in K$ tetszőlegesek és $\alpha > 0$ adott szám, ha pedig $n \in \mathbb{N}$ és megvan $u_n \in H$ és $p_n \in K$, akkor*

$$\begin{aligned} a(u_{n+1}, v) + b(p_n, v) &= \phi v \qquad (\forall v \in H), \\ \langle p_{n+1}, q\rangle &= \langle p_n, q\rangle + \alpha\big(b(q, u_{n+1}) - \psi q\big) \qquad (\forall q \in K) \end{aligned} \tag{6.23}$$

(azaz u_{n+1} és p_{n+1} ezeknek a feladatnak a megoldásai).

Ekkor van olyan $\alpha_0 > 0$, hogy $0 < \alpha < \alpha_0$ esetén a fenti iteráció lineárisan konvergál, vagyis alkalmas $c_1, c_2 > 0$ és $q < 1$ mellett

$$\|u_n - u^*\| \le c_1 q^n, \qquad \|p_n - p^*\| \le c_2 q^n \qquad (n \in \mathbb{N}).$$

Ha a két forma normájára a $\beta := \|b\|$ és $M := \|a\|$, ill. az a forma koercivitási konstansára az m jelölést használjuk, akkor $\alpha_0 = \frac{2m}{\beta^2}$ és az optimális paraméter $\alpha_{opt} = \frac{2}{\Lambda+\lambda}$, ahol $\lambda := \frac{\gamma^2}{M^2}$ és $\Lambda := \frac{\beta^2}{m}$; ekkor $q = \frac{\Lambda-\lambda}{\Lambda+\lambda}$.

Uzawa-iteráció a Stokes-feladatra. A fentiekből már közvetlenül származtatható az iteráció a (6.1) Stokes-feladatra. Láttuk, hogy ennek (6.15) gyenge megoldása éppen (6.22) alakú az ott tett feltételekkel, így alkalmazható rá a 6.12. tétel. A (6.16) bilineáris formák határaira az alábbi konkrét értékek vonatkoznak. Egyrészt

$$a(\mathbf{u},\mathbf{v}) := \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v} = \langle \mathbf{u},\mathbf{v}\rangle_{H_0^1},$$

azaz épp a $H_0^1(\Omega)^N$ tér skalárszorzata, így normájára és koercivitási konstansára $M = m = 1$. Másrészt a 6.10. megjegyzés révén

$$|b(p,\mathbf{v})| = \left|\int_\Omega p\,(\operatorname{div}\mathbf{v})\right| \le \|p\|_0 \|\operatorname{div}\mathbf{v}\|_0 \le \|p\|_0 |\mathbf{v}|_1 \qquad (\forall p\in \dot L^2(\Omega),\ \mathbf{v}\in H_0^1(\Omega)^N)$$

és ez éles becslés, így $\beta := \|b\| = 1$. A γ inf-sup-konstans a (6.17)-ben szereplő érték, lásd 6.8. megjegyzés.

6.13. Tétel. *Tekintsük a Stokes-feladatra az alábbi iterációt. Legyenek $\mathbf{u}_0 \in H_0^1(\Omega)^N$, $p_0 \in \dot L^2(\Omega)$ tetszőlegesek és $\alpha > 0$ adott szám, ha pedig $n\in\mathbb{N}$ és megvan $\mathbf{u}_n \in H_0^1(\Omega)^N$ és $p_n \in \dot L^2(\Omega)$, akkor*

$$\begin{aligned} \int_\Omega \nabla\mathbf{u}_{n+1} : \nabla\mathbf{v} - \int_\Omega p_n\,(\operatorname{div}\mathbf{v}) &= \int_\Omega \mathbf{f}\cdot\mathbf{v} \qquad (\forall v\in H_0^1(\Omega)^N),\\ \int_\Omega p_{n+1}\,q = \int_\Omega p_n\,q - \alpha\int_\Omega (\operatorname{div}\mathbf{u}_{n+1})\,q &\qquad (\forall q\in \dot L^2(\Omega)) \end{aligned} \tag{6.24}$$

(azaz u_{n+1} és p_{n+1} ezeknek a feladatnak a megoldásai).

Ha teljesül a (6.17) inf-sup-feltétel, akkor $0<\alpha<2$ esetén a (6.24) iteráció lineárisan konvergál, vagyis alkalmas $c_1,c_2>0$ és $q<1$ mellett

$$|\mathbf{u}_n - \mathbf{u}^*|_1 \le c_1 q^n, \qquad \|p_n - p^*\|_0 \le c_2 q^n \qquad (n\in\mathbb{N}).$$

Itt az optimális paraméter és hozzátartozó konvergenciahányados rendre $\alpha_{opt} = \frac{2}{1+\gamma^2}$ és $q = \frac{1-\gamma^2}{1+\gamma^2}$,

Bizonyítás. A (6.24) iteráció megegyezik a (6.15) feladatra alkalmazott (6.23) iterációval, így érvényes rá a 6.12. tétel a $H = H_0^1(\Omega)^N$ és $K = \dot L^2(\Omega)$ terekben. Mivel most $M = m = 1$ és $\beta = 1$, így $\alpha_0 = 2$, ill. $\lambda = \gamma^2$ és $\Lambda = 1$, melyekkel $\alpha_{opt} = \frac{2}{1+\gamma^2}$ és $q = \frac{1-\gamma^2}{1+\gamma^2}$. □

6.14. Megjegyzés. A (6.24) iteráció nem más, mint a

$$\begin{aligned} -\Delta\mathbf{u}_{n+1} + \nabla p_n &= \mathbf{f}, \qquad \mathbf{u}_{n+1\,|\partial\Omega} = 0,\\ p_{n+1} &= p_n - \alpha\,\operatorname{div}\mathbf{u}_{n+1} \end{aligned}$$

feladat gyenge alakja. ♢

6.3. A Stokes-feladat végeselemes megoldása

A Stokes-feladat végeselemes diszkretizációjához alkalmas $V_h \subset H_0^1(\Omega)^N$ és $P_h \subset \dot{L}^2(\Omega)$ véges dimenziós altereket választunk, és ezekben keressük a (6.15) feladat megoldását, azaz olyan $(\mathbf{u}^h, p^h) \in V_h \times P_h$ függvénypárt, melyre

$$\begin{cases} \displaystyle\int_\Omega \nabla \mathbf{u}^h : \nabla \mathbf{v}^h - \int_\Omega p^h \,(\operatorname{div} \mathbf{v}^h) = \int_\Omega \mathbf{f} \cdot \mathbf{v}^h & (\forall \mathbf{v}^h \in V_h), \\ \displaystyle\int_\Omega q^h \,(\operatorname{div} \mathbf{u}^h) = 0 & (\forall q^h \in P_h). \end{cases} \tag{6.25}$$

Először vizsgáljuk meg ennek megoldhatóságát. Szeretnénk a 6.5. tételt használni. Ez részben jónak látszik, mert a bilineáris formák korlátossága ugyanúgy teljesül, mint az eredeti $H_0^1(\Omega)^N$ és $\dot{L}^2(\Omega)$ terekben, hiszen ugyanazt a becslést most az eredeti Szoboljev-terek altereire használhatjuk.

A diszkretizáció kulcsproblémája, hogy a fentiekkel szemben a (6.17) inf-sup-feltétel megfelelője nem öröklődik automatikusan az alterekre. A $V_h \subset H_0^1(\Omega)^N$ altéren vett szuprémum akár 0 is lehet, ha a (V_h, P_h) párt nem jól választjuk. Ezért (6.17) megfelelőjét külön fel kell tennünk az elméleti vizsgálathoz. A szakasz végén kitérünk a feltétel teljesíthetőségére.

6.15. Definíció. A V_h és P_h alterek teljesítik a (6.25) feladathoz tartozó *LBB-feltételt* (Ladüzsenszkaja-Babuška-Brezzi-feltételt) vagy *diszkrét inf-sup-feltételt*, ha van olyan $\gamma_0 > 0$ h-tól független állandó, hogy

$$\inf_{p^h \in P_h \setminus \{0\}} \sup_{\mathbf{u}^h \in V_h \setminus \{0\}} \frac{-\int_\Omega p^h \,(\operatorname{div} \mathbf{u}^h)}{\|p^h\|_0 |\mathbf{u}^h|_1} \geq \gamma_0. \tag{6.26}$$

♢

A feltétel lényege, hogy V_h és P_h nem választható egymástól függetlenül: az u_h-ra vett supremum pozitív korlát fölöttisége azt követeli meg, hogy adott P_h esetén a V_h altér „elég nagy” legyen.

6.16. Tétel. *Ha teljesül a (6.26) LBB-feltétel, akkor a (6.25) feladatnak létezik egyetlen* $(\mathbf{u}_*^h, p_*^h) \in V_h \times P_h$ *megoldása.*

Bizonyítás. Ez (6.26) révén már megegyezik a 6.5. tétel bizonyításával $H_0^1(\Omega)^N \times \dot{L}^2(\Omega)$ helyett $(V_h \times P_h)$-ban. □

A megfelelő lineáris algebrai egyenletrendszer az alábbi alakú lesz:

$$\begin{bmatrix} A_h & B_h \\ B_h^T & 0 \end{bmatrix} \begin{bmatrix} c^h \\ r^h \end{bmatrix} = \begin{bmatrix} f^h \\ g^h \end{bmatrix}, \tag{6.27}$$

ahol c^h és r^h jelöli rendre $\mathbf{u}^h$ és p^h együtthatóvektorát a megfelelő bázisban. A rendszer alakja analóg a (6.19) feladatéval.

Az LBB-feltétel alapján emellett igazolható a Céa-lemma megfelelője: létezik olyan $c > 0$ állandó, hogy

$$|\mathbf{u}^* - \mathbf{u}_*^h|_1 + \|p^* - p_*^h\|_0 \leq c \left(\min_{\mathbf{v}_h \in V_h} |\mathbf{u}^* - \mathbf{v}^h|_1 + \min_{q_h \in P_h} \|p^* - q_h\|_0 \right),$$

ahol a $c > 0$ állandó γ_0-tól függ, de h-tól nem. (A bizonyítás megtalálható a [10, 27] könyvekben.) Ez függetleníti a hibabecslést a feladattól, vagyis a 3.4. szakasz interpolációs becsléseiből adódik a konvergencia.

Térjünk rá a (6.25) feladat iterációs megoldására! Az Uzawa-algoritmus a fenti helyzetben az LBB-feltétel révén a függvénytérbeli esettel teljesen analóg módon használható, azaz a 6.13. tételben $H_0^1(\Omega)^N \times P_h$ helyett $(V_h \times P_h)$ írandó:

6.17. Tétel. *Tekintsük a diszkretizált Stokes-feladatra az alábbi iterációt. Legyenek $\mathbf{u}_0^h \in V_h$, $p_0^h \in P_h$ tetszőlegesek és $\alpha > 0$ adott szám, ha pedig $n \in \mathbb{N}$ és megvan $\mathbf{u}_n^h \in V_h$ és $p_n^h \in P_h$, akkor*

$$\begin{aligned} \int_\Omega \nabla \mathbf{u}_{n+1}^h : \nabla \mathbf{v}^h - \int_\Omega p_n^h \, (\operatorname{div} \mathbf{v}^h) &= \int_\Omega \mathbf{f} \cdot \mathbf{v}^h \qquad (\forall v^h \in V_h), \\ \int_\Omega p_{n+1}^h \, q^h &= \int_\Omega p_n^h \, q^h - \alpha \int_\Omega (\operatorname{div} \mathbf{u}_{n+1}^h) \, q^h \qquad (\forall q^h \in P_h) \end{aligned} \tag{6.28}$$

(azaz u_{n+1}^h és p_{n+1}^h ezeknek a feladatnak a megoldásai).

Ha teljesül a (6.26) LBB-feltétel, akkor $0 < \alpha < 2$ esetén a (6.28) iteráció lineárisan konvergál, vagyis alkalmas $c_1, c_2 > 0$ és $q < 1$ mellett

$$|\mathbf{u}_n^h - \mathbf{u}_*^h|_1 \leq c_1 q^n, \qquad \|p_n^h - p_*^h\|_0 \leq c_2 q^n \qquad (n \in \mathbb{N}).$$

Itt az optimális paraméter és hozzátartozó konvergenciahányados rendre $\alpha_{opt} = \frac{2}{1+\gamma_0^2}$ és $q = \frac{1-\gamma_0^2}{1+\gamma_0^2}$,

Bizonyítás. Azonos a 6.13. tételével. □

6.18. Megjegyzés. A (6.28) Uzawa-iteráció lépéseiben a 6.14. megjegyzésbeli segédfeladatok végeselemes megoldását kell elvégezni, a (6.27) lineáris algebrai egyenletrendszer (6.21) alakú. ♢

Összességében tehát a végeselemes megvalósítás fő pontja olyan V_h és P_h alterek választása, melyek teljesítik a (6.26) LBB-feltételt. Az ilyen altereket *stabil térpároknak* hívjuk. A fő nehézség az, hogy egyrészt (mint láttuk) ha V_h túl szűk, azaz nem elég

nagy dimenziós, akkor a (6.26)-beli szuprémum sem elég nagy, azaz vagy 0, vagy h csökkentésével 0-hoz tart. Ha viszont P_h túl szűk, akkor p^* közelítése nem elég jó.

Gyakori ötlet háromszögrácson, hogy V_h bővítéséhez háromszögenként hozzáveszünk egy-egy új bázisfüggvényt, amely 0 a háromszög határán. Az E referenciaháromszögön ez $\varphi(\xi, \eta) = \xi\eta(1 - \xi - \eta)$. Alakja miatt szokás buborékfüggvénynek hívni.

Két gyakran használt példa stabil térpárra:

- V_h a szakaszonként lineáris folytonos függvények és a buborékfüggvények által kifeszített altér, P_h pedig azon szakaszonként lineáris folytonos függvényekből áll, melyek (6.14) miatt nulla átlagúak.

- V_h a szakaszonként másodfokú folytonos függvények és a buborékfüggvények által kifeszített altér, P_h pedig azon szakaszonként lineáris, de nem feltétlenül folytonos függvényekből áll, melyek nulla átlagúak.

Ezekről és további stabil térpárok konstrukciójáról olvashatunk a [10] könyvben és a [27] könyv III. 17.5.4. szakaszában.

7. fejezet

Nemlineáris elliptikus feladatok megoldása

7.1. Néhány egyszerű modellfeladat

Nemlineáris elliptikus feladatok sokféle helyzetben és szerkezettel felbukkannak. Ilyen feladatokról és numerikus megoldásukról bővebben a [12, 18] könyvekben olvashatunk. Itt néhány egyszerű feladatra adunk példát, ezen belül az egész fejezetben egy jól áttekinthető osztályra fogjuk ismertetni a megoldhatóságot és a numerikus megoldás menetét végeselemes diszkretizációval és alkalmas iterációval. Az elméleti hátteret [16] megfelelő részei alapján írjuk le.

Skalár-nemlinearitás. Legyen $\Omega \subset \mathbb{R}^n$ korlátos tartomány. Tekintsük először a

$$\begin{cases} -\operatorname{div}\big(a(|\nabla u|^2)\,\nabla u\big) = g, \\ u_{|\partial\Omega} = 0 \end{cases} \tag{7.1}$$

peremértékfeladatot, ahol $a : \mathbb{R}^+ \to \mathbb{R}^+$ adott korlátos C^1-beli függvény. Ez a feladat az (1.13)-belihez hasonló, de most ∇u együtthatója maga is ∇u-tól függ, ettől a feladat nemlineáris. Ilyen feladat lép fel például a Maxwell-egyenletek speciális stacionárius síkbeli esetében, melyet a bevezető 1.1. szakaszban említettünk. Itt általánosabb esetben az (1.5) egyenlőség (vagyis az egyenes arányosság) helyett H nemlineárisan függhet B-től úgy, hogy az irányuk azonos marad, azaz

$$H = a(|B|^2)\,B, \tag{7.2}$$

ahol $a : \mathbb{R}^+ \to \mathbb{R}^+$ adott korlátos és monoton C^1-beli függvény. Ez a (7.1)-beli egyenlethez vezet. A peremfeltétel a mező peremen előírt viselkedéséből adódik, a modellt részletesebben lásd a [18] könyvben. Az a függvény például néhány konkrét modellben

$$a(r) = \frac{r^k + \varrho}{r^k + \tau} \tag{7.3}$$

alakú, ahol $\varrho, \tau > 0$ és $k \in \mathbb{N}^+$ állandók. Egy másik fontos példa ilyen alakú egyenletre a képlékeny torzió egy modellje, ahol a feszültség és nyírás erőssége között (7.2) típusú nemlineáris összefüggés áll fenn.

A lineáris esethez hasonlóan célszerű értelmezni a fenti feladat gyenge alakját a szokásos módon: az egyenletet megszorozzuk egy $v \in H_0^1(\Omega)$ tesztfüggvénnyel, majd integrálunk. Egy $u \in H_0^1(\Omega)$ függvényt tehát a (7.1) feladat gyenge megoldásának nevezünk, ha

$$\int_\Omega a(|\nabla u|^2)\nabla u \cdot \nabla v = \int_\Omega gv \qquad (\forall\ v \in H_0^1(\Omega)). \tag{7.4}$$

A bal oldali integrál értelmes $H_0^1(\Omega)$-n az a függvény korlátossága miatt. A gyenge megoldás létezése a lineáris esetben azon múlt, hogy a bal oldalon szereplő formula egy bilineáris formát definiált, most a nemlineáris esetben ez nem járható. Ehelyett a (7.4) egyenletet alkalmas

$$A(u) = b$$

operátoregyenletnek szeretnénk tekinteni, amit a tesztfüggvények jelenléte miatt a vele ekvivalens

$$\langle A(u), v\rangle_{H_0^1} = \langle b, v\rangle_{H_0^1} \qquad (\forall\ v \in H_0^1(\Omega)) \tag{7.5}$$

„gyenge” alakban célszerű felírni. Ezt a jobb oldalra tudjuk a Riesz-féle reprezentációs tétel alapján: létezik olyan $b \in H_0^1(\Omega)$, hogy

$$\int_\Omega gv = \langle b, v\rangle_{H_0^1} \qquad (\forall\ v \in H_0^1(\Omega)).$$

Szintén a Riesz-tételből igazolható [16, 11.1. állítás], hogy létezik olyan $A : H_0^1(\Omega) \to H_0^1(\Omega)$ operátor, melyre

$$\langle A(u), v\rangle_{H_0^1} = \int_\Omega a(|\nabla u|^2)\nabla u \cdot \nabla v \qquad (\forall\ v \in H_0^1(\Omega)) \tag{7.6}$$

így (7.4) valóban ekvivalens az $A(u) = b$ operátoregyenlettel a $H_0^1(\Omega)$ térben.

Főrészében nemlineáris divergencia-alakú feladatok. A fenti (7.1) helyett érdemes egy kicsit általánosabb szerkezetű feladatosztállyal foglalkoznunk, a fejezetben erre részletezzük majd a továbbiakat. Tekintsük a

$$\begin{cases} -\operatorname{div} f(x, \nabla u) = g, \\ u_{|\partial\Omega} = 0 \end{cases} \tag{7.7}$$

peremérték-feladatot, ahol $f : \Omega \times \mathbb{R}^n \to \mathbb{R}^n$ adott vektorértékű függvény. (Megjegyezzük, hogy itt a szokásos $\operatorname{div} f(x, \nabla u)$ jelölés a precízebb $\operatorname{div}\big(f \circ (id, \nabla u)\big)$ helyett áll, ill. hogy az $f(x, \eta) := a(|\eta|^2)\eta$ esetben visszakapjuk a (7.1) szerkezetű egyenletet.) A feladat gyenge alakja

$$\int_\Omega f(x, \nabla u) \cdot \nabla v = \int_\Omega gv \qquad (\forall\ v \in H_0^1(\Omega)). \tag{7.8}$$

Az f nemlinearitást az alábbi feltételekkel vizsgáljuk:

7.1. Feltevés.

(i) f egyenletesen monoton, azaz létezik $m > 0$, hogy

$$\big(f(x,\eta) - f(x,\tilde{\eta})\big) \cdot (\eta - \tilde{\eta}) \geq m\,|\eta - \tilde{\eta}|^2 \qquad (\forall x \in \Omega,\ \eta, \tilde{\eta} \in \mathbb{R}^n);$$

(ii) f Lipschitz-folytonos, azaz létezik $M > 0$, hogy

$$|f(x,\eta) - f(x,\tilde{\eta})| \leq M\,|\eta - \tilde{\eta}| \qquad (\forall x \in \Omega,\ \eta, \tilde{\eta} \in \mathbb{R}^n);, \qquad \diamondsuit$$

A gyenge alak most is (7.5) típusú feladatként írható, ahol az $A : H_0^1(\Omega) \to H_0^1(\Omega)$ operátor most

$$\langle A(u), v\rangle_{H_0^1} = \int_\Omega f(x, \nabla u) \cdot \nabla v \qquad (\forall\ v \in H_0^1(\Omega)). \tag{7.9}$$

Felmerülhet a kérdés, miért a gyenge alakot írjuk $A(u) = b$ operátoregyenletként az eredeti (7.7) egyenlet helyett. Utóbbiban a differenciáloperátor nem folytonos, míg a gyenge alakban (mint látni fogjuk) A Lipschitz-folytonos, ami mind a megoldhatóság, mind a végeselemes megoldás szempontjából lényeges lesz.

Szemilineáris konvekció-reakció-diffúziós feladatok. Legyen $\Omega \subset \mathbb{R}^2$ korlátos síkbeli tartomány, és tekintsük az alábbi feladatot:

$$\begin{cases} -\operatorname{div}(p\,\nabla u) + \mathbf{w} \cdot \nabla u + q(x,u) = g, \\ u_{|\partial\Omega} = 0. \end{cases} \tag{7.10}$$

7.2. Feltevés.

(i) $p \in L^\infty(\Omega)$, $p(x) \geq m > 0$ (m. m. $x \in \Omega$);

(ii) $\mathbf{w} \in C^1(\overline{\Omega}, \mathbb{R}^2)$, $\operatorname{div} \mathbf{w} = 0$ (azaz $\mathbf{w}$ divergenciamentes vektormező);

(iii) $q \in C^1(\overline{\Omega} \times \mathbb{R})$, és léteznek olyan $p \geq 2$, $\alpha, \beta \geq 0$ állandók, hogy

$$0 \leq \frac{\partial q(x,\xi)}{\partial \xi} \leq \alpha + \beta|\xi|^{p-2} \qquad (\forall x \in \Omega,\ \xi \in \mathbb{R}). \tag{7.11}$$

$\diamondsuit$

A (7.10) modellekben általában az operátor három tagja rendre a diffúziót, konvekciót és kémiai reakciót írja le, melyekből, mint itt is, legtöbbször az első kettő lineáris. Itt tehát a korábban vizsgált (3.43) konvekció-diffúziós egyenletet egészítettük ki egy nemlineáris taggal.

A (7.10) feladat gyenge megoldása olyan $u \in H_0^1(\Omega)$ függvény, melyre

$$\int_\Omega \Big(p\ \nabla u \cdot \nabla v + (\mathbf{w} \cdot \nabla u)v + q(x,u)v\Big) = \int_\Omega gv \qquad \forall\ v \in H_0^1(\Omega). \tag{7.12}$$

Ez a fentiekhez hasonlóan $A(u) = b$ operátoregyenletként írható fel $H_0^1(\Omega)$-ban.

7.2. Monoton operátorok, megoldhatóság

A vizsgált nemlineáris elliptikus feladatok megoldhatóságának alapja a monoton operátorok elméletéből ismert alábbi tétel:

7.3. Tétel. *Legyen H valós Hilbert-tér, $A : H \to H$ adott operátor. Tegyük fel, hogy*

(i) A egyenletesen monoton: létezik $m > 0$, hogy

$$\langle A(u) - A(v), u - v\rangle \geq m \|u - v\|^2 \qquad (\forall u, v \in H); \tag{7.13}$$

(ii) A Lipschitz-folytonos: létezik $M > 0$, hogy

$$\|A(u) - A(v)\| \leq M \|u - v\| \qquad (\forall u, v \in H). \tag{7.14}$$

Ekkor bármely $b \in H$ esetén az $A(u) = b$ egyenletnek egyértelműen létezik $u^* \in H$ megoldása.

A bizonyítás a Banach-féle fixponttételre vezethető vissza: lényege, hogy az $A(u) = b$ egyenlet ekvivalens az $u = u - \alpha(A(u) - b) =: G(u)$ egyenlettel, ahol (az egyenletes monotonitásnak köszönhetően) alkalmas $\alpha > 0$ állandó esetén G kontrakció (lásd [16, 13.5. tétel].) Megjegyezzük, hogy a fenti két feltétel a bilineáris formák koercivitásának és korlátosságának nemlineáris megfelelője.

7.4. Lemma. *Ha teljesülnek a 7.1. feltételek, akkor a (7.9)-ben definiált A operátor egyenletesen monoton és Lipschitz-folytonos a $H_0^1(\Omega)$ téren.*

Bizonyítás. A örökli f tulajdonságait. A Lipschitz-folytonosság:

$$\begin{aligned}
&|A(u) - A(v)|_1 = \\
&\quad = \sup_{\|z\|_{H_0^1}=1} \langle A(u) - A(v), z\rangle_{H_0^1} = \sup_{\|z\|_{H_0^1}=1} \int_\Omega (f(x, \nabla u) - f(x, \nabla v)) \cdot \nabla z \leq \\
&\quad \leq \sup_{\|z\|_{H_0^1}=1} \int_\Omega |f(x, \nabla u) - f(x, \nabla v)| \, |\nabla z| \leq \\
&\quad \leq \sup_{\|z\|_{H_0^1}=1} M \int_\Omega |\nabla u - \nabla v| \, |\nabla z| \leq \sup_{\|z\|_{H_0^1}=1} M \|\nabla u - \nabla v\|_0 \|\nabla z\|_0 = \\
&\quad = \sup_{\|z\|_{H_0^1}=1} M|u - v|_1 \, |z|_1 = M|u - v|_1.
\end{aligned}$$

Az egyenletes monotonitás hasonló, lásd 9.37. feladat. □

Ebből az előző pont alapján a 7.3. tételből rögtön adódik:

7.5. Következmény. *Ha teljesülnek a 7.1. feltételek, akkor bármely $g \in L^2(\Omega)$ esetén a (7.7) peremértékfeladatnak egyértelműen létezik $u^* \in H_0^1(\Omega)$ gyenge megoldása.*

A (7.10) szemilineáris feladatra hasonlóan igazolható a megfelelő A operátor egyenletes monotonitása és Lipschitz-folytonossága, és ebből a megoldhatóság [16, 13.13. tétel].

7.3. Végeselemes diszkretizáció

Galjorkin-módszer nemlineáris operátoregyenletekre. Először operátoregyenletekkel foglalkozunk az előző szakasz alapján. Legyen tehát H ismét valós Hilbert-tér, $A : H \to H$ adott lineáris operátor, amely egyenletesen monoton és Lipschitz-folytonos. Tekintsük az

$$A(u) = b \tag{7.15}$$

operátoregyenletet, ahol $b \in H$, ennek a 7.3. tétel szerint egyértelműen létezik $u^* \in H$ megoldása. Írjuk fel ezt a vele ekvivalens tesztfüggvényes alakban:

$$\langle A(u^*), v\rangle = \langle b, v\rangle \qquad (\forall v \in H). \tag{7.16}$$

Legyen $V_h = span\{\varphi_1, \varphi_2, \ldots, \varphi_n\} \subset H$ véges dimenziós altér, ahol $\varphi_1, \varphi_2, \ldots, \varphi_n$ lineárisan függetlenek. Az $u^h \in V_h$ közelítő megoldást az

$$\langle A(u^h), v^h\rangle = \langle b, v^h\rangle \qquad (\forall v^h \in V_h) \tag{7.17}$$

vetületi egyenlet definiálja, létezését és egyértelműségét pedig a 7.3. tétel garantálja a V_h térben. Az

$$u^h = \sum_{i=1}^{n} c_i\varphi_i \tag{7.18}$$

előállítás együtthatóit a következőképp kapjuk. Helyettesítsük a (7.18) alakot és a $v^h := \varphi_k$ függvényeket a (7.17) egyenletbe:

$$\Big\langle A\big(\sum_{i=1}^{n} c_i\varphi_i\big), \varphi_k\Big\rangle = \langle b, \varphi_k\rangle \qquad (k = 1, \ldots, n).$$

Vezessük be az

$$\mathcal{A}_k : \mathbb{R}^n \to \mathbb{R}, \qquad \mathcal{A}_k(c) = \mathcal{A}_k(c_1, \ldots, c_n) := \Big\langle A\big(\sum_{i=1}^{n} c_i\varphi_i\big), \varphi_k\Big\rangle$$

valós függvényeket és legyen $\beta_k := \langle b, \varphi_k\rangle \quad (k = 1, \ldots, n)$. Az ezekből összerakott $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$ függvény és $\beta \in \mathbb{R}^n$ vektor mellett tehát u^h együtthatóit az

$$\mathcal{A}(c) = \beta$$

nemlineáris algebrai egyenletrendszer megoldásával kapjuk.

7.6. Megjegyzés. (A reziduális hiba ortogonalitása.) A (7.16) és (7.17) egyenletekből következik, hogy

$$\langle A(u^*) - A(u^h), v^h\rangle = 0 \qquad (\forall v \in V_h), \tag{7.19}$$

azaz $A(u^*) - A(u^h) \perp V_h$. Jelölje most $r^h := A(u^h) - b$ a reziduális vektort. Mivel $A(u^*) = b$, a fentiekből $\langle r^h, v^h\rangle = 0$ minden $v^h \in V_h$ esetén, azaz r^h ortogonális V_h-re. $\diamondsuit$

A módszer konvergenciája a Céa-lemma (3.7. állítás) megfelelőjére alapul:

7.7. Állítás. ***(„Nemlineáris Céa-lemma")*** *Az $u^h \in V_h$ Galjorkin-megoldásra*

$$\|u^* - u^h\| \leq \frac{M}{m} \min\{\|u^* - v^h\| : \ v^h \in V_h\}.$$

Bizonyítás. A (7.13)–(7.14) és (7.19) összefüggésekből bármely $v^h \in V_h$ esetén

$$\begin{aligned} m\|u^* - u^h\|^2 &\leq \langle A(u^*) - A(u^h), u^* - u^h\rangle = \langle A(u^*) - A(u^h), u^* - v^h\rangle \\ &\leq \|A(u^*) - A(u^h)\|\|u^* - v^h\| \leq M\|u^* - u^h\|\|u^* - v^h\|, \end{aligned}$$

így $\|u^* - u^h\| \leq \frac{M}{m}\|u^* - v^h\|$. □

Innen a folytatás a lineáris esettel azonos: ha egy altércsaládra $dist(u^*, V_h) \to 0$, akkor $\|u^h - u^*\| \to 0$, a konkrét végeselemes megvalósításokban pedig a konvergencia rendje interpolációs becslésekből adódik.

Végeselemes megoldás nemlineáris elliptikus feladatra. Tekintsük a (7.7) feladatot:

$$\begin{cases} -\operatorname{div} f(x, \nabla u) \ = g, \\ \ u_{|\partial\Omega} = 0 \end{cases} \tag{7.20}$$

a 7.1. feltételekkel. Ennek $u^* \in H_0^1(\Omega)$ gyenge megoldására

$$\int_\Omega f(x, \nabla u^*) \cdot \nabla v = \int_\Omega gv \qquad (\forall\ v \in H_0^1(\Omega)). \tag{7.21}$$

Legyen $V_h \subset H_0^1(\Omega)$ valamely végeselemes altér, mint a lineáris esetben a 3.2. szakaszban. Ekkor az $u^h \in V_h$ közelítő megoldás teljesíti az

$$\int_\Omega f(x, \nabla u^h) \cdot \nabla v^h = \int_\Omega gv^h \qquad (\forall\ v^h \in V_h)$$

vetületi egyenletet, és u^h együtthatóit az

$$\mathcal{A}(c) = \beta$$

nemlineáris algebrai egyenletrendszer megoldásával kapjuk, ahol

$$\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n, \qquad \mathcal{A}_k(c) = \mathcal{A}_k(c_1, \ldots, c_n) = \int_\Omega f(x, \sum_{i=1}^n c_i \nabla\varphi_i) \cdot \nabla\varphi_k \qquad (k = 1, \ldots, n)$$

és $\beta_k := \int_\Omega g\varphi_k \quad (k = 1, \dots, n)$. Itt $\mathcal{A}$ örökli A egyenletes monotonitását és Lipschitz-folytonosságát V_h-ban, így az $\mathcal{A}(c) = \beta$ rendszerre is igaz a 7.5. következmény szerinti megoldhatóság. Emellett igaz a nemlineáris Céa-lemma:

$$|u^* - u^h|_1 \leq \frac{M}{m} \min\{|u^* - v^h|_1 : \ v^h \in V_h\},$$

amiből a lineáris esettel azonos módon haladhatunk tovább: először v^h helyére u^* interpoláltját írjuk, így (3.27) mintájára

$$|u^* - u^h|_1 \leq \frac{M}{m} |u^* - \Pi_h u^*|_1 ,$$

majd alkalmazzuk a 3.37. állítást, és a 3.40. tétellel azonosan kapjuk a konvergencia-becsléseket:

7.8. Tétel. *Legyen $k \in \mathbb{N}^+$, és*

(i) a (7.21) feladat megoldására $u^* \in H^{k+1}(\Omega)$;

(ii) az FEM-ben használt háromszögű triangulációreguláris;

(iii) a polinomokra érvényes $P(T) \supset P^k(T) \quad (\forall T \in \mathcal{T}_h, \forall \mathcal{T}_h \in \mathcal{F})$.

Ekkor

$$|u^* - u^h|_1 \leq c\, h^k\, |u^*|_{k+1},$$

ahol $c > 0$ független a triangulációtól.

Végül a (7.10) szemilineáris feladatra értelemszerű módosításokkal írható fel a végeselemes megoldás konstrukciója. A megfelelő A operátor egyenletes monotonitása és Lipschitz-folytonossága révén ugyanúgy igaz a nemlineáris Céa-lemma és ennek köszönhetően a 7.8. tételbeli becslések.

7.4. Newton-típusú iterációk

A végeselemes diszkretizációval az eredeti feladatot közelítőleg a

$$\langle A(u^h), v^h \rangle = \langle b, v^h \rangle \qquad (\forall v^h \in V_h) \tag{7.22}$$

vetületi egyenletre, ill. az ennek megfelelő $\mathcal{A}(c) = \beta$ nemlineáris algebrai egyenletrendszerre vezettük vissza. Ennek megoldása iteráció segítségével lehetséges. Itt röviden felidézzük a nemlineáris egyenletrendszerek leghatékonyabb és emiatt legelterjedtebb megoldási módszerét, a Newton-iterációt.

A Newton-módszer Kantorovics munkássága révén általános Banach-térbeli operátor-egyenletekre is kiterjed, és leírása azonosan megy a véges dimenziós esettel, ezért ebben a formában idézzük fel. Követjük [16, 18.2. fejezet] leírását. Legyenek tehát most X, Y Banach-terek. A Newton-módszerben hagyományosan zérushelyet keresünk (ez most az Y tér 0-vektora), azaz egy

$$F(u) = 0$$

egyenlet megoldását. Ez nem megszorítás, hiszen az eredeti $A(u) = b$ egyenlet $F(u) := A(u) - b$ mellett ilyen alakú lesz.

Az előzőek alapján olyan egyenletekkel foglalkozunk, ahol garantálható az egyértelmű megoldás. Az erre vonatkozó elég általános feltétel, hogy bármely $u, h \in X$ esetén

$$F'(u) : X \to Y \text{ bijekció és } \quad \|F'(u)h\| \geq m\|h\|, \tag{7.23}$$

ahol $m > 0$ független u, h-tól.

7.9. Tétel. *Legyenek X, Y Banach-terek és $F : X \to Y$ Fréchet-deriválható. Tegyük fel, hogy teljesül (7.23), valamint F' Lipschitz-folytonos L konstanssal. Ha $u_0 \in X$ tetszőleges, akkor az*

$$u_{n+1} = u_n - F'(u_n)^{-1}F(u_n) \qquad (n \in \mathbb{N})$$

iterációra az alábbiak teljesülnek:

(1) $\|F(u_{n+1})\| \leq \dfrac{L}{2m^2} \|F(u_n)\|^2 \quad (n \in \mathbb{N}).$

(2) Ha u_0 olyan, hogy

$$q := \frac{L}{2m^2} \|F(u_0)\| < 1, \tag{7.24}$$

akkor

$$m\|u_n - u^*\| \leq \|F(u_n)\| \leq \frac{2m^2}{L} q^{2^n} \to 0. \tag{7.25}$$

A bizonyításhoz lásd [16] 18.3. tételét.

7.10. Megjegyzés. Hilbert-térben a (7.23) feltétel garantálható az

$$\langle F'(u)h, h\rangle \geq m \|h\|^2 \qquad (\forall u, h \in H)$$

egyenlőtlenséggel, ekkor F egyenletesen monoton. Ha az $F(u) = 0$ egyenlet megoldhatóságához még feltesszük F Lipschitz-folytonosságát és a 7.3. tételt használjuk, akkor itt és a 7.9. tétel bizonyításában is elég a Gâteaux-deriválhatóság. ♢

A fenti Newton-módszer gyengéje, hogy csak lokálisan konvergál. Ezt orvosolja a *csillapított Newton-módszer*, lásd pl. [16, 18.11. tétel]:

7.11. Tétel. *Teljesüljenek a 7.9. tétel feltételei és legyen $u^* \in X$ az $F(u) = 0$ egyenlet megoldása. Legyen $u_0 \in X$ tetszőleges, és tekintsük az alábbi sorozatot:*

$$\begin{cases} u_{n+1} := u_n + \tau_n p_n \qquad (n \in \mathbb{N}), \quad \textit{ahol} \\ F'(u_n)p_n = -F(u_n) \qquad \textit{és} \quad \tau_n = \min\left\{1,\ \dfrac{m^2}{L\|F(u_n)\|}\right\}. \end{cases} \tag{7.26}$$

Ekkor

$$\|u_n - u^*\| \le \tfrac{1}{m}\|F(u_n)\| \ \to 0$$

monoton csökkenően és lokálisan másodrendben, azaz alkalmas $n_0 \in \mathbb{N}$ index után

$$\|F(u_{n+1})\| \le c_1\|F(u_n)\|^2 \qquad (n \ge n_0) \tag{7.27}$$

$$\textit{és} \qquad \|u_n - u^*\| \le \tfrac{1}{m}\|F(u_n)\| \le d_1\, q^{2^n} \qquad (n \ge n_0) \tag{7.28}$$

(ahol $0 < q < 1$, $c_1, d_1 > 0$).

7.12. Megjegyzés. A megadott iterációs lépés az alábbi alakba írható át:

$$\begin{cases} F'(u_n)p_n &= -F(u_n), \\ u_{n+1} &= u_n + \tau_n p_n\,, \end{cases}$$

sőt ez az, amit valójában használunk: nem kell meghatározni $F'(u_n)$ inverzét, hanem a megfelelő segédfeladatot kell megoldani p_n kiszámításához. ♢

Tekintsük most a (7.7) feladat végeselemes megoldását, melyre a Newton-módszert alkalmazni kívánjuk:

$$\int_\Omega f(x, \nabla u^h)\cdot\nabla v^h = \int_\Omega g v^h \qquad (\forall\ v^h \in V_h)$$

Írjuk ezt 0-ra rendezve:

$$\langle F_h(u^h), v^h\rangle_{H_0^1} := \int_\Omega \big(f(x, \nabla u^h)\cdot\nabla v^h - g v^h\big) = 0 \qquad (\forall v^h \in V_h). \tag{7.29}$$

A Newton-módszerben szereplő deriválhatósági feltételek miatt f-re is új feltételt kell tennünk:

7.13. Feltevés. $f \in C^1(\overline{\Omega}\times\mathbb{R}^N, \mathbb{R}^N)$, a $\frac{\partial f(x,\eta)}{\partial \eta}$ Jacobi-mátrixok szimmetrikusak, és létezik $M \ge m > 0$, hogy

$$m|\xi|^2 \le \frac{\partial f(x,\eta)}{\partial \eta}\,\xi\cdot\xi \le M|\xi|^2 \qquad (x \in \Omega,\ \xi,\eta \in \mathbb{R}^n), \tag{7.30}$$

valamint $\eta \mapsto \frac{\partial f(x,\eta)}{\partial \eta}$ Lipschitz-folytonos. ♢

7.14. Megjegyzés. A (7.30) becslés azt jelenti, hogy a Jacobi-mátrixok sajátértékei m és M közt vannak. Ez a feltétel emellett felülírja a 7.1. feltételeket abban az értelemben, hogy az ottani egyenletes monotonitás és Lipschitz-folytonosság rendre következik a (7.30)-beli alsó és felső becslésből. ♢

Igazolható [16, 11.2. szakasz] mintájára, hogy ekkor a (7.29)-ben bevezetett $F_h : V_h \to V_h$ operátor örökli f tulajdonságait: F_h Gâteaux-deriválható és

$$\langle F_h'(u^h)z^h, v^h\rangle_{H_0^1} = \int_\Omega \frac{\partial f}{\partial \eta}(x, \nabla u^h)\nabla z^h \cdot \nabla v^h \qquad (\forall u^h, z^h, v^h \in V_h),$$

amiből (7.30) miatt rögtön látható, hogy

$$m\,|v^h|_1^2 \le \langle F_h'(u^h)v^h, v^h\rangle_{H_0^1} \le M\,|v^h|_1^2 \qquad (\forall u^h, v^h \in V_h),$$

valamint F_h' Lipschitz-folytonos alkalmas L_h konstanssal. Így érvényes a 7.11. tétel a (7.29) feladatra.

Írjuk fel a 7.12. megjegyzés alapján az iteráció algoritmusát! Ha megvan u_n^h, akkor gyenge alakot használva az alábbi segédfeladatot kell megoldanunk:

$$\langle F'(u_n^h)p_n^h, v^h\rangle_{H_0^1} = -\langle F(u_n^h), v^h\rangle_{H_0^1} \qquad (\forall v^h \in V_h),$$

és az ebből kapott p_n^h-nel javítunk:

$$u_{n+1}^h := u_n^h + \tau_n p_n^h.$$

Az algoritmus tehát az alábbi, ahol egyszerűség kedvéért most elhagyjuk a felső h indexet:

$$\begin{cases} (a)\ u_0 \in V_h \text{ tetszőleges;} \\ \quad \text{bármely } n \in \mathbb{N} \text{ esetén: ha megvan } u_n \in V_h, \text{ akkor} \\ (b1)\ p_n \in V_h \text{ az alábbi feladat megoldása:} \\ \quad \displaystyle\int_\Omega \frac{\partial f}{\partial \eta}(x, \nabla u_n)\,\nabla p_n \cdot \nabla v = -\int_\Omega \big(f(x, \nabla u_n)\cdot \nabla v - gv\,\big) \qquad (v \in V_h); \\ (b2)\ \tau_n := \min\{\,1,\ \dfrac{\mu_1}{L|p_n|_1}\} \in (0,1]\,, \\ (b3)\ u_{n+1} := u_n + \tau_n p_n\,. \end{cases} \tag{7.31}$$

A segédfeladatok lineáris elliptikus feladatok végeselemes diszkretizáltjai, melyek a korábban látott módszerek valamelyikével megoldhatók. Az iteráció konvergenciáját a 7.11. tétel írja le.

Hasonlóan építhető fel a csillapított Newton-iteráció a (7.10) szemilineáris konvekció-reakció-diffúziós feladat végeselemes diszkretizációjára, lásd 9.38. feladat.

7.15. Példa. Tekintsük a (7.1) feladatot:

$$\begin{cases} -\operatorname{div}\Big(a(|\nabla u|^2)\,\nabla u\Big) = g, \\ u_{|\partial\Omega} = 0, \end{cases} \tag{7.32}$$

ahol legyen $a : \mathbb{R}^+ \to \mathbb{R}^+$ adott C^2-beli függvény, és tegyük fel, hogy léteznek olyan $M \geq m > 0$ konstansok, hogy

$$0 < m \leq a(r^2) \leq \big(a(r^2)r\big)' \leq M \qquad (\forall r \geq 0), \tag{7.33}$$

valamint $\big(a(r^2)r\big)''$ korlátos. Ez a feladat olyan alakú, mint (7.7), ha

$$f(x,\eta) = a\big(|\eta|^2\big)\eta \qquad (x \in \overline{\Omega},\ \eta \in \mathbb{R}^N),$$

sőt, itt f független x-től. A fenti példa általánosítható úgy, hogy f függhet x-től is:

$$f(x,\eta) = a\big(x, |\eta|^2\big)\eta, \qquad \text{ahol} \qquad 0 < m \leq a(x,r^2) \leq \frac{\partial}{\partial r}\Big(a(x,r^2)r\Big) \leq M$$

$(\forall x \in \overline{\Omega},\ \eta \in \mathbb{R}^N,\ r \geq 0)$.

Könnyen látható, hogy ekkor teljesülnek a 7.13. feltételek (lásd [16, 13.4.1. szakasz]). Így a (7.32) feladat végeselemes diszkretizáltjára alkalmazott (7.31) iteráció a 7.11. tétel szerint konvergál. ♢

8. fejezet

Számítógépes alkalmazások

8.1. Programok

Ebben a fejezetben bemutatjuk, hogy a korábban részletesen vizsgált eljárások hogyan valósíthatók meg a gyakorlatban. Az eljárások megírására a MATLAB programcsomagot választottuk. A MATLAB programozási nyelvének egyszerűsége és áttekinthetősége miatt az eljárások könnyen átültethetők más programozási nyelvekre is. Azon olvasóknak, akik még nem használták a MATLAB-ot, e fejezet elolvasása előtt ajánljuk áttanulmányozásra a program honlapját vagy pl. az alábbi magyar nyelvű rövid leírást a http://www.mit.bme.hu/oktatas/targyak/vimia305/matlab címen. Megjegyezzük, hogy bár mi a továbbiakban részletes MATLAB kódokat fogunk megadni, a MATLAB rendelkezik egy pdetools nevű alkalmazással, melyben egy grafikus felhasználói felületen adhatjuk meg a megoldandó differenciálegyenletet, a megoldási tartományt és a kezdő- és peremfeltételeket. Ezek után a feladat megoldásához és a megoldás ábrázolásához szinte egy gombnyomás elegendő.

8.1.1. Hasznos MATLAB parancsok

Először áttekintjük azokat a hasznos MATLAB parancsokat, melyekre gyakran szükségünk lehet parciális differenciálegyenletek numerikus megoldása során.

A korábbi fejezetekben láttuk, hogy a numerikus megoldások során mindig ritka mátrixokkal kell dogoznunk. Ezeket a mátrixokat érdemes a gyakorlatban is ritka mátrixként definiálni, hiszen ezzel memóriát takaríthatunk meg és a mátrixműveletek is gyorsabban végrehajthatók. Ha egy mátrixot ritka mátrixként definiálunk, akkor a MATLAB csak a nemnulla elemeket, és azok helyét tárolja el, ellentétben egy teljes mátrixszal, amikor a mátrix minden elemét tároljuk. Az $n \times n$-es egységmátrix pl. ritka mátrixként az

```
I=speye(n)
```

módon definiálható, amely $n = 1000$ esetén csak 16kb tárhelyet foglal (teljes mátrixként 8Mb-ot). Egy nullákat tartalmazó B mátrixot a

```
A=sparse(B)
```

paranccsal alakíthatunk ritka mátrixszá. A visszaalakítás a

```
B=full(A)
```

paranccsal történik. Egy A mátrix nemnulla elemeinek szerkezetét a

```
spy(A)
```

paranccsal tudjuk ábrán szemléltetni. Az ábra alján nz a nemnulla elemek számát jelöli.

Ritka mátrixokat természetesen érdemes közvetlenül definiálni. Ezt megtehetjük pl. az

```
A=sparse(sorindexek,oszlopindexek,elemek,m,n)
```

paranccsal, ahol a nemnulla elemek az *elemek* vektorban vannak felsorolva, a sor és oszlopindexeiket a *sorindexek* és *oszlopindexek* vektorok adják meg, és a mátrix $m \times n$-es. Hasonlóan használható még ritka mátrixok megadására az

```
A=spdiags(B,index,m,n)
```

parancs is, ahol a B mátrix oszlopai rendre az *index* vektorban megadott sorszámú diagonálisba kerülnek (0. a főátló, -1. a subdiagonál, 1. a szuperdiagonál, stb.) úgy, hogy a keletkező mátrix $m \times n$ méretű legyen (a kilógó elemeket nem vesszük figyelembe).

Blokkmátrixok konstrukciójánál jól használható még a

```
kron(X,Y)
```

Kronecker-féle szorzat, melynek blokkjai rendre az X mátrix elemeinek az Y mátrixszal alkotott szorzatai.

A fenti parancsokkal pl. a kétdimenziós Laplace-operátor az alábbi A mátrixszal diszkretizálható az egységnégyzeten homogén Dirichlet peremfeltétel esetén egy $n \times n$-es négyzetrácson:

```
I = speye(n);
E = sparse(2:n,1:n-1,1,n,n);
D = E+E'-2*I;
A = kron(D,I)+kron(I,D);
```

A mátrixok méretének megváltoztatása a

```
B = reshape(A,m,n)
```

paranccsal történik. Ekkor a B mátrix egy $m \times n$-es mátrix lesz, melybe oszloponként haladva kerülnek bele A elemei.

Parciális differenciálegyenletek numerikus megoldása során sokszor nagy méretű lineáris egyenletrendszereket kell megoldanunk. Ezek megoldását sosem az együtthatómátrix inverzének kiszámításával határozzuk meg, mert ennél a módszernél vannak sokkal gyorsabb és pontosabb eljárások is. Általában speciálisan ritka mátrixokra kifejlesztett direkt megoldási eljárásokat vagy iterációs módszereket érdemes használni (lásd pl. [11]). Ezek közül érdemes kiemelni a Thomas-algoritmust, amely a Gauss-módszer speciális változata tridiagonális együtthatómátrixú egyenletrendszerek megoldására. A MATLAB beépített egyenletrendszer-megoldó parancsa

```
A\b
```

amely mindig az egyenletrendszer típusához leginkább megfelelő direkt módszerrel oldja meg a feladatot.

8.1.2. Az eredmények szemléltetése

A feladatok numerikus megoldását követően fontos feladat a megoldás vizualizációja. Ennek segítségével szemléletes képet kaphatunk a megoldásfüggvény viselkedéséről, de akár segíthet a programozási hibák felfedezésében is.

A MATLAB grafikai függvényei kihasználják, hogy a numerikus eljárások során általában a megoldás rácspontokbeli függvényértékekként adott.

Így pl. egy egyváltozós függvényt, amely az *xvektor* vektorban megadott pontokban rendre az *yvektor* vektor elemeinek értékeit veszi fel a

```
plot(xvektor,yvektor)
```

módon ábrázolhatjuk. Itt harmadik argumentumként megadhatjuk a grafikon színét, vonaltípusát és a pontokat jelző szimbólumokat is. Lásd részletesen a plot parancs leírását a MATLAB honlapján.

Kétváltozós függvények ábrázolása téglalaprácson hasonlóan történik. Először megadjuk a téglalap x- és y-tengelyekkel párhuzamos oldalainak felosztásait.

```
x=xmin:xlépés:xmax;
y=ymin:ylépés:ymax;
```

majd ezen felosztásokból megkonstruáljuk a téglalaprács pontjainak x és y koordinátáit tartalmazó mátrixokat.

```
[xi,yi]=meshgrid(x,y);
```

Ezek után, ha a rácspontokbeli értékeket a zi mátrixban tároljuk, akkor az ábrázolás a

```
mesh(xi,yi,zi);
surf(xi,yi,zi);
trisurf(tri,xi,yi,zi)
```

parancsokkal történhet. Az első két esetben négyszögrácson ábrázolunk drótháló vagy kitöltött grafikonokkal, a harmadikban pedig a *tri* mátrixszal adott háromszögrácson kitöltött grafikonnal.

Mi a véges differenciák módszerek esetén az első, a végeselem-módszerek esetén a harmadik módot használtuk.

8.1.3. Két mintaprogram

Példaként megadjuk a numerikus megoldást előállító programokat a véges differenciás és a végeselem-módszerek esetén egy egyszerű peremérték-feladatra.

A $-\Delta u = f$ Poisson-egyenletet oldjuk meg az egységnégyzeten. Az $y = 0$ peremen a peremfeltétel $g(x) := u(x, 0) = \sin(\pi x)$, a többi élen homogén Dirichlet-peremfeltétel adott, továbbá $f(x, y) = \sin(\pi x)(2 + (1 - y^2)\pi^2)$. A feladat pontos megoldása $u(x, y) = (1 - y^2)\sin(\pi x)$.

8.1. Példa. A véges differenciák módszerével történő megoldás az alábbi kóddal nyerhető. Itt n a belső osztópontok száma (ugyanannyi x és y irányban is), g a peremfeltételt megadó függvény, f pedig a forrás.

A program futása után a megoldásfüggvény közelítő értékei az *ugrid* mátrixban találhatók. Az elkészült 8.1. ábra pedig az együtthatómátrix szerkezetét és a numerikus megoldást mutatják.

```
clear all; close all; % minden korábbi adat törlése

% Input adatok megadása

n=5; h=1/(n+1); % rácspontok száma x és y irányban, rácstávolság
x=h*[1:n]; y=x; % oldalak ekvidisztáns felosztása
[xi,yi]=meshgrid(x,y); % a rácspontok koordinátáinak kiszámítása
g = sin(pi*x); % perem y=0-nál a sin(pi*x) peremfeltétel esetén
f =(2+pi^2*(1-yi.^2)).*sin(pi*xi); % forrás (2+pi^2(1-y^2))*sin(pi*x)

% A diszkretizációs mátrix konstrukciója

I = speye(n);
E = sparse(2:n,1:n-1,1,n,n);
D = E+E'-2*I;
A = -kron(D,I)-kron(I,D);
```

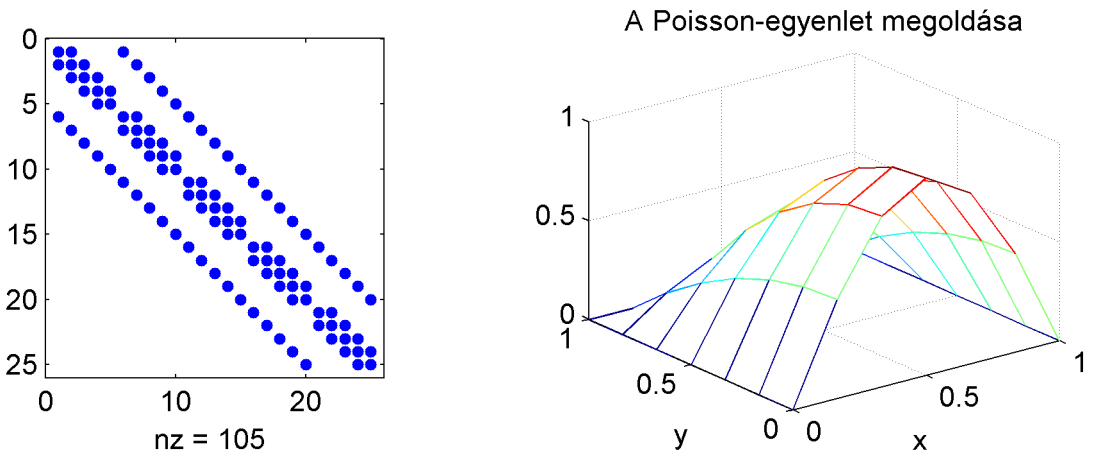
```
A = A/h^2; % Laplace-operátor (-1)-szeresének diszkr.
spy(A); pause % A mátrix szerkezetének kirajzoltatása

% A jobboldali vektor konstrukciója

b=reshape(f,n^2,1); % a forrás miatt
b(1:n:n^2)=b(1:n:n^2)+g'/h^2; % a peremfeltétel módosítja b-t

% Numerikus megoldás kiszámítása az egyenletrendszer megoldásával

uapprox=A\b;

% A numerikus megoldás ábrázolása

ugridin = reshape(uapprox,n,n); % megoldásvektor mátrixszá alakítása
ugrid = zeros(n+2,n+2); % a peremfeltételek figyelembevétele
ugrid(2:n+1,2:n+1)=ugridin;
ugrid(1,2:n+1)=g;

mesh(0:h:(n+1)*h',0:h:(n+1)*h',ugrid);
axis([0,1,0,1,0,1])
xlabel('x','FontSize',14)
ylabel('y','FontSize',14)
title('A Poisson-egyenlet megoldása.','FontSize',14)
```

8.2. Példa. A végeselem-módszer esetén szintén n jelöli a belső osztópontok számát, g a peremfeltétel és f a forrás. A rács egy négyzetrács négyzeteinek ugyanolyan irányú átlókkal való darabolásából származó háromszögrács. Az *ugrid* mátrix tartalmazza a megoldásfüggvény értékeit a rácspontokban.

Figyeljük meg, hogy mennyire hasonló az egyenletrendszer konstrukciója a véges differenciás esethez. A tulajdonképpeni különbség az, hogy a h^2 szorzó nem a mátrixban oszt, hanem a jobb oldalon szoroz.

A program futása után a 8.2 ábrát kapjuk.

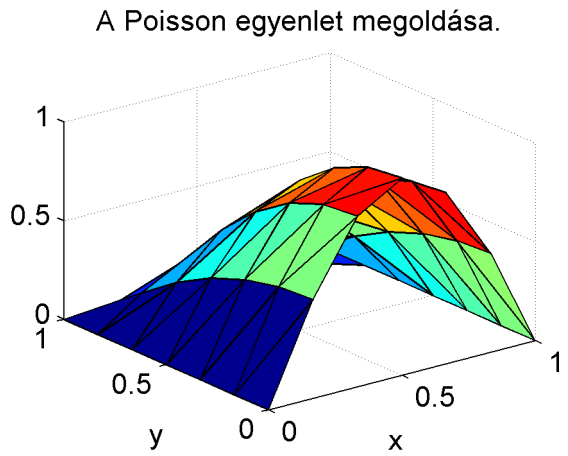
```
clear all; close all; % minden korábbi adat törlése

% Input adatok megadása

n=5; h=1/(n+1); % osztópontok száma x és y irányban, rácstávolság
x=h*[1:n]; y=x; % oldalak ekvidisztáns felosztása
```



8.1. ábra. Az együtthatómátrix és a numerikus megoldás szemléltetése a véges differenciák módszere esetén.

```
[xi,yi]=meshgrid(x,y); % a rácspontok koordinátáinak kiszámítása
g = sin(pi*x); % perem y=0-nál a sin(pi*x) peremfeltétel esetén
f =(2+pi^2*(1-yi.^2)).*sin(pi*xi); % forrás (2+pi^2(1-y^2))*sin(pi*x)

% A merevségi mátrix konstrukciója

I = speye(n);
E = sparse(2:n,1:n-1,1,n,n);
D = E+E'-2*I;
A = -kron(D,I)-kron(I,D);

% A terhelési vektor konstrukciója

b=reshape(f,n^2,1)*h^2; % a forrás miatt (a bázisfüggvények és f
                        % szorzatainak integráljának közelítésére
                        % a 3h^2*f/3=h^2*f képletet
                        % használjuk)

b(1:n:n^2)=b(1:n:n^2)+g'; % a peremfeltétel módosítja b-t

% Az egyenletrendszer megoldása

uapprox=A\b;
```

```
% A numerikus megoldás ábrázolása

ugridin = reshape(uapprox,n,n); % megoldásvektor mátrixszá alakítása
ugrid = zeros(n+2,n+2); % a peremfeltételek figyelembevétele
ugrid(2:n+1,2:n+1)=ugridin;
ugrid(1,2:n+1)=g;

tri=[0 0 0]; % A háromszögek konstrukciója a hozzájuk tartozó csúcsokból.
for i=1:(n+1)*(n+2)
    if mod(i,n+2)~=0
        tri(length(tri(:,1))+1,:)=[i,i+1,i+n+2];
        tri(length(tri(:,1))+1,:)=[i+1,i+1+n+2,i+n+2];
    end
end

tri(1,:)=[];
[xii,yii]=meshgrid(0:h:1,0:h:1);
trisurf(tri,xii,yii,ugrid)
axis([0,1,0,1,0,1])
xlabel('x','FontSize',14)
ylabel('y','FontSize',14)
title(['A Poisson egyenlet megoldása.'],'FontSize',14)
```

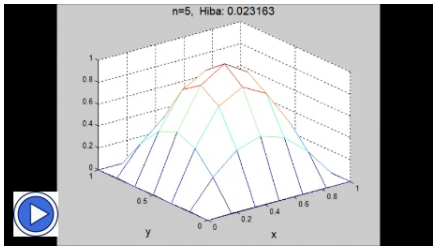


8.2. ábra. A numerikus megoldás szemléltetése a végeselem-módszer esetén.

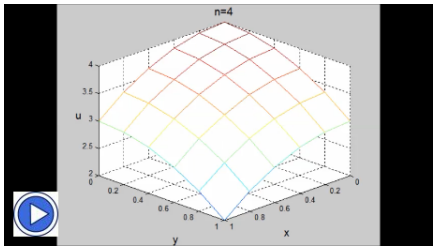
8.2. Animációk

8.2.1. A Poisson-egyenlet numerikus megoldása a véges differenciák módszerével

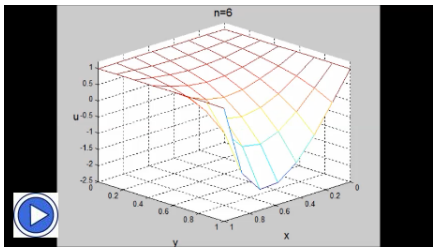
8.2.1 Animáció. Tekintsük a $-\Delta u = 2\pi^2 \sin(\pi x)\sin(\pi y)$ Poisson-egyenletet az egységnégyzeten homogén Dirichlet-peremfeltétellel! Könnyen látható, hogy a feladat pontos megoldása $u(x,y) = \sin(\pi x)\sin(\pi y)$. Az alábbi animáció a véges differenciás numerikus megoldásokat és a rácspontokbeli legnagyobb eltérést mutatja feleződő rácstávolság esetén. n a belső rácspontok száma x és y irányban. Figyeljük meg, hogy feleződő rácstávolság esetén a hiba kb. negyedelődik, ami mutatja az elméletben igazolt másodrendű konvergenciát!



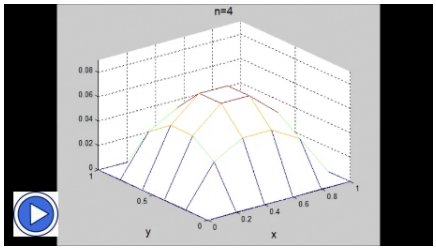
8.2.2 Animáció. Tekintsük a $-\Delta u = 4$ Poisson-egyenletet az egységnégyzeten az alábbi peremfeltételekkel: $u(x,0) = 4 - x^2$, $u(0,y) = 4 - y^2$, $u(1,y) = 3 - y^2$ (Dirichlet-peremfeltételek), $\partial u/\partial n(x,1) = -2$ (Neumann-peremfeltétel)! Az alábbi animáció a feladat véges differenciás numerikus megoldásait mutatja azokban az esetekben, amikor az egységnégyzeten x és y irányban is ekvidisztáns módon rendre $n = 2, 4, \ldots, 60$ belső rácspontot veszünk fel. A feladat pontos megoldása $u(x,y) = 4 - x^2 - y^2$.



8.2.3 Animáció. Tekintsük a $-\Delta u = y$ Poisson-egyenletet az egységnégyzeten az alábbi peremfeltételekkel: $u(x,0) = 1$, $u(0,y) = 1$, $u(1,y) = 1$ (Dirichlet-peremfeltételek), $\partial u/\partial n(x,1) = -15x$ (Neumann-peremfeltétel)! Az alábbi animáció a feladat véges differenciás numerikus megoldásait mutatja azokban az esetekben, amikor az egységnégyzeten x és y irányban is ekvidisztáns módon rendre $n = 2, 4, \ldots, 60$ belső rácspontot veszünk fel.

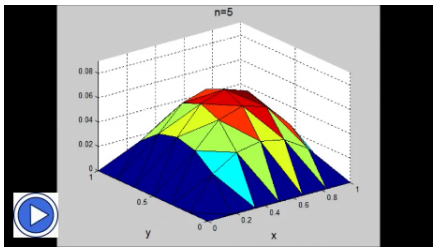


8.2.4 Animáció. Tekintsük a $-\Delta u = 1$ Poisson-egyenletet az egységnégyzeten homogén Dirichlet-peremfeltétellel! Az alábbi animáció a feladat véges differenciás numerikus megoldásait mutatja azokban az esetekben, amikor az egységnégyzeten x és y irányban is ekvidisztáns módon rendre $n = 2, 3, \ldots, 30$ belső rácspontot veszünk fel.

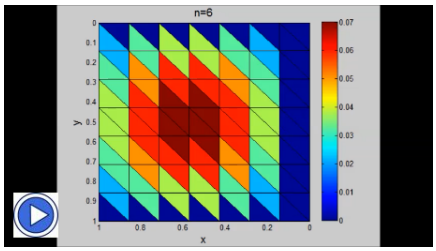


8.2.2. A Poisson-egyenlet numerikus megoldása a végeselem-módszerrel

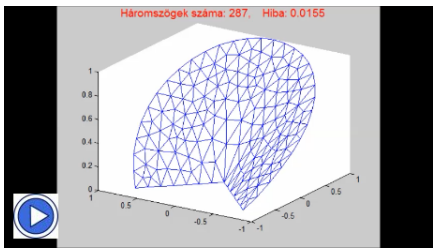
8.2.5 Animáció. Tekintsük a $-\Delta u = 1$ Poisson-egyenletet az egységnégyzeten homogén Dirichlet-peremfeltétellel! Az alábbi animáció a végeselem-módszerrel nyert numerikus megoldásokat mutatja egyre finomodó egyenletes háromszögrácsokon azokban az esetekben, amikor az egységnégyzet oldalain ekvidisztáns módon rendre $n = 2, 3, \ldots, 30$ belső rácspontot veszünk fel.



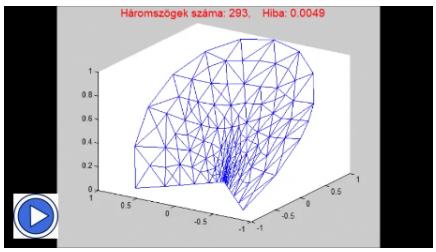
8.2.6 Animáció. Tekintsük a $-\Delta u = 1$ Poisson-egyenletet az egységnégyzeten homogén Dirichlet-peremfeltétellel! Az alábbi animáció a végeselem-módszerrel nyert numerikus megoldásokat mutatja a z-tengely irányából egyre finomodó egyenletes háromszögrácsokon azokban az esetekben, amikor az egységnégyzet oldalain ekvidisztáns módon rendre $n = 2, 3, \ldots, 30$ belső rácspontot veszünk fel.



8.2.7 Animáció. Tekintsük a $\Delta u = 0$ Laplace-egyenletet az $r \in [0,1]$, $\phi \in [-3\pi/4, 3\pi/4]$ polárkoordinátás paraméterezésű egységsugarú körcikken az alábbi peremfeltételekkel: a köríven $u(x,y) = \cos(2 \operatorname{arctg} 2(y,x)/3)$, az egyenes szakaszokon pedig $u(x,y) = 0$! A lenti animáció a végeselem-módszerrel nyert numerikus megoldásokat mutatja a háromszögrács egyenletes finomítása esetén. Felül a háromszögek száma és a maximumnormabeli hiba látható.



8.2.8 Animáció. Tekintsük a $\Delta u = 0$ Laplace-egyenletet az $r \in [0,1]$, $\phi \in [-3\pi/4, 3\pi/4]$ polárkoordinátás paraméterezésű egységsugarú körcikken az alábbi peremfeltételekkel: a köríven $u(x,y) = \cos(2 \operatorname{arctg} 2(y,x)/3)$, az egyenes szakaszokon pedig $u(x,y) = 0$! A lenti animáció a végeselem-módszerrel nyert numerikus megoldásokat abban az esetben, ha a konkáv sarok körül finomítjuk a háromszögfelosztást. Felül a háromszögek száma és a maximumnormabeli hiba látható. Figyeljük meg, hogy sokkal kevesebb háromszög elegendő egy adott pontosság eléréséhez, mint az előző animációban!



9. fejezet

Feladatok

9.1. Feladat. Igazoljuk, hogy egy megengedett előjeleloszlású négyzetes mátrix pontosan akkor diagonálisan domináns, ha M-mátrix a $g := e$ vektor mellett!

9.2. Feladat. Igazoljuk, hogy ha $\bar{A}$ diagonálisan domináns mátrix, akkor reguláris!

(Útmutatás: indirekt tegyük fel, hogy $\bar{A}$ egy $x \neq 0$ vektort 0-ba visz, és írjuk fel a definíciót x legnagyobb abszolút értékű koordinátájára.)

9.3. Feladat. Igazoljuk, hogy ha az $\bar{A}$ megengedett előjeleloszlású mátrix diagonálisan domináns mátrix, akkor $\bar{A}^{-1} \geq 0$!

(Útmutatás: legyen D a mátrix főátlója, és $A = D(I - B)$. Mutassuk meg, hogy $B \geq 0$ és $\|B\|_\infty = \max_j \sum_j b_{ij} < 1$, majd írjuk fel $A = D(I - B)$ inverzét Neumann-sor alapján.)

9.4. Feladat. Igazoljuk, hogy ha A M-mátrix, akkor reguláris!

(Útmutatás: legyen $G := diag(g_i)$. Mutassuk meg, hogy $\bar{A} := AG$ diagonálisan domináns és alkalmazzuk a 9.2. feladatot.)

9.5. Feladat. Igazoljuk, hogy ha A M-mátrix, akkor $A^{-1} \geq 0$!

(Útmutatás: az előző feladatbeli $\bar{A} := AG$ és a 9.3. feladat alapján.)

9.6. Feladat. Legyen A adott M-mátrix valamely $g > 0$ vektor mellett. Igazoljuk, hogy ekkor $\|A^{-1}\|_\infty \leq \frac{\max g}{\min Ag}$!

(Útmutatás: mutassuk meg, hogy bármely b vektorra $\pm b \leq \|b\|_\infty (Ag/\min Ag)$ (koordinátánként értve), és így A^{-1} monotonitása miatt az $Ax = b$ lineáris algebrai egyenletrendszer megoldására $\pm x \leq \|b\|_\infty (g/\min Ag)$.)

9.7. Feladat. Igazoljuk, hogy ha A diagonálisan domináns, akkor pozitív szemidefinit!

(Útmutatás: a kvadratikus alak alsó becslése Cauchy–Schwarz-egyenlőtlenségek révén lesz 0.)

9.8. Feladat. Tekintsünk egy 3×2 belső pontból álló, a két irányban egyforma lépésközű rácsot, azaz $\omega_h := \{(ih,\, jh) :\ \ i = 1,2,3,\ \ j = 1,2\}$. Írjuk fel a homogén Dirichlet-peremfeltételű $-\Delta_h$ leképezés A_h mátrixát!

(Útmutatás: a peremmel szomszédos pontokban a peremfeltétel alapján kell.)

9.9. Feladat. Igazoljuk a (2.10) diszkrét Green-formulát!

9.10. Feladat. Igazoljuk az egydimenziós diszkrét Laplace-operátor sajátértékeit és sajátvektorait meghatározó a (2.11) egyenlőségeket!

(Útmutatás: addíciós tételből.)

9.11. Feladat. Igazoljuk a kétdimenziós diszkrét Laplace-operátor sajátértékeit és sajátvektorait meghatározó képleteket az előző feladat alapján!

9.12. Feladat. Írjuk fel a (2.17)-beli $\tilde{A}_h$ mátrixot téglalapon adott inhomogén Dirichlet-feladat esetén!

9.13. Feladat. Írjuk fel a hibafüggvény Galjorkin-ortogonalitásának konkrét képletét a homogén Dirichlet-feladat (3.15)-beli végeselemes megoldására!

9.14. Feladat. Tekintsük a (3.17) vegyes feladatot az ott tett feltételekkel, azaz $s, q \geq 0$, $q \in L^\infty(\Omega)$, $s \in L^\infty(\Gamma_N)$, $\gamma \in L^2(\Gamma_N)$, ill. Γ_D és Γ_N a perem mérhető felbontását alkotják és Γ_D pozitív mértékű. Adjunk becslést a megfelelő

$$a(u,v) := \int_\Omega (p\ \nabla u \cdot \nabla v + quv) + \int_{\Gamma_N} suv$$

bilineáris forma határaira!

(Útmutatás: Cauchy–Schwarz és Szoboljev-féle beágyazás révén $m \geq \operatorname{ess\,inf} p$ és $M \leq \operatorname{ess\,sup} p + C_\Omega \operatorname{ess\,sup} q + C_{\Gamma_N} \operatorname{ess\,sup} s$.)

9.15. Feladat. Igazoljuk, hogy az 1.8 Poincaré–Friedrichs-egyenlőtlenségben C_Ω éles értéke $\frac{1}{\sqrt{\lambda_1}}$, ahol $\lambda_1 > 0$ a Laplace-operátor legkisebb sajátértéke az Ω tartományon homogén Dirichlet-peremfeltétellel!

(Útmutatás: Green-formula és Fourier-sorfejtés, lásd [16, (8.10)].)

9.16. Feladat. Igazoljuk, hogy

$$\lambda_1 \geq \frac{n\pi^2}{diam(\Omega)^2},$$

ahol $\lambda_1 > 0$ a Laplace-operátor legkisebb sajátértéke az Ω tartományon homogén Dirichlet-peremfeltétellel, n a tér dimenziója és $diam(\Omega)$ a tartomány átmérője!

(Útmutatás: foglaljuk az Ω tartományt téglalapba és használjuk fel, hogy ekkor $\lambda_1 := \lambda_1(\Omega) \geq \lambda_1(T)$, lásd [25]. Itt $\lambda_1(T)$-re explicit képletet tudunk a szinuszos sajátfüggvényekből.)

9.17. Feladat. Tekintsük a (3.51) nem szimmetrikus feladatot vegyes peremfeltétellel. Legyen $\varphi_1, \dots, \varphi_n$ bázis a $V_h \subset H^1_D(\Omega)$ altérben, és tekintsük a végeselemben megoldandó $A_h c = b_h$ lineáris algebrai egyenletrendszert. Írjuk fel az A_h mátrix a_{ij} együtthatóit és a jobboldal b_i koordinátáit!

9.18. Feladat. Tekintsük a Poisson-egyenletet Dirichlet- peremfeltétellel téglalapon. Igazoljuk, hogy az egyenletes (azaz négyzetrács egyirányú átlós felezéseiből kapott) háromszögrácshoz tartozó Courant-elemek merevségi mátrixa a 3.14. megjegyzésben megadott alakú!

9.19. Feladat. Igazoljuk, hogy a (3.61) feladat gyenge alakja (3.62)!

(Útmutatás: először mutassuk meg az $u_{|\partial\Omega} = 0$ peremfeltételből, hogy a τ érintőirányban $\partial_\tau u\big|_{\partial\Omega} = 0$, így a $\partial_\nu u\big|_{\partial\Omega} = 0$ feltétellel együtt $\nabla u\big|_{\partial\Omega} = \mathbf{0}$. Ezután kétszer alkalmazzuk a Gauss–Osztrogradszkij-tételt.)

9.20. Feladat. Igazoljuk, hogy a (3.64) bilineáris forma korlátos és koercív a $H^2_0(\Omega)$ téren a (3.65) skalárszorzatra nézve $M = 2$ és $m = 1$ határokkal!

(Útmutatás: a két kifejezés majdnem ugyanaz, csak a vegyes deriváltak szerepelnek duplán a bilineáris formában.)

9.21. Feladat. Igazoljuk a (3.42) becslést!

(Útmutatás: $R(u_h)v = a(u^* - u_h, v)$, és legyen $v := u^* - u_h$.)

9.22. Feladat. Igazoljuk, hogy ha A_h a Poisson-egyenlet FDM-es diszkretizációjának (2.7) mátrixa, akkor

$$\kappa(A_h) := O(h^{-2})$$

nagyságrendű.

(Útmutatás: explicit képlet van a szélső sajátértékekre.)

9.23. Feladat. Tekintsük a Poisson-egyenletet Dirichlet-peremfeltétellel téglalapon. Igazoljuk, hogy ha A_h egyenletes (azaz négyzetrács egyirányú átlós felezéseiből kapott) háromszögrácshoz tartozó Courant-elemekkel felírt merevségi mátrix, akkor

$$\kappa(A_h) := O(h^{-2})$$

nagyságrendű.

(Útmutatás: az előző feladatból és a 3.14. megjegyzésből következik.)

9.24. Feladat. Írjuk fel FDM esetén a lineáris kiterjesztés prolongációs mátrixát egydimenziós esetben (intervallumon)!

9.25. Feladat. Írjuk fel FDM esetén a megszorítás restrikciós mátrixát egydimenziós esetben (intervallumon)!

9.26. Feladat. Írjuk fel FDM esetén a lineáris kiterjesztés prolongációs mátrixának transzponáltját egydimenziós esetben (intervallumon), és vizsgáljuk meg, milyen súlyozott megszorításnak felel ez meg!

9.27. Feladat. Írjuk fel FDM esetén a lineáris kiterjesztés prolongációs mátrixát kétdimenziós esetben (téglalapon)!

9.28. Feladat. Írjuk fel FEM esetén a beágyazás prolongációs mátrixát egydimenziós esetben (intervallumon)!

9.29. Feladat. Írjuk fel FEM esetén a beágyazás prolongációs mátrixát kétdimenziós esetben (téglalapon)!

9.30. Feladat. Legyen $f \in V_h = span\{\varphi_1, \dots, \varphi_n\}$ végeselemes altér $\mathbf{T_1}$- vagy $\mathbf{R_1}$-elemekkel, és legyen a trianguláció reguláris. Legyen M_h a tömegmátrix, azaz $M_{ij} := \int_\Omega \varphi_i \varphi_j$. Igazoljuk, hogy $M_h d \cdot d = O(h^2)\,|d|^2$ $(\forall d \in \mathbb{R}^n)$.

(Útmutatás: írjuk fel előbb egy háromszögön, majd összegezzük.)

9.31. Feladat. Legyen $f \in V_h = span\{\varphi_1, \dots, \varphi_n\}$ végeselemes altér $\mathbf{T_1}$- vagy $\mathbf{R_1}$-elemekkel, és legyen a trianguláció reguláris. Legyen $f_h \in \mathbb{R}^n$ az a vektor, melyre $(f_h)_i = \int_\Omega f\varphi_i$ $(i = 1, ..., n)$. Igazoljuk, hogy $|f_h| = O(h)\|f\|_0$ (ahol $\|f\|_0 := \|f\|_{L^2}$)!

(Útmutatás: Igazoljuk, hogy $\|f\|_0^2 = M_h c \cdot c$, ill. hogy ha c az f együtthatóvektora V_h-ban, akkor $f_h = Mc$, végül használjuk a 9.30. feladatot $d = M^{1/2}c$-re.)

9.32. Feladat. Igazoljuk, hogy az (5.1) formulában bevezetett f_ω paraméterbecslő függvényre $\max_{x\in[\frac{1}{2},1]} |f_\omega(x)| = \max\{|1-\omega|,\ |2\omega - 1|\}$!

9.33. Feladat. Igazoljuk, hogy az 5.8./1. példában bevezetett f_ω paraméterbecslő függvényre $\max\{|f_\omega(x,y)| :\ \ x \geq \frac{1}{2}$ vagy $y \geq \frac{1}{2}\ \} = \max\{|1 - \frac{\omega}{2}|,\ |2\omega - 1|\}$! Határozzuk meg ebből az optimális iterációs paramétert a $[0,1]^2$ tartományon egyenletes háromszögrácson!

9.34. Feladat. Igazoljuk a 6.1. állítást!

(Útmutatás: Gauss–Osztrogradszkij-tételből.)

9.35. Feladat. Igazoljuk, hogy (6.9) és (6.10) ekvivalens!

9.36. Feladat. Igazoljuk 2D esetben a (6.18) becslést, és hogy egyenlőség is fennállhat!

(Útmutatás: először két parciális integrálással a Gauss–Osztrogradszkij-tételnek megfelelően $\int_\Omega \partial_1 u_1\ \partial_2 u_2 = \int_\Omega \partial_1 u_2\ \partial_2 u_1$, majd a definíciókból és a Cauchy–Schwarz-egyenlőtlenségből kapjuk (6.18)-t. Egyenlőség az $(u_1, u_2) = \nabla p$ esetben adódik.

9.37. Feladat. Igazoljuk, hogy ha teljesül a 7.1. feltétel, akkor a (7.9)-ben definiált F operátor egyenletesen monoton a $H_0^1(\Omega)$ téren!

9.38. Feladat. Írjuk fel (7.31) mintájára a csillapított Newton-iteráció algoritmusát a (7.10) szemilineáris konvekció-reakció-diffúziós feladat végeselemes diszkretizációjára!

II. rész

Időfüggő parciális differenciálegyenletek numerikus módszerei

10. fejezet

Szükséges ismeretek rövid áttekintése

Az $u : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ típusú függvényekre vonatkozó alábbi alakú általános egyenletet fogjuk vizsgálni:

$$\partial_t u(t, \mathbf{x}) = Lu(t, \mathbf{x}), \quad t \in (0, T), \mathbf{x} \in \Omega, \tag{10.1}$$

illetve ezen egyenletre vonatkozó

$$\begin{cases} \partial_t u(t, \mathbf{x}) = Lu(t, \mathbf{x}), & t \in (0, T), \mathbf{x} \in \Omega \\ u(0, \mathbf{x}) = u_0(\mathbf{x}), & \mathbf{x} \in \Omega \end{cases} \tag{10.2}$$

kezdetiérték- vagy Cauchy-feladatokat, ahol $\Omega \subset \mathbb{R}^d$ egy adott lokálisan Lipschitz-tartomány, L pedig egy differenciáloperátor. Ennek tulajdonságait a következő szakaszban részletezzük. A véges differenciákkal történő közelítő módszerek felírásához külön figyelembe kell vennünk a peremfeltételeket is, amelyeket részletesen kiírva a (10.2) feladat a következő alakú:

$$\begin{cases} \partial_t u(t, \mathbf{x}) = \tilde{L}u(t, \mathbf{x}) + f(t, \mathbf{x}), & t \in (0, T), \mathbf{x} \in \Omega \\ \tilde{\mathbf{D}}u(t, \mathbf{x}) = \mathbf{g}(t, \mathbf{x}), & t \in (0, T), \mathbf{x} \in \Gamma_j,\ j = 1, 2, \ldots, M \\ u(0, \mathbf{x}) = u_0(\mathbf{x}), & \mathbf{x} \in \Omega, \end{cases} \tag{10.3}$$

ahol $j = 1, 2, \ldots, M$ esetén

$$\mathbf{D}u = (D_1 u, D_2 u, \ldots, D_M u),\ D_j u : (0, T) \times \Gamma_j \to \mathbb{R}$$

valamint

$$\mathbf{g} = (g_1, g_2, \ldots, g_M)(t, \mathbf{x}),\ g_j : (0, T) \times \Gamma_j \to \mathbb{R}$$

az Ω halmaz határának M darab diszjunkt $\Gamma_j, \Gamma_2, \ldots, \Gamma_M \subset \partial\Omega$ részén, továbbá

$$\tilde{L}u(t, \mathbf{x}) = \sum_s a_s \partial^{(0, \boldsymbol{\alpha}_s)} u(t, \mathbf{x}) \tag{10.4}$$

és

$$\tilde{D}_j u(t,\mathbf{x}) = \sum_s b_s \partial^{(0,\boldsymbol{\beta}_{j,s})} u(t,\mathbf{x}) \tag{10.5}$$

megfelelő $\{\boldsymbol{\alpha}_s\}$ és $\{\boldsymbol{\beta}_{1,s}\}, \{\boldsymbol{\beta}_{2,s}\}, \ldots, \{\boldsymbol{\beta}_{M,s}\}$ multiindex-halmazokkal. Itt a peremfeltételek az ismert g_j függvényekkel adottak.

10.1. Példa. Tekintsük a $\partial_t u(t,x,y) = \partial_{xx} u(t,x,y) + \partial_{yy} u(t,x,y)$ hővezetési egyenletet az $Q = (0,0),(0,1),(1,0)$ csúcsokkal rendelkező háromszögön annak alsó és bal oldalsó lapján adott konstans 4 peremfeltétellel, az oldalsó lapon pedig homogén Neumann-peremfeltétellel, valamint az adott $u(0,x,y) = 1 - x - y$ kezdeti feltétellel!

Ekkor Γ_1 az alsó, Γ_2 a bal oldalsó, Γ_3 pedig a harmadik oldal, továbbá

$$\begin{cases} \tilde{L}u(t,x,y) = \partial_{xx} u(t,x,y) + \partial_{yy} u(t,x,y) \\ \tilde{D}_1 u(t,x,y) = \tilde{D}_2 u(t,x,y) = u(t,x,y),\ g_1(t,x,y) = g_2(t,x,y) = 4, \\ \tilde{D}_3 u(t,x,y) = \partial_x u(t,x,y) + \partial_y u(t,x,y),\ g_3(t,x,y) = 0 \end{cases}$$

azaz $a_1 = a_2 = 1$, $\boldsymbol{\alpha}_1 = (0,2,0), \boldsymbol{\alpha}_2 = (0,0,2)$, $b_{11} = b_{21} = 1$, $\boldsymbol{\beta}_{11} = \boldsymbol{\beta}_{21} = (0,0)$, továbbá $b_{31} = b_{32} = 1$, $\boldsymbol{\beta}_{31} = (1,0), \boldsymbol{\beta}_{32} = (0,1)$. ♢

10.1. Időfüggő parciális differenciálegyenletekkel kapcsolatos ismeretek

A (10.1) egyenletben az ismeretlen u függvény

$$u : (0,T) \times \Omega \to \mathbb{R} \quad \text{vagy} \quad \mathbf{u} : (0,T) \times \Omega \to \mathbb{R}^d$$

típusú; a második típus esetén rendszerről beszélünk.

Az L (differenciál)operátor

$$L : L_2(\Omega) \to L_2(\Omega) \quad \text{vagy} \quad L : [L_2(\Omega)]^d \to [L_2(\Omega)]^d$$

típusú, amely mindig sűrűn definiált és általában nem korlátos. Az L operátor értelmezési tartományát az eredeti feladatban szereplő peremfeltételek határozzák meg. Fontos, hogy az elméletben a differenciáloperátorokat általában egy altéren értelmezzük, ezért az esetleges nem homogén peremfeltételekkel kitűzött feladatot olyan alakra kell átírni, hogy ezt teljesítse. A magyarul használt terminológiával ellentétben nem vegyes feladatról beszélünk.

10.2. Példa. A továbbiakban is gyakran idézett hővezetési/diffúziós egyenletre vonatkozó homogén Dirichlet-peremfeltétellel adott kezdetiérték-feladat az alábbi:

$$\begin{cases} \partial_t u(t,\mathbf{x}) = \Delta u(t,\mathbf{x}) + f(t,\mathbf{x}), \quad t \in (0,T),\ \mathbf{x} \in \Omega \\ u(t,\mathbf{x}) = 0, \quad t \in (0,T), \mathbf{x} \in \partial\Omega \\ u(0,\mathbf{x}) = u_0(\mathbf{x}), \quad \mathbf{x} \in \Omega. \end{cases} \tag{10.6}$$

Itt $\mathcal{D}(L) = H_0^1(\Omega) \cap H^2(\Omega)$, $Lu = \Delta u$, valamint $f \in L_2(\Omega)$ adott. ♢

Fontos tudni, hogy a felírt feladatoknak létezzen megoldása és az egyértelmű legyen. Nem idézünk erre vonatkozó állításokat, de megjegyezzük, hogy az általunk a továbbiakban vizsgált esetekben ez teljesül. A numerikus módszerektől is akkor várhatunk pontos közelítést, ha a feladatban szereplő összes függvénytől folytonosan függ a megoldás. Gyakran hasznos az ezen függést kifejező képletek ismerete.

10.3. Példa. A fenti példában a (10.6) feladat megoldása a $C((0,T);\mathcal{D}(L))$ térben létezik, egyértelmű, és a megoldás az alábbi képlet szerint függ a kezdeti feltételektől:

$$\|u(t,\cdot)\|_{H^1(\Omega)} \leq e^{-t}\|u_0\|_{H^1(\Omega)} + \|f\|_{L_2(\Omega)}.$$ ♢

10.4. Megjegyzés. A fenti (10.2) feladat egy általánosított differenciálegyenlet, amit általánosított Cauchy-problémának neveznek a félcsoportelmélet nyelvén. A félcsoportelmélet kapcsolatot teremt az L differenciáloperátor (pontosabban az ezzel felírt peremértékfeladatok megoldóoperátorainak) tulajdonságai és a (10.2) feladat megoldhatósága, valamint a megoldás tulajdonságai között. ♢

10.2. Néhány fogalom a numerikus módszerek köréből

Használni fogunk a közönséges differenciálegyenletekre vonatkozó néhány egyszerűbb módszert: az *explicit és implicit Euler-módszert* és a többlépéses módszerek levezetésének elvét.

Alapvető kvadratúraképleteket is alkalmazunk, a középpont-szabályt és a trapézformulát fogjuk használni közelítések levezetéséhez. Ismertnek tételezzük fel az ezek rendjére vonatkozó állításokat.

A rend fogalmát az általunk vizsgált esetre ismét felírjuk.

10.3. Véges differenciák egyenletes rácsfelosztásokon

A jegyzet első felében azokat a módszereket tárgyaljuk, amelyekben a véges differenciák módszerével közelítjük a feladat megoldását. Egymást követő rögzített időpontokban közelítjük a megoldás értékét az $\bar{\Omega}$ tartomány egyes pontjaiban. Ennek leírásához vezetjük be a következő jelöléseket:

- $h_x, h_y, h_z, \ldots$ - térbeli diszkretizációban a rácshossz az egyes irányokban.
- $\mathbf{h} = (h_x, h_y, h_z, \ldots) = (h_1, h_2, \ldots, h_d)$ a rácshosszokat tartalmazó vektor, $\mathbf{h} = h_x$ esetén ezek helyett egyszerűen a h jelölést használjuk.

- $\mathbf{k} \otimes \mathbf{h} = (k_1 h_1, k_2 h_2, \ldots, k_d h_d)$, ahol $\mathbf{k} = (k_1, k_2, \ldots, k_d) \in \mathbb{Z}^d$ az egyes rácspontokat megadó vektor.
- δ - időlépés, ahol a megoldást egymás után a $\delta, 2\delta, \ldots$ időpontokban közelítjük.
- $\mathbf{x} = (x, y, z)$ vagy $\mathbf{x} = (x, y)$.
- u_h (ill. $u_{\mathbf{h}}$) az u megoldás közelítése a h-val (ill. $\mathbf{h}$-val) paraméterezett rácson.
- $u^n_{\mathbf{k}} = u_h(n\delta, \mathbf{k} \otimes \mathbf{h})$ a numerikus megoldás értéke az n-edik időlépésben a $\mathbf{k} \otimes \mathbf{h}$ helyen.
- $\mathbf{u}^n = \{u_h(n\delta, \mathbf{x}_h) : \mathbf{x}_h \text{ egy rácspont}\}$, amely az n-edik időpontban a numerikus megoldás értékeit tartalmazza $\bar{\Omega}$-nak a h-hoz tartozó felosztás rácspontjaiban.
- $\mathbf{u}(t, \cdot) = \{u(t, \mathbf{x}_h) : \mathbf{x}_h \text{ egy rácspont}\}$, amely a t időpontban a pontos megoldás értékeit tartalmazza $\bar{\Omega}$-nak a h-hoz tartozó felosztás rácspontjaiban.

10.5. Megjegyzés.

1. A h, h_x, h_y, h_z felosztásparaméterek felülről korlátosak, egy természetes korlát az Ω halmaz átmérője.
2. Az utolsó három jelölés az egyszerűség kedvéért nem tartalmazza a h paramétert, de ezeket mindig rögzített felosztás mellett használjuk, ezért nem fog félreértést okozni. ♢

A következőkben a felosztások konstrukciójához szükséges definíciókat részletezzük.

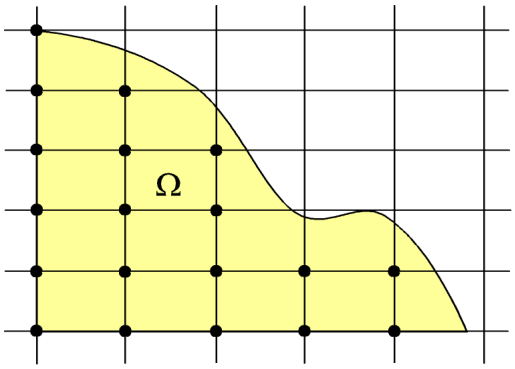
10.6. Definíció. Tetszőleges egész $k_1, k_2, \ldots, k_d$ esetén a $[k_1 h_1, (k_1+1)h_1] \times [k_2 h_2, (k_2+1)h_2] \times \cdots \times [k_d h_d, (k_d+1)h_d]$ halmazt $\mathbf{h}$-téglának nevezzük. ♢

Először egy egyenletes rács azon pontjait adjuk meg, amelyek az Ω tartomány lezártjába esnek, majd az Ω halmaz ennek megfelelő módosítását.

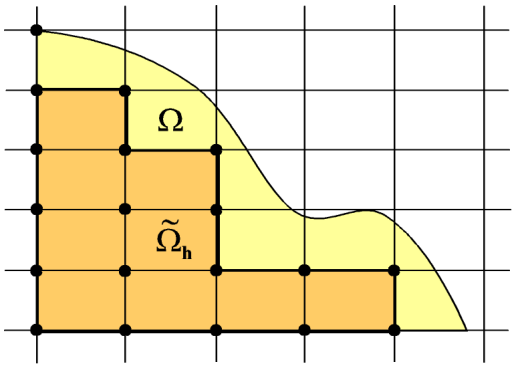
10.7. Definíció. Legyen az $\Omega_{\mathbf{h}} = \{(j_1 h_1, j_2 h_2, \ldots, j_d h_d) \in \bar{\Omega} : (j_1, j_2, \ldots, j_d) \in \mathbb{Z}^d\}$, vagyis azon rácspontok halmaza, amelyek $\bar{\Omega}$-ban vannak (10.1. ábra).
Jelölje $\tilde{\Omega}_h$ azon h-téglák uniójának belsejét, amelyeknek az összes csúcsa Ω_h-ban van (10.2. ábra).

Ha például $d = 2$, akkor

$$\tilde{\Omega}_{\mathbf{h}} = \operatorname{int}\Big\{\bigcup [jh_x, jh_x + h_x] \times [kh_y, kh_y + h_y] : (jh_x, kh_y), (jh_x, kh_y + h_y), (jh_x + h_x, kh_y), (jh_x, kh_y + h_y) \in \Omega_{\mathbf{h}}\Big\}.$$



10.1. ábra. Egy lehetséges Ω tartomány, a hozzá tartozó rácsháló és az $\Omega_{\mathbf{h}}$ rácspontok halmaza.



10.2. ábra. Az $\tilde{\Omega}_{\mathbf{h}}$ halmaz szemléltetése.

10.8. Megjegyzés.

1. Ha Ω nem korlátos, akkor az $\Omega_{\mathbf{h}}$ halmaz végtelen lehet.

2. A $<$ reláció a vektorok közt részbenrendezést jelent, azaz $\mathbf{h} < \mathbf{h}_0$ pontosan akkor teljesül, ha minden komponensre igaz az egyenlőtlenség.

3. Mivel az Ω halmaz nyílt, ezért Lebesgue-mérhető, vagyis a fenti definíció alapján

$$\lim_{\mathbf{h}\to 0} \lambda(\tilde{\Omega}_{\mathbf{h}}) = \lambda(\Omega), \tag{10.7}$$

   ahol λ a megfelelő dimenzióban vett Lebesgue-mértéket jelöli.

4. Az egyszerűség kedvéért az egyes számítási algoritmusokat mindig Ω_h típusú tartományokon adjuk meg. A könnyebb érthetőség kedvéért a következő két definíciót $d = 2$ esetére írjuk fel. ♢

10.9. Definíció. Legyen valamilyen $\mathcal{H} \subset \mathbb{Z}^3$ indexhalmazra $\{(t+i\delta, x+jh_x, y+kh_y) : (i,j,k) \in \mathcal{H}\} \subset [0,T] \times \bar{\Omega}$, emellett a fenti halmaz elemszámát jelölje K. Ekkor a

$$D_{\delta,h_x,h_y} u(t,x,y) = \sum_{s=1}^{K} a_s u(t+i\delta, x+jh_x, y+kh_y)$$

alakú összeget a (t,x,y) pontokhoz tartozó véges differenciának nevezzük, ahol a_s is függ (δ, h_x, h_y)-tól. A lehetséges i,j,k indexek számát az egyes irányokhoz, változókhoz tartozó lépésszámnak nevezzük. ♢

10.10. Definíció. Azt mondjuk, hogy a $\mathcal{D}_{\tilde{L}}$ véges differencia a $(t,\mathbf{x})$ pontban $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_d)$ rendben közelíti az $\tilde{L}$ differenciáloperátort, ha minden

$$u \in C^{(\gamma_1+1, \gamma_2+1, \ldots, \gamma_d+1)}(\Omega)$$

függvényre

$$\tilde{L}u(t,\mathbf{x}) = \mathcal{D}_{\tilde{L}} u(t,\mathbf{x}) + \mathcal{O}(h_1^{\gamma_1}) + \cdots + \mathcal{O}(h_d^{\gamma_d}). \tag{10.8}$$

Az egyes γ_j, $j = 1,2,\ldots,d$ számokat a megfelelő változókhoz tartozó rendnek nevezzük. ♢

10.11. Megjegyzés.

1. Hasonlóan definiálható a ∂_t differenciáloperátor közelítésének rendje is.

2. A továbbiakban is gyakran használjuk a nagyságrendekre vonatkozó $\mathcal{O}(\cdot)$ jelölést. Az előző definícióban ez azt jelenti, hogy az $\tilde{L}$ és a $\mathcal{D}_{\tilde{L}}$ operátorok alkalmazásával kapott két mennyiség különbsége felülről becsülhető egy olyan függvénnyel, amely legfeljebb $C(t,\mathbf{x}) \cdot (h_1^{\gamma_1} + \cdots + h_d^{\gamma_d})$, ahol $C(t,\mathbf{x})$ a t és $\mathbf{x}$ értékektől függő konstans. Az itt szereplő $\gamma_1, \ldots, \gamma_d$ exponensek mutatják, hogy a közelítésből kapott hiba milyen nagyságrendben csökken.

3. Mindkét fenti definícióban az L operátor tartalmazza a peremfeltételeket, így annak véges differenciás közelítéséhez a peremfeltételekre vonatkozó közelítéseket is meg kellene adnunk. Ezt itt nem részletezzük, viszont abban a konkrét formában, ahogy arra szükségünk lesz, a konzisztenciarendről szóló fejezetben tárgyaljuk. ♢

Gyakran pontosan adjuk meg az f függvényt, de lehet, hogy $f(t,x,y,\ldots)$ értékét is közelítenünk kell.

Gyakran használt véges differenciás közelítések, a megfelelő közelítések rendje

Néhány fontos példát sorolunk fel, amelyek többségét a 2.1.1 szakaszban már bevezettük. Minden esetben feltesszük, hogy a képletben szereplő $v : \mathbb{R}^d \to \mathbb{R}$ típusú függvények mindegyike annyiszor deriválható, hogy a 10.10. definíció alkalmazható legyen rá. Az egyszerűség kedvéért a t változót egyetlen példában sem írjuk ki.

- jobb oldali differencia 1 dimenzióban:

$$D_+v(x) = v(x+h_x) - v(x), \qquad D_+v(x) = h\partial_x v(x) + \mathcal{O}(h^2)$$

- bal oldali differencia 1 dimenzióban:

$$D_-v(x) = v(x) - v(x-h_x), \qquad D_-v(x) = h\partial_x v(x) + \mathcal{O}(h^2)$$

- centrális differencia 1 dimenzióban és egy változóra 2 dimenzióban:

$$D_0v(x) = \frac{1}{2}(v(x+h_x) - v(x-h_x))$$
$$D_{0,x}v(x,y) = \frac{1}{2}(v(x+h_x,y) - v(x-h_x,y)),$$

  ahol

$$D_0v(x) = h\partial_x v(x) + \mathcal{O}(h^3) \qquad \text{és} \qquad D_{0,x}v(x,y) = h_x\partial_x v(x,y) + \mathcal{O}(h^3)$$

- centrális differencia a második derivált közelítésére 1 dimenzióban és egy változóra 3 dimenzióban:

$$D_0^2v(x) = v(x+h_x) - 2v(x) + v(x-h_x)$$
$$D_{0,y}^2v(x,y,z) = v(x,y+h_y,z) - 2v(x,y,z) + v(x,y-h_y,z),$$

  ahol

$$D_0^2v(x) = h^2\partial_{xx}v(x) + \mathcal{O}(h^4) \quad \text{és} \quad D_{0,y}^2v(x,y,z) = h_y^2\partial_{yy}v(x,y,z) + \mathcal{O}(h^4).$$

  Nyilván

$$D_0^2v(x) = D_+v(x) - D_-v(x).$$

- Egy összetett centrális differencia a $\partial_{xx}\partial_{yy}$ negyedrendű derivált közelítésére 2 dimenzióban:

$$\begin{aligned} D_{0,x}^2D_{0,y}^2v(x,y) = {} & v(x+h_x,y+h_y) - 2v(x+h_x,y) + v(x+h_x,y-h_y) - \\ & - 2(v(x,y+h_y) - 2v(x,y) + v(x,y-h_y)) + \\ & + v(x-h_x,y+h_y) - 2v(x-h_x,y) + v(x-h_x,y-h_y), \end{aligned}$$

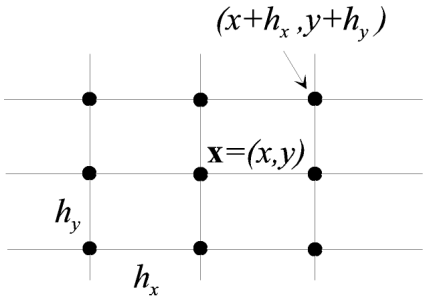
  ahol

$$D_{0,x}^2D_{0,y}^2v(x,y) = h_x^2h_y^2\partial_{yy}\partial_{xx}v(x,y) + \mathcal{O}(h_x^4) + \mathcal{O}(h_y^4).$$

Ugyanezt a jelölést használjuk a rácspontokon értelmezett v függvényekre is. Ezzel a fenti utolsó példa a következő alakú:

$$D^2_{0,x}D^2_{0,y}v^n_{j,k} = v^n_{j+1,k+1} - 2v^n_{j+1,k} + v^n_{j+1,k-1} - 2(v^n_{j,k+1} - 2v^n_{j,k} + v^n_{j,k-1}) + \\ + v^n_{j-1,k+1} - 2v^n_{j-1,k} + v^n_{j-1,k-1}.$$

Az $\Omega_{\mathbf{h}}$ halmaz azon pontjait, amelyek az $\mathbf{x}$-beli véges differencia kiszámításához szükségesek, az $\mathbf{x}$-hez és a sémához tartozó alappontoknak vagy az $\mathbf{x}$-hez és a sémához tartozó differenciacsillagnak nevezzük (10.3. ábra).



10.3. ábra. Az $\mathbf{x}$ ponthoz és a $D^2_{0,x}D^2_{0,y}$ véges differenciához tartozó alappontok.

10.12. Megjegyzés. Ha a fenti véges differenciákat h megfelelő hatványával osztjuk, akkor az első, illetve második derivált közelítését kapjuk. A 2.1.1. szakaszban ilyen módon definiáltunk különböző differenciasémákat. ♢

10.13. Megjegyzés. Ezeknek a véges differenciáknak azonban nem minden $\Omega_{\mathbf{h}}$ pontban van értelme, hiszen lehet, hogy a szükséges alappontok nincsenek $\Omega_{\mathbf{h}}$-ban. Ekkor a megadott peremfeltételeket használjuk az egyenletekben megadott differenciáloperátorok közelítéséhez. ♢

Szintén a fenti jelölést használjuk azokra az $l_\infty \to l_\infty$ típusú lineáris transzformációkra is, amelyeket az alábbi hozzárendeléssel adunk meg:

- $(D_+\mathbf{v})_k = \mathbf{v}_{k+1} - \mathbf{v}_k$
- $(D_-\mathbf{v})_k = \mathbf{v}_k - \mathbf{v}_{k-1}$
- $(D_0\mathbf{v})_k = \mathbf{v}_{k+1} - \mathbf{v}_{k-1}$
- $D_0^2\mathbf{v} = D_+\mathbf{v} - D_-\mathbf{v}$.

Természetesen merül fel a kérdés, hogy lehet mindezt ellenőrizni, valamint, hogy ha nem csak az $\mathbf{x}$ pontban, hanem valamilyen értelemben egyenletesen szeretnénk ezt tudni, akkor milyen feltétel teljesülését kell megvizsgálni.

Az ezzel kapcsolatos állításokat nem részletezzük, de utalunk rá, hogy mindezt Taylor-sorfejtések segítségével kaphatjuk. Az itt kapott maradéktag jelenik meg hibatagként a közelítésben, és ha a magasabbrendű deriváltakra valamilyen egyenletes felső korlát is létezik, akkor t-től és x-től független hibabecslés is adható.

További példákat látunk a 21.1. és a 21.2. feladatokban.

10.4. Az egyenletek megoldásának véges differenciás módszerrel való közelítése

10.14. Definíció. A $\partial_t u = Lu + f,\ u(0,\cdot) = u_0$ feladat megoldására felírt véges differencia közelítést (vagy annak egy olyan alakját, amelyből $\mathbf{u}^{n+1}$ megadható $\mathbf{u}^n, \mathbf{u}^{n-1}, \dots, \mathbf{u}^{n-k}$ függvényében) a fenti feladat megoldására vonatkozó sémának nevezzük. ♢

A megoldást közelítő sémákat többféle elven kaphatunk:

- A legkézenfekvőbb módszer az, hogy mind a ∂_t, mind az $L + f$ inhomogén differenciáloperátort (az L által adott mellékfeltételekkel együtt) valamilyen véges differenciával közelítjük, majd ezeket egyenlővé téve a kapott egyenletrendszert megoldjuk.

- Először a feladat jobb oldalát „diszkretizáljuk”, azaz véges sok változóval írjuk le minden rögzített időpontban az u függvényt. Ennek megfelelően helyettesítjük az $L + f$ inhomogén differenciáloperátort valamilyen véges differenciával. A véges sok változó időbeli értékeire vonatkozólag ekkor egy differenciálegyenlet-rendszert kapunk, amelynek megoldását valamilyen szokásos differenciálegyenlet megoldó módszerrel közelítjük.

Mindkét módszerre több példát látunk a továbbiakban. Habár a második módszer kevésbé tűnik természetesnek, kiderül, hogy ennek alapján kis változtatással könnyen kaphatunk végeselemes diszkretizáción alapuló közelítő megoldást is.

10.5. Két konstrukció a hővezetési egyenlet numerikus megoldására

A hővezetési egyenletre vonatkozó alábbi két modellfeladatot vizsgáljuk egydimenziós esetben:

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x) & t \in \mathbb{R}^+,\ x \in \mathbb{R} \\ u(0,x) = f(x) & x \in \mathbb{R}, \end{cases} \tag{10.9}$$

valamint

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x) \quad t \in \mathbb{R}^+,\ x \in (0,\pi) \\ u(t,0) = u(t,\pi) = 0 \quad t \in \mathbb{R}^+ \\ u(0,x) = \sin x \quad x \in (0,\pi). \end{cases} \tag{10.10}$$

A megfelelő differenciáloperátorok 10.4. szakaszban leírt elven történő közelítésével két véges differenciás módszert adunk, amelyekkel a numerikus közelítés kiszámítható.

10.5.1. Az első elv: az egyenlet két oldalának közelítése

Az idő szerinti deriváltra vonatkozó közelítés:

$$\partial_t u(t,x) \approx \frac{1}{\delta}(u(t+\delta,x) - u(t,x)), \tag{10.11}$$

ugyanez a megfelelő hibataggal:

$$\partial_t u(t,x) = \frac{1}{\delta}(u(t+\delta,x) - u(t,x)) + \mathcal{O}(\delta),$$

sőt, még pontosabban

$$\partial_t u(t,x) = \frac{1}{\delta}(u(t+\delta,x) - u(t,x)) - \frac{\delta}{2}\partial_{tt} u(t,x) + \mathcal{O}(\delta^2). \tag{10.12}$$

A térváltozó szerinti deriváltra vonatkozó közelítés:

$$\sigma_D \partial_{xx} u(t,x) \approx \sigma_D \frac{1}{h^2} D_0^2 u(t,x). \tag{10.13}$$

A (10.11) és a (10.13) formulákat egyenlővé téve kapjuk, hogy $x-h \in [0,\pi]$ és $x+h \in [0,\pi]$ esetén

$$\frac{1}{\delta}(u(t+\delta,x) - u(t,x)) \approx \sigma_D \frac{1}{h^2}(u(t,x+h) - 2u(t,x) + u(t,x-h)), \tag{10.14}$$

tehát a közelítés:

$$\begin{aligned} u(t+\delta,x) &\approx u(t,x) + \sigma_D \frac{\delta}{h^2}(u(t,x+h) - 2u(t,x) + u(t,x-h)) \\ &= u(t,x) + r D_0^2 u(t,x), \end{aligned} \tag{10.15}$$

azaz komponensenként felírva:

$$u_k^{n+1} = u_k^n + \sigma_D \frac{\delta}{h^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n) = u_k^n + r(u_{k-1}^n - 2u_k^n + u_{k+1}^n),$$

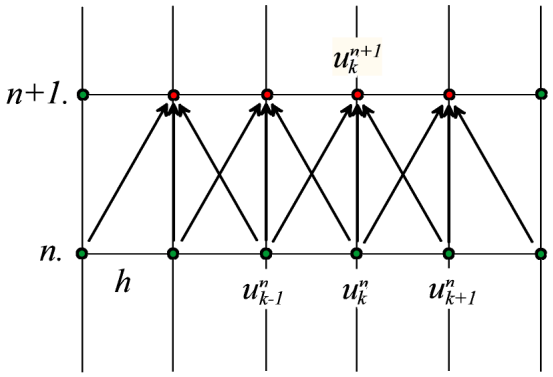
ahol $r = \frac{\sigma_D \delta}{h^2}$, valamint $u_k^n = u_h(n \cdot \delta, k \cdot h),\ k = 1, 2, \ldots, N$. Itt N a $[0, \pi]$ intervallum belső osztópontjainak száma, azaz $h = \frac{\pi}{N+1}$, emellett az eredeti (10.10) feladatnak megfelelően legyen

$$u_0^{n+1} = u_{N+1}^{n+1} = 0.$$

Összefoglalva, az 10.4. szakasz első pontjában leírt módszer alapján a (10.10) feladatra vonatkozó lehetséges séma a következő:

$$\begin{cases} u_k^0 = \sin(kh),\ k = 0, 1, \ldots, N, N+1 \\ u_k^{n+1} = u_k^n + r(u_{k-1}^n - 2u_k^n + u_{k+1}^n),\ k = 1, 2, \ldots, N \\ u_0^{n+1} = u_{N+1}^{n+1} = 0. \end{cases} \tag{10.16}$$

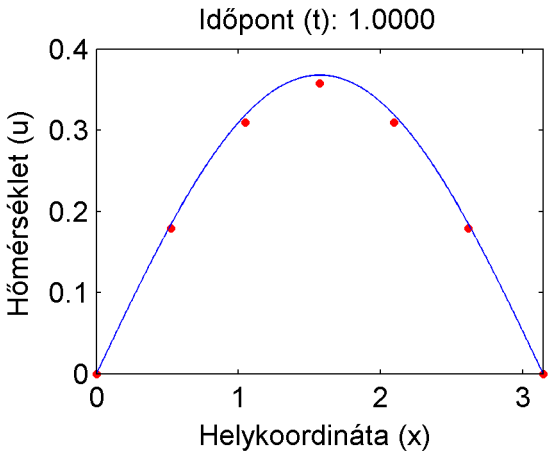
A 10.4. ábra azt mutatja, hogy az $(n+1)$-edik időpontbeli értékeket mely n-edik időpontbeli értékekből tudjuk közvetlenül kiszámolni a séma alapján. A 10.5 és a 10.6 ábrákon pedig a $t = 1$ időponthoz tartozó időrétegen adtunk meg egy-egy konkrét rácson a numerikus megoldást.



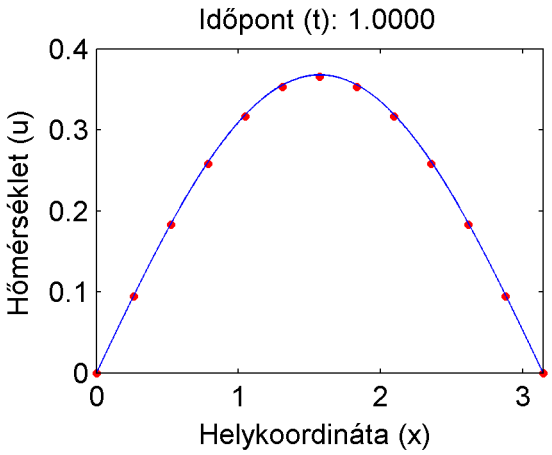
10.4. ábra. Az n-edik időrétegen adott közelítésekből az $(n+1)$-edik időréteg közelítései közvetlenül meghatározhatók.

Ahhoz, hogy (számítógéppel) a számítást végrehajtsuk, a (10.16) sémát a következő egyenletrendszer alakjában adjuk meg:
Jelölje most az $\mathbf{u}^n = (u_1^n, u_2^n, \ldots, u_N^n)$ vektor az intervallum belsejében levő közelítendő értékeket az $n\delta$ időpontban! Ekkor (10.16) második sora alapján az $\mathbf{u}^{n+1}$ vektor k-adik komponensét úgy kapjuk, hogy $\mathbf{u}^n$ vektor k-adik komponensét $1 - 2r$-rel, a $k-1$-edik és a $k+1$-edik komponensét r-rel szorozzuk. Ha $k = 1$ vagy $k = N$, akkor csak azt a komponenst kell vennünk, ami $\mathbf{u}^n$-ben szerepel, mert a többi hozzáadandó a (10.16) harmadik sora szerint nulla. Azaz

$$\mathbf{u}^{n+1} = Q\mathbf{u}^n, \quad \text{ahol} \quad Q = \texttt{tridiag}[r, 1-2r, r],$$



10.5. ábra. A (10.16) séma eredménye a $t = 1$ pontban (piros pontok). $\sigma_D = 1$, $h = \pi/6$, $\delta = 0.1$. A pontos megoldást kék vonal szemlélteti. A legnagyobb eltérés a középső rácspontban van, értéke 0.01033.



10.6. ábra. A (10.16) séma eredménye a $t = 1$ pontban (piros pontok). $\sigma_D = 1$, $h = \pi/12$, $\delta = 0.025$. A pontos megoldást kék vonal szemlélteti. A legnagyobb eltérés értéke a középső rácspontban van, értéke 0.002518.

ahol a fenti szimbólum azt a tridiagonális mátrixot jelöli, amelynek főátlójában mindenhol $1 - 2r$, a mellette levő átlókban mindenhol r áll. Lásd még a 20.1. példában szereplő programot és a 20.2.1-20.2.4. animációkat!

Hasonló példákat látunk a 21.14. és a 21.15. feladatokban, ahol a felírt sémák inhomogén lineáris rendszerekre vezetnek.

Pontosabb közelítést is megadhatunk (10.13) helyett a következő formulákkal:

$$\begin{aligned}
D_+u(t,x) &= h\partial_x u(t,x) + \frac{h^2}{2!}\partial_{xx}u(t,x) + \frac{h^3}{3!}\partial_{xxx}u(t,x) + \frac{h^4}{4!}\partial_{xxxx}u(t,x) + \\
&+ \frac{h^5}{5!}\partial_{xxxx}u(t,x) + \mathcal{O}(h^6) \\
D_-u(t,x) &= h\partial_x u(t,x) - \frac{h^2}{2!}\partial_{xx}u(t,x) + \frac{h^3}{3!}\partial_{xxx}u(t,x) - \frac{h^4}{4!}\partial_{xxxx}u(t,x) + \\
&+ \frac{h^5}{5!}\partial_{xxxx}u(t,x) + \mathcal{O}(h^6)
\end{aligned}$$

amelyeket felhasználva kapjuk, hogy

$$\begin{aligned}
\sigma_D\partial_{xx}u(t,x) &= \sigma_D\frac{1}{h^2}(D_+u - D_-u) + \sigma_D\frac{h^2}{12}\partial_{xxxx}u(t,x) + \mathcal{O}(h^4) \\
&= \sigma_D\frac{1}{h^2}(D_0^2u) + \sigma_D\frac{h^2}{12}\partial_{xxxx}u(t,x) + \mathcal{O}(h^4).
\end{aligned} \tag{10.17}$$

10.15. Megjegyzés. Az itt szereplő h és δ diszkretizációs paraméterek megválasztásának fontos szerepe lesz a konvergencia analízisében. Konkrétabban, az itt bevezetett r paraméter értékétől függ, hogy a fenti sémával kapott közelítő megoldás valóban tart-e az igazihoz. Az analízis egyik lényeges pontja éppen az, hogy kiderül, a h és δ értékét *nem lehet egymástól függetlenül* elegendően kicsinek választani. ♢

Példa a közelítések rendjének növelésére

Egy magasabb rendű módszert ismertetünk, ahol a diszkretizációs paramétereket speciálisan választjuk.

10.16. Állítás. *Ha a (10.16) sémában $\delta = \frac{h^2}{6}$ teljesül, akkor a (10.16) sémával adott véges differenciás közelítés h-ban negyed, t-ben pedig másodrendű.*

Bizonyítás. Használjuk a (10.12) és a (10.17) képletekben megadott pontosabb közelítéseket! Így azt kapjuk, hogy

$$\begin{aligned}
\partial_t u(t,x) - \partial_{xx}u(t,x) &= \frac{1}{\delta}(u(t+\delta,x) - u(t,x)) - \frac{\delta}{2}\partial_{tt}u(t,x) + \mathcal{O}(\delta^2) \\
&- \sigma_D\frac{1}{h^2}(D^2u) + \sigma_D\frac{h^2}{12}\partial_{xxxx}u(t,x) + \mathcal{O}(h^4).
\end{aligned}$$

Vagyis ha itt

$$\sigma_D\frac{h^2}{12}\partial_{xxxx}u(t,x) = \frac{\delta}{2}\partial_{tt}u(t,x),$$

akkor a pontosság h-ban negyed, t-ben pedig másodrendű. Ez a feltétel $\sigma_D \partial_{xx} u = \partial_t u$ miatt éppen azt jelenti, hogy

$$\frac{h^2}{12} = \frac{\delta}{2},$$

ahogy az állításban szerepelt. □

10.17. Megjegyzés. A fenti példa arra világít rá, hogy ha a módszerektől valamilyen extra tulajdonságot várunk, akkor a lehetséges diszkretizációs paraméterek halmaza egyre szűkebb lesz. ♢

10.5.2. A másik megközelítés: szemidiszkretizáció

A 10.4. szakasz második pontjában szereplő elvet követjük. Röviden összefoglalva az eljárás lényege az, hogy a diszkretizációt először csak a térbeli koordinátákra hajtjuk végre, és a kapott véges sok (az időtől folytonosan függő) függvényre egy közönséges differenciálegyenlet-rendszert kapunk, amelynek megoldását ismét numerikusan közelítjük.

A módszert általánosan írjuk le, hogy aztán egyszerűen kiterjeszthető legyen a későbbiekben vizsgált végeselemes diszkretizáció esetére is.

Adott tehát a $\Pi_{\Omega_\mathbf{h}} : C^1(\overline{\Omega}) \to \mathbb{R}^s$ projekció, amelyre a véges differenciák módszere esetén

$$\Pi_{\Omega_\mathbf{h}} u(t, \cdot) = \mathbf{u}(t, \cdot).$$

Ekkor az eredeti feladat helyett a

$$\begin{cases} \partial_t (t \to \Pi_{\Omega_\mathbf{h}} u(t, \cdot)) = L_h \Pi_{\Omega_\mathbf{h}} u(t, \cdot) \\ \Pi_{\Omega_\mathbf{h}} u(0, \cdot) \text{ adott} \end{cases} \tag{10.18}$$

szemidiszkretizált feladatot tekintjük, ahol az $L_h : \mathbb{R}^s \to \mathbb{R}^s$ az L operátor valamilyen „diszkretizációja”. Ez pontosabban azt jelenti, hogy

$$L_h \Pi_{\Omega_\mathbf{h}} u(t, \cdot) = \Pi_{\Omega_\mathbf{h}} L u(t, \cdot) + \mathcal{R}(h), \tag{10.19}$$

ahol valamilyen $(\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{Z}^+$ vektorra

$$\mathcal{R}(h) = \mathcal{O}(h_1^{\alpha_1}) + \mathcal{O}(h_2^{\alpha_2}) + \cdots + \mathcal{O}(h_d^{\alpha_d}).$$

Az eredeti feladat megoldására vonatkozó numerikus eljárást úgy kapunk, ha a (10.18) feladat megoldását valamilyen numerikus módszerrel közelítjük.

10.18. Példa. A szemidiszkretizáció első lépésében a

$$t \to \mathbf{u}(t, \cdot) = (u(t, h), u(t, 2h), \ldots, u(t, \pi - h))$$

függvényt tekintjük ismeretlennek. Ennek megfelelően az eredeti feladatot úgy diszkretizáljuk, hogy minden időpontban a $(0, h, 2h, \dots, \pi)$ pontokban felvett értékekkel adjuk meg, azaz

$$\Pi_{(h,2h,\dots,\pi-h)}u(t,\cdot) = (u(t,h), u(t,2h), \dots, u(t,\pi-h)).$$

Itt természetesen a $(t,0)$ és (t,π) osztópontok is a számításban használt $\Omega_{\mathbf{h}}$ rácshoz tartoznak, azonban $u(t,0)$ és $u(t,\pi)$ nem ismeretlenek.

Az egyenlet jobb oldalán levő differenciáloperátort ismét a fenti másodrendű véges differenciával közelítjük (amely most egy függvényekből álló vektorra vonatkozik), azaz

$$\begin{aligned}&[L_h(u(t,h), u(t,2h), \dots, u(t,\pi-h))]_k = \\ &\frac{1}{h^2}(u(t,(k+1)h) - 2u(t,kh) + u(t,(k-1)x_h)), \quad k = 1,2,\dots,N,\end{aligned}$$

és ezzel a (10.10) feladat megoldásának közelítésére a

$$\begin{cases} \partial_t u_k(t) = \sigma_D \frac{1}{h^2}(u_{k+1}(t) - 2u_k(t) + u_{k-1}(t,(k-1)h)), \quad k = 1,2,\dots,N \\ u_0(t) = u_{N+1}(t) = 0 \\ u_k(0) = \sin kh, \quad k = 1,2,\dots,N \end{cases} \tag{10.20}$$

közönséges differenciálegyenlet-rendszert kapjuk, ahol $u_k(t) \approx u(t,kh)$ (10.7. ábra).

A (10.20) rendszer felírható az

$$\begin{cases} \dot{\mathbf{u}}(t) = A\mathbf{u}(t) \\ \mathbf{u}(0) \quad \text{adott} \end{cases} \tag{10.21}$$

alakba is, ahol $A = \texttt{tridiag}[1,-2,1]$ mátrix és $\mathbf{u}(t) = (u_1(t), u_2(t), \dots u_n(t))$. Ezzel pedig

$$\mathbf{u}(t) = e^{At}\mathbf{u}(0), \tag{10.22}$$

amelyet ismét közelíthetünk numerikusan úgy, hogy (10.21) megoldására vagy (10.22) kiszámítására valamilyen numerikus módszert alkalmazunk.
Válasszuk a legkézenfekvőbbet: közelítsük (10.21) megoldását az explicit Euler-módszerrel! Ekkor

$$\mathbf{u}(t+\delta) = \mathbf{u}(t) + \delta A\mathbf{u}(t),$$

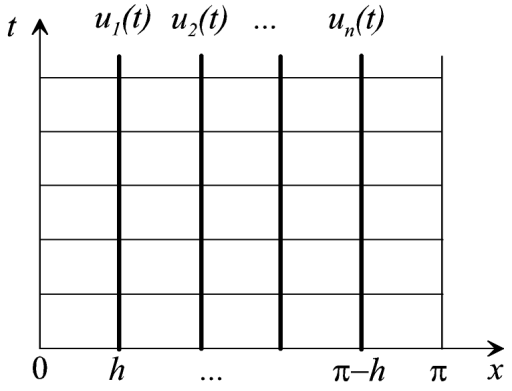
amelybe a (10.20) formulát helyettesítve az $\mathbf{u}(t+\delta)$ vektor $k = 1,2,\dots,N$ komponenseire

$$\mathbf{u}_k(t+\delta) = \mathbf{u}_k(t) + \delta\sigma_D\frac{1}{h^2}(u_{k+1}(t) - 2u_k(t) + u_{k-1}(t)),$$

valamint

$$\mathbf{u}(t+\delta, 0) = \mathbf{u}(t+\delta, (N+1)h) = 0,$$

ami pontosan ugyanazt a közelítést adja, mint amit a (10.15) formulában kaptunk. $\diamondsuit$



10.7. ábra. A szemidiszkretizáció után nyert $u_1, u_2, \ldots, u_n$ függvények értelmezési tartományai.

10.6. Függvényterek a közelítéshez

Különböző $\mathbf{h}$ diszkretizációs paraméterekre az $u_\mathbf{h}$ közelítések más pontokban adottak. Valamilyen struktúrába kellene ezeket beilleszteni, hogy konvergenciáról beszélhessünk majd. Sőt, a differenciáloperátorok közelítéséhez (ami a bevezető példa eszköze is volt) is szükségünk van egy megfelelő konvergenciafogalomra.

10.19. Definíció. Legyen $f : \Omega_\mathbf{h} \to \mathbb{R}$ függvény, $p \in \mathbb{R}^+$, továbbá jelölje

$$l_{\mathbf{h},p} = \{f_\mathbf{h} : \Omega_\mathbf{h} \to \mathbb{R} : \sum_{\mathbf{x}\in\Omega_\mathbf{h}} |f(\mathbf{x})|^p \text{ véges}\},$$

valamint

$$\|f_\mathbf{h}\|_{\mathbf{h},p} := |h_1 h_2 \cdot \cdots \cdot h_n| \left(\sum_{\mathbf{x}\in\Omega_\mathbf{h}} |f(\mathbf{x})|^p \right)^{\frac{1}{p}},$$

amely normát definiál a fenti vektortéren. Ezzel ellátva az $l_{\mathbf{h},p}$ vektorteret Banach-teret kapunk, amelyet szintén szintén $l_{\mathbf{h},p}$-vel jelölünk.
Ha $p = \infty$, akkor $l_{\mathbf{h},\infty} = \{f_\mathbf{h} : \Omega_\mathbf{h} \to \mathbb{R} : \sup_{\mathbf{x}\in\Omega_\mathbf{h}} |f(\mathbf{x})| \text{ véges}\}$, és ekkor legyen

$$\|f_\mathbf{h}\|_{\mathbf{h},\infty} := |h_1 h_2 \cdot \cdots \cdot h_n| \sup_{\mathbf{x}\in\Omega_\mathbf{h}} |f(\mathbf{x})|$$

és a kapott Banach-teret szintén $l_{\mathbf{h},\infty}$-vel jelöljük. ♢

A fenti norma az L_p norma „diszkrét verziója”, amit pontosabban a következő állításban fogalmazunk meg.

10.20. Lemma. *Legyen $f \in C(\overline{\Omega})$, és tekintsük az Ω halmaz egy finomodó $\Omega_\mathbf{h}$ felosztását. Ekkor minden $p \in (0, \infty]$ esetén* $\lim_{\mathbf{h}\to 0} \|f|_{\Omega_\mathbf{h}}\|_{\mathbf{h},p} = \|f\|_p$.

Bizonyítás. Először megjegyezzük, hogy $\overline{\Omega}$ korlátos, ezért az ott folytonos f maga is korlátos.
Terjesszük ki f-et arra az $\Omega_+ \supset \Omega$ halmazra azonosan nullaként, amely határának $\partial\Omega$-tól mért távolsága legalább h_0. A kiterjesztést jelölje f_+. A Riemann-integrál tulajdonságai alapján kapjuk, hogy $\int_\Omega f^p = \int_{\Omega_+} f_+^p$.
Tekintsünk minden $h < h_0$ esetén egy olyan $\mathcal{F}$ felosztást, amelyhez tartozó pontok az $\Omega_{\mathbf{h}}$-ra illeszkedő rács pontjai, és $\mathbf{h}$ méretű téglákból állnak. Nyilván f_+^p Riemann-integrálható, és f^p Riemann integrálja előáll a fenti felosztáshoz tartozó limeszként, ahol a felosztás finomsága $\mathbf{h}$, amely 0-hoz tart. Ekkor

$$\int_\Omega f^p = \int_{\Omega_+} f_+^p = \lim_{\mathbf{h}\to 0} |h_1 h_2 \cdot \dots \cdot h_n| \sum_{\mathbf{x}\in\mathcal{F}} |f_+(\mathbf{x})| = \lim_{\mathbf{h}\to 0} |h_1 h_2 \cdot \dots \cdot h_n| \sum_{\mathbf{x}\in\Omega_{\mathbf{h}}} |f_+(\mathbf{x})| =$$
$$= \lim_{\mathbf{h}\to 0} |h_1 h_2 \cdot \dots \cdot h_n| \sum_{\mathbf{x}\in\Omega_{\mathbf{h}}} |f(\mathbf{x})|,$$

ahogy azt állítottuk. □

A különböző $\|\cdot\|_{\mathbf{h},p}$ normák közti összefüggéssel kapcsolatban a 21.3. feladatra utalunk.

10.21. Megjegyzés. Az állítás nem igaz minden $f \in L_p(\Omega)$ függvényre. Ez is mutatja a véges differenciás közelítéseket használó elmélet korlátait. ♢

11. fejezet

A közelítő megoldás konvergenciája lineáris feladatok esetében

11.1. Konvergenciafogalmak a közelítésre

Először a pontonkénti konvergencia fogalmát adjuk meg.

11.1. Definíció. Azt mondjuk, hogy az u_h közelítés a t időpontban az $\mathbf{x}$ helyen konvergál az u megoldáshoz, ha

$$\lim_{\substack{n\delta\to t,\delta\to 0\\ \mathbf{k}\otimes\mathbf{h}\to\mathbf{x},\mathbf{h}\to 0}} u_{\mathbf{k}}^n = u(t,\mathbf{x}).$$

♢

Ez gyakran nem ad elegendő információt, hiszen többször az a cél, hogy az adott feladat megoldását egy t időpontban valamilyen (az alkalmazásokban például energiából származtatott) normában közelítsük.

11.2. Definíció. Azt mondjuk, hogy az u_h közelítés a t időpontban a $\|\cdot\|_{\mathbf{h},p}$ normában konvergál az u megoldáshoz, ha

$$\lim_{\substack{n\delta\to t,\delta\to 0\\ \mathbf{h}\to 0}} \|\mathbf{u}^n - \mathbf{u}(t,\cdot)\|_{\mathbf{h},p} = 0.$$

♢

Olyan feltételeket keresünk, amellyel a fenti értelemben vett konvergencia teljesül.

Jelölés: Jelölje megfelelő p esetén $Q_j : l_{\mathbf{h},p} \to l_{\mathbf{h},p}$ azt az operátort, amelyre $Q_j(\mathbf{u}^{j-1}) = \mathbf{u}^j$, ahol $\mathbf{u}^j$ a fenti sémával adott. Ezt a megfelelő sémához tartozó *j-edik lépésoperátornak* nevezzük.
Gyakran $Q_1 = Q_2 = \cdots = Q_N$, ekkor egyszerűen az összeset Q-val jelöljük, és *lépésoprátornak* nevezzük. Tipikusan ez az eset fordul elő, ha L homogén és állandó együtthatós. Természetesen a lépésoperátorok függnek $\mathbf{h}$-tól, azonban ezt az egyszerűség kedvéért nem jelöljük.

11.2. A konzisztencia és a konzisztenciarend fogalma

A konvergencia bizonyításához a séma pontosságának olyan definíciójára lesz szükség, amely az abban szereplő ∂_t és $\tilde{L}$ alakú (lásd a (10.4) formulát) differenciáloperátorokat együttesen tartalmazza. Ekkor az u függvény diszkretizációjához hasonlóan az

$$\mathbf{f}^n = \{f(n\delta, \mathbf{x}_h) : \mathbf{x}_h \text{ egy rácspont}\},$$

valamint ennek valamilyen közelítésére az $\hat{\mathbf{f}}^n$ jelölést is használjuk.

Nyilván szeretnénk, ha a kapott séma a differenciáloperátorok valamilyen pontos közelítéséből adódna; csak ekkor várhatjuk, hogy az ebből számolt közelítő megoldás pontos lesz.

11.3. Definíció. Azt mondjuk, hogy egy $0 = D_{\delta,\mathbf{h}}v$ séma pontonként konzisztens a $0 = \partial_t u - Lu$ egyenlettel, ha annak pontos u megoldására teljesül, hogy minden felosztásban tetszőleges (t, x) osztópont esetén

$$\lim_{\delta\to 0, \mathbf{h}\to 0} D_{\delta,\mathbf{h}} u(t, x) = 0.$$

♢

11.4. Megjegyzés.

1. Az eredeti differenciálegyenletben szereplő minden tagot, így például a nulladrendű forrástagot is közelítenünk kell, a fenti definíció ezt a követelményt is tartalmazza.

2. Természetesen az átrendezéssel kapott eredeti feladattal is konzisztensnek mondjuk a definícióban szereplő séma bármilyen átrendezettjét is. Nem vonatkozik ez azonban arra az esetre, ha mindkét oldalt szorozzuk valamilyen diszkretizációs paraméterrel. Emiatt nem is definiálunk konzisztenciarendet ebben az általános esetben.

3. A sémák jelölésénél sehol sem tüntetjük fel azok h-tól való függését. ♢

A konvergencia bizonyításához azonban ennél a természetesnek tűnő konzisztenciafogalomnál erősebb tulajdonságra van szükség, szeretnénk továbbá a konzisztencia rendjére is következtetni.

11.5. Definíció. Azt mondjuk, hogy az

$$\mathbf{u}^{n+1} = Q_{j+1}\mathbf{u}^n + \delta\hat{\mathbf{f}}^n \tag{11.1}$$

egylépéses séma a $\|\cdot\|_{\mathbf{h},p}$ normában $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_d)$ rendben konzisztens a $\partial_t u = L_0 u + f$ feladattal, ha annak u megoldására teljesül, hogy az L_0 operátorhoz tartozó minden Q_j lépésoperátorra

$$\mathbf{u}(j\delta, \cdot) = Q_j\mathbf{u}((j-1)\delta, \cdot) + \delta\hat{\mathbf{f}}^n + \delta\mathcal{R}_j(\delta\cdot), \tag{11.2}$$

ahol a j indextől, valamint h-tól és δ-tól független C konstanssal

$$\|\mathcal{R}_j(\delta)\|_{\mathbf{h},p} \leq C(\delta^{\alpha_0} + h_1^{\alpha_1} + \cdots + h_d^{\alpha_d}). \tag{11.3}$$

♢

A fenti (11.3) helyett egyszerűen a

$$\|\mathcal{R}_j(\delta)\|_{\mathbf{h},p} = \mathcal{O}(\delta^{\alpha_0}) + \mathcal{O}(h_1^{\alpha_1}) + \cdots + \mathcal{O}(h_d^{\alpha_d})$$

egyenlőséget fogjuk majd írni.

A gyakorlatban alkalmazott sémák rendszerint nem olyan egyszerű alakúak, mint a (11.1)-beli, másrészt a keresett u megoldást nem ismerjük, így azt be sem tudjuk helyettesíteni, hogy ellenőrizzük az ott szereplő feltételt.
Ezért egy általános konstrukciót ismertetünk, és megmutatjuk, hogy a sémának a kívánt konzisztenciarend eléréséhez milyen, egyszerűen ellenőrizhető követelményt kell teljesítenie.
Használjuk a (10.3) alakú egyenletben szereplő tagok következő közelítéseit:

$$\begin{aligned}
&\partial_t u(n\delta, \mathbf{k}\otimes\mathbf{h}) \approx \frac{u((n+1)\delta, \mathbf{k}\otimes\mathbf{h}) - u(n\delta, \mathbf{k}\otimes\mathbf{h})}{\delta}\\
&\tilde{L}u(n\delta, \mathbf{k}\otimes\mathbf{h}) \approx \mathcal{D}_{\tilde{L}}(\mathbf{u}(n\delta,\cdot), \mathbf{u}((n+1)\delta,\cdot))_{\mathbf{k}}\\
&f(n\delta, \mathbf{k}\otimes\mathbf{h}) \approx (\hat{\mathbf{f}}_{n,n+1})_{\mathbf{k}}\\
&g_j(n\delta, \mathbf{k}\otimes\mathbf{h}) \approx \hat{g}_j(n\delta, \mathbf{k}\otimes\mathbf{h}),\ j = 1, 2, \ldots, M\\
&\mathbf{D}u(n\delta, \mathbf{k}\otimes\mathbf{h}) \approx \mathcal{D}_{\mathbf{D}}(\mathbf{u}(n\delta,\cdot), \mathbf{u}((n+1)\delta,\cdot))_{\mathbf{k}},\ j = 1, 2, \ldots, M,
\end{aligned} \tag{11.4}$$

ahol $\mathcal{D}_{\tilde{L}}$ és $\mathcal{D}_{\mathbf{D}}$ lineáris operátorok.
A séma pontos felírásához kétféle rácspontot különböztetünk meg. Azon rácspontok halmazát, ahol az $\tilde{L}$ operátor közelítését alkalmazzuk, $\Omega_{\mathbf{h},b}$-vel jelöljük, és belső rácspontoknak nevezzük, továbbá használjuk az

$$\mathbf{u}_b^n = \{u_{\mathbf{k}}^n : \mathbf{k} \in \Omega_{\mathbf{h},b}\}$$

jelölést is. A többi rácspontot pedig $\Omega_{\mathbf{h},p}$-vel jelöljük, és perem rácspontoknak nevezzük, azaz

$$\Omega_{\mathbf{h}} = \Omega_{\mathbf{h},b} \cup \Omega_{\mathbf{h},p},$$

valamint a fentieknek megfelelően

$$\mathbf{u}_p^n = \{u_{\mathbf{k}}^n : \mathbf{k} \in \Omega_{\mathbf{h},p}\}.$$

11.6. Megjegyzés. A „perempontok” nem feltétlenül az Ω_h által megadott rács peremén helyezkednek el, pozíciójuk a sémától függ. A fejezet végén a korlátos tartományon adott sémákra vonatkozó példákban mutatjuk azt be konkrét esetekben. ♢

Ezek felhasználásával kapjuk az eredeti (10.3) alakú egyenletre vonatkozó

$$\begin{cases} \frac{\mathbf{u}_{\mathbf{k}}^{n+1}-\mathbf{u}_{\mathbf{k}}^{n}}{\delta} = \mathcal{D}_{\tilde{L}}(\mathbf{u}^n, \mathbf{u}^{n+1}) + \hat{\mathbf{f}}_{n,n+1} \quad \mathbf{k} \in \Omega_{\mathbf{h},b} \\ \hat{\mathbf{g}}_{\mathbf{k}}^{n+1} = \mathcal{D}_{\mathbf{D}}(\mathbf{u}^n, \mathbf{u}^{n+1})_{\mathbf{k}} = \mathcal{D}_{\mathbf{D}_1}\mathbf{u}_p^{n+1} + \mathcal{D}_{\mathbf{D}_2}\mathbf{u}_b^{n+1} + \mathcal{D}_{\mathbf{D}_3}\mathbf{u}^n \quad \mathbf{k} \in \Omega_{\mathbf{h},p}. \end{cases} \tag{11.5}$$

sémát, ahol a második sorban $\mathcal{D}_{\mathbf{D}_1}, \mathcal{D}_{\mathbf{D}_1}$, illetve $\mathcal{D}_{\mathbf{D}_3}$ jelöli a $\mathcal{D}_{\mathbf{D}}$ operátor $\mathbf{u}_p^{n+1}, \mathbf{u}_b^{n+1}$, illetve $\mathbf{u}^n$ típusú vektorok által generált altérre vett megszorítását.
Nyilvánvaló követelmény ezzel szemben az, hogy a perempontokban felvett értékeket meg lehessen határozni a peremfeltételek diszkretizációjának, azaz a (11.5) séma második sorában megadott peremfeltételek segítségével. Sőt, a peremfeltételben szereplő differenciáloperátor közelítése attól rendszerint nem változik, ha a felosztást sűrítjük. Ezzel összhangban használjuk a következő feltevést:

11.7. Feltevés. $\mathbf{u}_p^{n+1}$ kifejezhető $\mathbf{u}_b^{n+1}$ és $\mathbf{u}^n$ segítségével, azaz

$$\mathbf{u}_p^{n+1} = \mathcal{D}_{\mathbf{D}_1}^{-1}(\hat{\mathbf{g}}_{\mathbf{k}}^{n+1} - \mathcal{D}_{\mathbf{D}_2}\mathbf{u}_b^{n+1} - \mathcal{D}_{\mathbf{D}_3}\mathbf{u}^n), \tag{11.6}$$

továbbá $\|\mathcal{D}_{\mathbf{D}_1}^{-1}\mathcal{D}_{\mathbf{D}_2}\|_{\max}$ felülről korlátos. ♢

A feltevést használva átalakítjuk a kapott séma jobb oldalának első tagját a következőképpen:

$$\begin{aligned} \mathcal{D}_{\tilde{L}}(\mathbf{u}^n, \mathbf{u}^{n+1}) &= \mathcal{D}_{\tilde{L}}(\mathbf{u}^n, \mathbf{u}_b^{n+1}, \mathbf{u}_p^{n+1}) = \\ &= \mathcal{D}_{\tilde{L}}(\mathbf{u}^n, \mathbf{u}_b^{n+1}, \mathcal{D}_{\mathbf{D}_1}^{-1}(\hat{\mathbf{g}}_{\mathbf{k}}^{n+1} - \mathcal{D}_{\mathbf{D}_2}\mathbf{u}_b^{n+1} - \mathcal{D}_{\mathbf{D}_3}\mathbf{u}^n)). \end{aligned}$$

Itt a 11.7. feltevéssel összhangban a (11.5) séma második sorát is felhasználtuk, vagyis a megoldandó rendszer a következő:

$$\frac{\mathbf{u}_{\mathbf{k}}^{n+1} - \mathbf{u}_{\mathbf{k}}^{n}}{\delta} = \mathcal{D}_{\tilde{L}}(\mathbf{u}^n, \mathbf{u}_b^{n+1}, \mathcal{D}_{\mathbf{D}_1}^{-1}(\hat{\mathbf{g}}_{\mathbf{k}}^{n+1} - \mathcal{D}_{\mathbf{D}_2}\mathbf{u}_b^{n+1} - \mathcal{D}_{\mathbf{D}_3}\mathbf{u}^n)) + \hat{\mathbf{f}}_{n,n+1} \quad \mathbf{k} \in \Omega_{\mathbf{h},b}, \tag{11.7}$$

ahol az ismeretlenek a belső rácspontokban vett közelítő értékek. Természetesen csak akkor van értelme az eredeti séma használatának, ha az így kapott rendszer egyértelműen megoldható. Ezt rögzítjük a következő feltevésben:

11.8. Feltevés. A fenti (11.7) rendszer egyértelműen megoldható. ♢

A konzisztenciarend speciális esetekben

Különböző általános eseteket vizsgálunk, amikor a konzisztenciarend egyszerűen meghatározható.

Az első esetben legyen az eredeti séma

$$\begin{cases} \frac{\mathbf{u}_{\mathbf{k}}^{n+1}-\mathbf{u}_{\mathbf{k}}^{n}}{\delta} = (\mathcal{D}_{\tilde{L}}(\mathbf{u}^n))_{\mathbf{k}} + (\hat{\mathbf{f}}_{n,n+1})_{\mathbf{k}} \quad \mathbf{k} \in \Omega_{\mathbf{h},b} \\ \hat{g}_{\mathbf{k}} = (\mathcal{D}_{D_j}(\mathbf{u}^n))_{\mathbf{k}} \quad \mathbf{k} \in \Omega_{\mathbf{h},p} \end{cases} \tag{11.8}$$

alakú, vagyis minden explicit. Ekkor a 11.7. feltevésben szereplő kifejezésbe a pontos megoldást helyettesítve annak tetszőleges $\mathbf{k} \in \Omega_p$ indexű komponensére

$$\mathbf{u}(n\delta,\cdot)_{\mathbf{k}} = (B_n \mathbf{u}(n\delta,\cdot)_b)_{\mathbf{k}} + \mathcal{R}_0(h,\delta)_{\mathbf{k}} \tag{11.9}$$

alakú az $\mathcal{R}_0(h,\delta)$ hibatag-vektorral.

11.9. Állítás. *Tegyük fel, hogy a $\mathcal{D}_{\tilde{L}}$ véges differencia $\mathcal{O}(h^{\boldsymbol{\alpha}})$ rendű közelítése $\tilde{L}$-nek, valamint a peremfeltételek (11.9) numerikus közelítésében*

$$\mathcal{R}_0(h,\delta)_{\mathbf{k}} = \mathcal{O}(h_{s_{\mathbf{k}}}^{\beta_{\mathbf{k}}}),$$

továbbá, hogy $\mathbf{k} \in \Omega_p$ esetén az $\tilde{L}$ véges differenciában az $u_{\mathbf{k}}^n$ tagot $h_{s_{\mathbf{k}}}^{\gamma_{\mathbf{k}}}$-val osztottuk. Ekkor a fenti (11.8) sémából kapott módszer térváltozók szerinti konzisztenciarendje

$$\min\left\{\boldsymbol{\alpha}, \min_{\mathbf{k}\in\Omega_p} \beta_{\mathbf{k}} - \gamma_{\mathbf{k}}\right\}.$$

Bizonyítás. A (11.8) sémába a pontos megoldást beírva, majd az egyes közelítésekre vonakozó rendeket figyelembe véve kapjuk, hogy

$$\begin{aligned}
&\frac{u((n+1)\delta,\mathbf{k}\otimes\mathbf{h}) - u(n\delta,\mathbf{k}\otimes\mathbf{h})}{\delta} = \partial_t u(n\delta,\mathbf{k}\otimes\mathbf{h}) + \mathcal{R}_{1,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) = \\
&= \tilde{L}u(n\delta,\cdot)_{\mathbf{k}} + f(n\delta,\mathbf{k}\otimes\mathbf{h}) + \mathcal{R}_{1,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) = \\
&= \mathcal{D}_{\tilde{L}}(\mathbf{u}(n\delta,\cdot))_{\mathbf{k}} + \mathcal{R}_{2,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) + \\
&+ (\hat{f}_{n,n+1})_{\mathbf{k}} + \mathcal{R}_{3,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) + \mathcal{R}_{1,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) \\
&= \mathcal{D}_{\tilde{L}_1}(\mathbf{u}(n\delta,\cdot)_b)_{\mathbf{k}} + \mathcal{D}_{\tilde{L}_2}(\mathbf{u}(n\delta,\cdot)_p)_{\mathbf{k}} + \mathcal{R}_{2,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) + \\
&+ (\hat{f}_{n,n+1})_{\mathbf{k}} + \mathcal{R}_{3,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) + \mathcal{R}_{1,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) \\
&= \mathcal{D}_{\tilde{L}_1}(\mathbf{u}(n\delta,\cdot)_b)_{\mathbf{k}} + \mathcal{D}_{\tilde{L}_2}(B_n u(n\delta,\cdot)_b + \mathcal{R}_0(h,\delta))_{\mathbf{k}} + \mathcal{R}_{2,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) + \\
&+ (\hat{f}_{n,n+1})_{\mathbf{k}} + \mathcal{R}_{3,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) + \mathcal{R}_{1,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) \\
&= \left(\mathcal{D}_{\tilde{L}_1}(\mathbf{u}(n\delta,\cdot)_b) + \mathcal{D}_{\tilde{L}_2}(B_n u(n\delta,\cdot)_b) + \hat{f}_{n,n+1}\right)_{\mathbf{k}} + \\
&+ \mathcal{D}_{\tilde{L}_2}\mathcal{R}_0(h,\delta))_{\mathbf{k}} + \mathcal{R}_{1,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) + \mathcal{R}_{2,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) + \mathcal{R}_{3,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}),
\end{aligned}$$

amit átrendezve kapjuk, hogy a keresett konzisztenciarend az

$$\mathcal{D}_{\tilde{L}_2}\mathcal{R}_0(h,\delta)_{\mathbf{k}} + \mathcal{R}_{1,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) + \mathcal{R}_{2,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h}) + \mathcal{R}_{3,\mathbf{h},\delta}(n\delta,\mathbf{k}\otimes\mathbf{h})$$

hibatag nagyságrendje. A tételben szereplő feltevések miatt

$$\mathcal{R}_{2,\mathbf{h},\delta}(n\delta,\cdot)=\mathcal{R}_{3,\mathbf{h},\delta}(n\delta,\cdot)=\mathcal{O}(\mathbf{h}^{\boldsymbol{\alpha}}) \quad \text{és} \quad \mathcal{R}_{1,\mathbf{h},\delta}(n\delta,\cdot)=\mathcal{O}(\delta),$$

továbbá az

$$\mathcal{D}_{\tilde{L}_2}\mathcal{R}_0(h,\delta)_{\mathbf{k}}$$

tagot $h_{s_{\mathbf{k}}}^{\gamma_{s_{\mathbf{k}}}}$-val osztottuk. Ennek nagyságrendje tehát $h_{s_{\mathbf{k}}}^{\beta_{s_{\mathbf{k}}}-\gamma_{s_{\mathbf{k}}}}$, amiből azonnal adódik a tétel állítása. □

11.10. Példa. Tekintsük a

$$\begin{cases}\partial_t u(t,x)=\sigma_D\partial_{xx}u(t,x) & t\in\mathbb{R}^+,\ x\in(0,\pi)\\ \partial_x u(t,0)=\partial_x u(t,\pi)=0 & t\in\mathbb{R}^+\\ u(0,x)=\sin x & x\in(0,\pi).\end{cases} \tag{11.10}$$

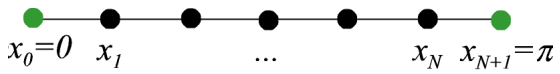
feladat közelítésére vonatkozó

$$\begin{cases}u_k^{n+1}=u_k^n+r(u_{k-1}^n-2u_k^n+u_{k+1}^n) & k=1,2,\dots,N\\ u_0^n=u_1^n,\ u_N^n=u_{N+1}^n,\\ u_k^0=\sin\frac{k\pi}{N+1} & k=0,1,\dots,N,N+1,\end{cases}$$

illetve az ezzel ekvivalens és a (11.8) feladatnak megfelelő alakú

$$\begin{cases}\frac{u_k^{n+1}-u_k^n}{\delta}=\frac{\sigma_D}{h^2}(u_{k-1}^n-2u_k^n+u_{k+1}^n) & k=1,2,\dots,N\\ u_0^n=u_1^n,\ u_N^n=u_{N+1}^n,\\ u_k^0=\sin\frac{k\pi}{N+1} & k=0,1,\dots,N,N+1\end{cases} \tag{11.11}$$

sémát (11.1. ábra)!
A feltételek nyilvánvalóan teljesülnek, hiszen a perempontok a (11.11) séma második sorában eleve a belső pontok segítségével adottak. Másrészt a perempontok eliminációjával kapott rendszer mátrixáról látható, hogy az nem szinguláris.



11.1. ábra. A belső- (fekete) és peremrácspontok (zöld) elhelyezkedése.

Itt a bal oldal közelítése δ szerint első rendű, a jobb oldal közelítése h szerint másod rendű, azaz a 11.9. állításban $\boldsymbol{\alpha}=\alpha_1=2$.

Továbbá a pontos megoldást (vagy akár egy tetszőleges u függvényt, amelyre a peremfeltételek teljesülnek) helyettesítve

$$u(n\delta, 0) = u(n\delta, h) - h \cdot \partial_x u(n\delta, 0) + \mathcal{R}(n\delta, h),$$

ahol $\mathcal{R}(n\delta, h) = \mathcal{O}(h^2)$, amennyiben $\partial_{xx} u$ korlátos. Ez hasonlóan érvényes a perem jobb oldalán is. Azaz a 11.9. állításban $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = 2$.
Világos továbbá, hogy a (11.11) képletben a 0 és $N+1$ indexű perempontokat a séma első sorába helyettesítve h^2-tel osztjuk (mint minden más tagot is). Emiatt $\gamma_1 = \gamma_2 = 2$, vagyis

$$\min\left\{\boldsymbol{\alpha}, \min_{\mathbf{k}\in\Omega_p} \beta_{\mathbf{k}} - \gamma_{\mathbf{k}}\right\} = \min\{2, \min\{0, 0\}\} = 0.$$

Tehát a módszer nem konzisztens. ♢

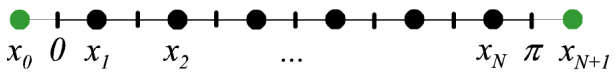
11.11. Példa. Tekintsük az

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x) \quad t \in \mathbb{R}^+,\ x \in (0,\pi) \\ \partial_x u(t,0) = 1,\ \partial_x u(t,\pi) = -1 \quad t \in \mathbb{R}^+ \\ u(0,x) = \sin x \quad x \in (0,\pi) \end{cases} \tag{11.12}$$

feladat közelítését a következő rácson:
A $(0,\pi)$ intervallumot N egyenlő részre osztjuk, mindegyik középpontjában kijelölünk egy rácspontot, ezek lesznek az $\{x_1, x_2, \ldots, x_N\}$ belső rácspontok, továbbá $x_0 = -\frac{\pi}{2N}$ és $x_{N+1} = \pi + \frac{\pi}{2N}$ a perem rácspontok (11.1. ábra). Erre vonatkozóan definiáljuk a (11.12) feladatra vonatkozó a (11.8) feladatnak megfelelő alakú

$$\begin{cases} \frac{u_k^{n+1} - u_k^n}{\delta} = \frac{\sigma_D}{h^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n) \quad k = 1, 2, \ldots, N \\ \frac{u_1^n - u_0^n}{h} = 1,\ \frac{u_{N+1}^n - u_N^n}{h} = -1 \\ u_k^0 = \sin\frac{k\pi}{N} - \frac{\pi}{2N} \quad k = 0, 1, \ldots, N, N+1 \end{cases} \tag{11.13}$$

sémát!
A feltételek nyilvánvalóan teljesülnek, hiszen a perempontok a (11.13) séma második sorában eleve a belső pontok segítségével adottak. Másrészt a perempontok eliminációjával kapott rendszer mátrixáról látható, hogy az nem szinguláris.



11.2. ábra. A belső- (fekete) és peremrácspontok (zöld) elhelyezkedése.

Itt a bal oldal közelítése δ szerint első rendű, a jobb oldal közelítése h szerint másod rendű, azaz a 11.9. állításban $\boldsymbol{\alpha} = \alpha_1 = 2$.

Továbbá a pontos megoldást (vagy akár egy tetszőleges u függvényt, amelyre a peremfeltételek teljesülnek) helyettesítve

$$u(n\delta, -\frac{\pi}{2N}) = u(n\delta, \frac{\pi}{2N}) - h \cdot \partial_x u(n\delta, 0) + \mathcal{R}(n\delta, h),$$

ahol $\mathcal{R}(n\delta, h) = \mathcal{O}(h^3)$, amennyiben $\partial_{xxx} u$ korlátos. Ez hasonlóan érvényes a perem jobb oldalán is. Azaz a 11.9. állításban $\beta_1 = \beta_2 = 3$.
Világos továbbá, hogy a (11.13) képletben a 0 és $N+1$ indexű perempontokat a séma első sorába helyettesítve h^2-tel osztjuk (mint minden más tagot is). Emiatt $\gamma_1 = \gamma_2 = 2$, vagyis

$$\min \left\{ \boldsymbol{\alpha}, \min_{\mathbf{k} \in \Omega_p} \beta_{\mathbf{k}} - \gamma_{\mathbf{k}} \right\} = \min \{2, \min\{1, 1\}\} = 1.$$

Tehát a módszer az x változó szerint első rendben konzisztens. ♢

További hasonló példák találhatók a 21.4. és a 21.5. feladatokban.

11.12. Megjegyzés.

1. Az inhomogén feladatokban megjelenő forrástagot is többféle módon építhetjük be a közelítésként felírt sémába. A konzisztenciarend szempontjából azonban nem mindegy, hogy ezt hogyan tesszük. Egy egyszerű példát találunk a 21.8. feladatban. A többdimenziós parabolikus feladatok vizsgálata esetén részletesen is foglalkozunk majd ezzel a kérdéssel.

2. Implicit sémák normában vett konzisztenciájának vizsgálata összetettebb lehet. Ilyen példa található a 21.9. feladatban.

3. Előfordulhat az is, hogy a konzisztencia csak akkor teljesül, ha a diszkretizációs paraméterekre valamilyen összefüggés áll fenn. Erre mutat rá a 21.10. feladat. ♢

A konzisztenciarend vizsgálata implicit módszerek esetében összetett lehet, lásd [29].

11.3. A stabilitás fogalma

Ez a vizsgált elméleten belül a legnehezebb fogalom. A benne szereplő feltétel ellenőrzésével külön fejezetben foglalkozunk. A további alkalmazásoknál ki fogjuk használni, hogy egylépéses sémákról van szó, tehát át kell majd fogalmaznunk mindezt, ha többlépéses sémákat vizsgálunk. Az itteni definíció nem magától értetődő, úgy mondjuk ki, hogy ennek ismeretében a konzisztens sémák konvergenciáját tudjuk igazolni.

11.13. Definíció. Azt mondjuk, hogy a

$$\begin{cases} \partial_t u = L_0 u \\ u(0,\cdot) \quad \text{adott} \end{cases}$$

feladat megoldására felírt séma a t időpontig feltétel nélkül stabil, ha létezik $\mathbf{h}_0, \delta_0$ és $K(t)$, hogy minden $\mathbf{h} < \mathbf{h}_0, \delta < \delta_0$ esetén és minden olyan n-re, amelyre $n\delta < t$, teljesül, hogy

$$\|\mathbf{u}^n\|_{\mathbf{h},p} \leq K(t)\|\mathbf{u}^0\|_{\mathbf{h},p}$$

valamilyen t-től függő K számmal. ♢

11.14. Megjegyzés.

1. Gyakran explicitebb alakban is megadják a K mennyiség függését - rendszerint exponenciális függést vesznek.

2. A definícióban szereplő feltétel ellenőrzése nem tűnik egyszerűnek első ránézésre, mert az egyenlőtlenségben szereplő n tetszőlegesen nagy lehet. Ha ugyanis $\delta \to 0$, akkor $n \to \infty$.

3. A definícióban szereplő feltételek közül a megfelelően kis δ_0 és h_0 választása rendszerint nem okoz problémát.

4. A definícióban az $\|\mathbf{u}^0\|_{\mathbf{h},p}$ norma helyett mindenhol egyszerűen az $\|\mathbf{u}^0\|_p$ normát vehetjük, mert az összehasonlítás során a $\mathbf{h}$ mennyiséggel úgyis egyszerűsítünk.

5. Ha a feladatban szereplő differenciáloperátor nem lenne lineáris, (hanem egy forrástag is szerepelne benn, amely nem homogén peremfeltételből is adódhat), akkor $\mathbf{u}(0,\cdot) = \mathbf{0}$ kezdeti feltétel esetén valamilyen t-re $\mathbf{u}(t,\cdot) \neq \mathbf{0}$ lehetne. Ekkor még a pontos megoldást megadó séma sem lehetne stabil, hiszen erre a t-re minden $K \in \mathbb{R}^+$ esetén $\|\mathbf{u}(t,\cdot)\|_{\mathbf{h},p} \geq K\|\mathbf{u}(0,\cdot)\|_{\mathbf{h},p}$. ♢

11.15. Definíció. Azt mondjuk, hogy egy séma feltételesen stabil a t időpontig, ha a 11.13. definíció feltételein kívül valamilyen $H : \mathbb{R}^+ \times \mathbb{R}^k$ függvényre $H(\delta, \mathbf{h}) < 0$ teljesül.♢

11.16. Megjegyzés.

1. Hasonlóan definiálhatjuk a feltételes exponenciális stabilitás fogalmát is.

2. Látni fogjuk majd az explicit sémák vizsgálata során, hogy azok feltételesen stabilak. A fenti H függvénnyel adott feltétel határozza meg a kapcsolatot a megfelelő $\mathbf{h}$ és δ paraméterek között.

3. A továbbiakban több állításban is egyszerűen stabilitást írunk. Ez minden esetben feltétel nélküli stabilitást jelent majd. ♢

11.17. Definíció. Azt mondjuk, hogy egy séma a t időpontig feltétel nélkül exponenciálisan stabil a $\|\cdot\|_{\mathbf{h},p}$ normában, ha van olyan $\beta \in \mathbb{R}$, hogy minden $j \in \mathbb{N}$ esetén (azaz minden lehetséges időlépéshez) a megfelelő Q_j operátorra teljesül, hogy

$$\|Q_j\| \leq e^{\beta\delta}. \tag{11.14}$$

♢

11.18. Példa. Vizsgáljuk meg a (10.9) közelítésére felírt

$$\begin{cases} u_k^{n+1} = u_k^n + \sigma_D \frac{\delta}{h^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n) = u_k^n + r(u_{k-1}^n - 2u_k^n + u_{k+1}^n), \quad k \in \mathbb{Z} \\ \mathbf{u}^0 \text{ adott} \end{cases}$$

séma $\|\cdot\|_{\mathbf{h},\infty}$ normában vett stabilitását abban az esetben, ha $r \leq \frac{1}{2}$ teljesül!

A sémában szereplő időlépés az

$$u_k^{n+1} = (1-2r)u_k^n + ru_{k-1}^n + ru_{k+1}^n,$$

alakba írható, ahol mindkét oldal szuprémumát véve k-ban majd a $0 < r \leq \frac{1}{2}$ feltételből kapott

$$|1-2r| + r + r = 1$$

egyenlőséget alkalmazva nyerjük az

$$\|\mathbf{u}^{n+1}\|_\infty \leq |1-2r| \cdot \|\mathbf{u}^n\|_\infty + r\|\mathbf{u}^n\|_\infty + r\|\mathbf{u}^n\|_\infty = \|\mathbf{u}^n\|_\infty$$

egyenlőtlenséget, amivel a fenti séma feltétel nélküli stabilitását igazoltuk. ♢

További példákat, stabilitási feltételeket a 3. fejezet végén látunk.
A fent bevezetett konzisztencia- és stabilitásfogalom mindenképp szükségesnek látszik, ha egy $Lu = f$ egyenletre vonatkozó feladat megoldását akarjuk numerikusan közelíteni. Ha ugyanis nem lenne az erre szolgáló séma konzisztens, akkor a pontos megoldásból kiindulva attól eltérő eredményt kapnánk. Ha pedig a séma nem lenne stabil, akkor a kis kezdeti hiba egy t időpontig tetszőlegesen naggyá válhat, ha elég sok időlépésben érjük el a közelítésben a t időpontot.
A következő fontos eredmény szerint a fenti két tulajdonság teljesülése elegendő is a konvergenciához. Természetesen a fenti szemléletes magyarázat nem bizonyítása ezen két tulajdonság szükségességének.

11.4. A Lax-féle konvergenciatétel

A tárgyalt anyagrész legfontosabb tételét mondjuk ki, amellyel a közelítő megoldások konvergenciáját lehet igazolni.

11.19. Tétel. *(Lax-féle ekvivalencia tétel.) Tegyük fel, hogy az $\partial_t u = L_0 u$ egyenlettel adott feladat megoldására felírt valamilyen $\mathbf{u}^{n+1} = Q_{n+1}\mathbf{u}^n$ séma a T időpontig exponenciálisan stabil, emellett a $\partial_t u = L_0 u + f$ egyenlettel adott feladat megoldására felírt*

$$\mathbf{u}^{j+1} = Q_{j+1}\mathbf{u}^j + \delta\hat{\mathbf{f}}^j \tag{11.15}$$

séma a $\|\cdot\|_{\mathbf{h},p}$ normára nézve normában $(1, \alpha_1, \ldots, \alpha_d)$ rendben konzisztens. Ekkor a (11.15) sémából kapott megoldás a $\|\cdot\|_{\mathbf{h},p}$ normában konvergens minden $t \leq T$ időpontban, és a konvergencia rendje megegyezik a fenti konzisztenciarenddel.

Bizonyítás. A bizonyítás lényege, hogy a hiba változását követjük nyomon, amelyet az

$$\mathbf{e}^j = \mathbf{u}^j - \mathbf{u}(j\delta, \cdot), \quad j = 1, 2, \ldots, N$$

egyenlőséggel definiálunk.
A (11.15) sémát és a konzisztenciára vonatkozó

$$\mathbf{u}((j+1)\delta, \cdot) = Q_{j+1}\mathbf{u}(j\delta, \cdot) + \delta\hat{\mathbf{f}}^j + \delta\mathcal{R}_{j+1}(\delta, \cdot)$$

egyenlőséget öszehasonlítva kapjuk, hogy

$$\mathbf{e}^{j+1} = Q_{j+1}\mathbf{e}^j + \delta\mathcal{R}_{j+1}(\delta, \cdot).$$

Innen teljes indukcióval egyszerűen nyerjük az

$$\mathbf{e}^n = Q_n Q_{n-1} \ldots Q_1 \mathbf{e}^0 + \delta Q_n Q_{n-1} \ldots Q_2 \mathcal{R}_1(\delta, \cdot) + \cdots + \delta Q_n \mathcal{R}_{n-1}(\delta, \cdot) + \delta\mathcal{R}_n(\delta, \cdot) \tag{11.16}$$

összefüggést, ahol a norma tagonkénti becsléséhez először megjegyezzük, hogy a stabilitás feltétele miatt a (11.14) formula alapján tetszőleges $j \in \mathbb{N}$ esetén

$$\|Q_n Q_{n-1} \ldots Q_j\| \leq e^{\beta n\delta} \leq e^{\beta t}.$$

Ennek felhasználásával a (11.16) egyenlőség következményeként kapjuk, hogy

$$\begin{aligned}
\|\mathbf{e}^n\|_{h,p} &\leq \|e^{\beta t}\|\|\mathbf{e}^0\|_{h,p} + \delta \sum_{j=1}^{n} e^{\beta n\delta}\|\mathcal{R}_j(\delta, \cdot)\|_{h,p} \leq \\
&\leq e^{\beta t}\|\mathbf{e}^0\|_{h,p} + n\delta e^{\beta t} \max_{j=1,2,\ldots,n-1} \|\mathcal{R}_j(\delta, \cdot)\|_{h,p} = \\
&= e^{\beta t}\|\mathbf{e}^0\|_{h,p} + t(\mathcal{O}(h_1^{\alpha_1}) + \cdots + \mathcal{O}(h_d^{\alpha_d}),
\end{aligned}$$

ahogy azt a tételben állítottuk. □

11.20. Megjegyzés.

1. A tétel igazából ekvivalenciát állít, de mi csak az egyik irányú állítással foglalkoztunk.

2. Csak lineáris differenciáloperátorokra használható. Ez a feltétel enyhíthető, de nem látszik, hogy a legtöbb nemlineáris probléma esetére hogyan lehetne a fenti elméletet alkalmazni.

3. A tétel értelmében elegendő az exponenciális stabilitás feltételét $\mathbf{u}^{n+1} = Q\mathbf{u}^n$ alakú sémákra ellenőrizni. Az elméletben ez a legnehezebb kérdés.

4. A normában vett konzisztencia bizonyításánál szükség volt a megoldás magas rendű deriváltjainak korlátosságára, ami korlátozó tényező lehet. ♢

Az elméleti bevezetéshez kapcsolódnak a [21], [28] és a [29] könyvek.

12. fejezet

Stabilitásvizsgálati módszerek

Először egy $\mathbb{R}$-en adott feladatokhoz tartozó módszert ismertetünk, a Neumann-féle stabilitásvizsgálatot az $l_{\mathbf{h},2}$ normában való exponenciális stabilitásra. Az itt tárgyalt elméletben az alább definiált sorozatok és függvények komplex értékűek lehetnek.

12.1. Diszkrét idejű Fourier-transzformáció, inverz transzformáció

Ha az L differenciáloperátor $\mathbb{R}$-en értelmezett, akkor a megoldás közelítése minden időpillanatban egy $\mathbf{v} = \ldots, v_{-1}, v_0, v_1, v_2, \ldots$ sorozat. Ennek Fourier-transzformáltját értelmezzük, ahol most l_2 téren a mindkét irányban végtelen és négyzetesen összegezhető sorozatokat értjük.

12.1. Definíció. (diszkrét idejű Fourier-transzformáció) Legyen az $\mathcal{F} : l_2 \to L_2[-\pi, \pi]$ az alábbi hozzárendeléssel értelmezett:

$$\mathcal{F}(\mathbf{u})(s) = \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} e^{-iks} u_k. \qquad \diamondsuit$$

Belátható, hogy erre teljesülnek az alábbiak:

- Az $\mathcal{F}$ transzformáció lineáris.
- Az $\mathcal{F}$ transzformáció izometria, azaz

$$\|\mathcal{F}(\mathbf{u})\|_2 = \|\mathbf{u}\|_2,$$

  ahol a két azonos $\|\cdot\|_2$ szimbólum más-más normát jelöl.

A következő állítás alapján egyszerűen megadható $\mathcal{F}$ inverze is:

12.2. Állítás. *Tekintsük az*

$$(\mathcal{F}^{-1}g)_k = \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi} e^{iks}g(s)\,\mathrm{d}s.$$

hozzárendeléssel adott $\mathcal{F}^{-1} : L_2[-\pi,\pi] \to l_2$ *függvényt. Ez valóban a fent definiált* $\mathcal{F}$ *inverze.*

12.1.1. Néhány nevezetes diszkrét idejű Fourier-transzformált

Olyan sorozatok diszkrét idejű Fourier-transzformáltját adjuk meg, amelyek véges differencia közelítésekben előfordulnak.

- Jobb oldali differenciával kapott sorozat esete:

$$\begin{aligned}\mathcal{F}(D_+\mathbf{v}) &= \frac{1}{\sqrt{2\pi}}\sum_{k=-\infty}^{\infty} e^{-iks}(v_{k+1}-v_k) = \\ &= \frac{1}{\sqrt{2\pi}}\sum_{k=-\infty}^{\infty} e^{is}e^{-i(k+1)s}v_{k+1} - \frac{1}{\sqrt{2\pi}}\sum_{k=-\infty}^{\infty} e^{-iks}v_k = (e^{is}-1)\mathcal{F}(\mathbf{v})\end{aligned}$$

- Bal oldali differenciával kapott sorozat esete:

$$\begin{aligned}\mathcal{F}(D_-\mathbf{v}) &= \frac{1}{\sqrt{2\pi}}\sum_{k=-\infty}^{\infty} e^{-iks}(v_k-v_{k-1}) = \\ &= \frac{1}{\sqrt{2\pi}}\sum_{k=-\infty}^{\infty} e^{-iks}v_k - \frac{1}{\sqrt{2\pi}}\sum_{k=-\infty}^{\infty} e^{-is}e^{-i(k-1)s}v_{k-1} = (1-e^{-is})\mathcal{F}(\mathbf{v})\end{aligned}$$

- Centrális differenciával kapott sorozat esete:

$$\begin{aligned}\mathcal{F}(D_0\mathbf{v}) &= \frac{1}{\sqrt{2\pi}}\sum_{k=-\infty}^{\infty} e^{-iks}(v_{k+1}-v_{k-1}) = \\ &= \frac{1}{\sqrt{2\pi}}\sum_{k=-\infty}^{\infty} e^{is}e^{-i(k+1)s}v_{k+1} - \frac{1}{\sqrt{2\pi}}\sum_{k=-\infty}^{\infty} e^{-is}e^{-i(k-1)s}v_{k-1} = \\ &= (e^{is}-e^{-is})\mathcal{F}(\mathbf{v}) = 2i\sin s\cdot\mathcal{F}(\mathbf{v}).\end{aligned}$$

- Centrális differencia másodrendű derivált közelítésére:

$$\begin{aligned}\mathcal{F}(D_0^2\mathbf{v}) &= \mathcal{F}(D_+\mathbf{v}-D_-\mathbf{v}) = ((e^{is}-1)-(1-e^{-is}))\mathcal{F}(\mathbf{v}) \\ &= (2\cos s-2)\mathcal{F}(\mathbf{v}) = -4\sin^2\frac{s}{2}\cdot\mathcal{F}(\mathbf{v})\end{aligned}$$

12.1.2. Alkalmazás a stabilitásvizsgálatban

Ebben a szakaszban mindenhol feltesszük, hogy a Q lépésoperátor minden lépésben azonos.

Jelölje most is $\mathbf{u}^n$ egy feladat megoldásának közelítését. Ekkor az alábbi hányadost fogjuk meghatározni:

$$\rho(s) = \frac{\mathcal{F}(\mathbf{u}^n)(s)}{\mathcal{F}(\mathbf{u}^{n-1})(s)},$$

ahol a $\rho : [-\pi, \pi] \to \mathbb{R}$ függvényt szorzófaktornak is nevezik.

12.3. Megjegyzés.

1. Természetesen ρ függ a sémától és a h paramétertől is, az egyszerűség kedvéért azonban ezt nem jelöljük.

2. A továbbiakban mindig feltesszük, hogy a h-tól, valamint a ρ-tól való függés folytonos.

3. A fenti formula ismételt alkalmazásával kapjuk azt is, hogy

$$\mathcal{F}(\mathbf{u}^n)(s) = \rho^n(s)\mathcal{F}(\mathbf{u}^0)(s). \tag{12.1}$$

♢

12.4. Állítás. *Ha itt $|\rho(s)| \leq 1$ minden $s \in [-\pi, \pi]$ esetén, akkor a megfelelő séma exponenciálisan stabil.*

Bizonyítás. Az $\mathcal{F}$ transzformáció tulajdonságai alapján és $|\rho(s)| \leq 1$ miatt ekkor nyilván

$$\|\mathbf{u}^n\|_2 = \|\mathcal{F}(\mathbf{u}^n)\|_2 \leq \|\mathcal{F}(\mathbf{u}^{n-1})\|_2 = \|\mathbf{u}^{n-1}\|_2,$$

vagyis ezt többször alkalmazva minden n-re

$$\|\mathbf{u}^n\|_2 \leq \|\mathbf{u}^0\|_2$$

teljesül, amiből a stabilitás azonnal következik. □

12.5. Megjegyzés. A fenti feltétel (és a belőle kapott norma-becslés) nagyon erős, igazoljuk majd ennek egy gyengítését. ♢

12.6. Példa. Vizsgáljuk a

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x), & x \in \mathbb{R},\ t \in \mathbb{R}^+ \\ u(0,x) = u_0(x), & x \in \mathbb{R} \end{cases}$$

feladat megoldására felírt

$$\begin{cases} u_k^0 = u_0(kh), \quad k = \ldots, -1, 0, 1, \ldots \\ u_k^{n+1} = u_k^n + \frac{\sigma_D \delta}{h^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n), \quad k = \ldots, -1, 0, 1, \ldots, \ n = 0, 1, 2, \ldots \end{cases}$$

séma stabilitását az l_2 norma szerint.

A szokásos $r = \frac{\sigma_D \delta}{h^2}$ jelöléssel a sémában szereplő egyenlőséget a

$$u_k^{n+1} = r u_{k-1}^n + (1-2r)u_k^n + r u_{k+1}^n$$

alakba írva kapjuk, hogy

$$\begin{aligned} \mathcal{F}(\mathbf{u}^{n+1})(s) &= \sum_{k=-\infty}^{\infty} e^{-iks} u_k^{n+1} \\ &= \sum_{k=-\infty}^{\infty} e^{-iks}(r u_{k-1}^n + (1-2r)u_k^n + r u_{k+1}^n) \\ &= \sum_{k=-\infty}^{\infty} e^{-is} e^{-i(k-1)s} r u_{k-1}^n + e^{-iks}(1-2r)u_k^n + e^{is} e^{-i(k+1)s} r u_{k+1}^n \\ &= r e^{-is} \mathcal{F}(\mathbf{u}^n)(s) + (1-2r)\mathcal{F}(\mathbf{u}^n)(s) + r e^{is} \mathcal{F}(\mathbf{u}^n)(s). \end{aligned}$$

Azaz ha $r \leq \frac{1}{2}$, akkor

$$\begin{aligned} |\mathcal{F}(\mathbf{u}^{n+1})(s)| &\leq |re^{-is}| \cdot |\mathcal{F}(\mathbf{u}^n)(s)| + |(1-2r)| \cdot |\mathcal{F}(\mathbf{u}^n)(s)| + |re^{is}| \cdot |\mathcal{F}(\mathbf{u}^n)(s)| \\ &= (|r| + |1-2r| + |r|)\mathcal{F}(\mathbf{u}^n)(s) = |\mathcal{F}(\mathbf{u}^n)(s)|, \end{aligned}$$

vagyis a 12.4. állítás értelmében a példában szereplő séma exponenciálisan stabil. ♢

12.7. Lemma. *A Q séma pontosan akkor exponenciálisan stabil, ha van olyan $\mathbf{h}_0, \delta_0$, hogy minden $\mathbf{h} \leq \mathbf{h}_0$ és $\delta \leq \delta_0$ esetén minden $s \in [-\pi, \pi]$-re teljesül a*

$$|\rho(s)|^{n+1} \leq K e^{\beta(n+1)\delta} \tag{12.2}$$

egyenlőtlenség.

Bizonyítás. A 12.4. állítás bizonyítását követve kapjuk, hogy ha a (12.2) egyenlőtlenség teljesül, akkor

$$\|\mathbf{u}^n\|_2 = \|\mathcal{F}(\mathbf{u}^n)\|_2 \leq e^{\beta\delta} \|\mathcal{F}(\mathbf{u}^{n-1})\|_2 = e^{\beta\delta} \|\mathbf{u}^{n-1}\|_2,$$

amiből nyilvánvalóan adódik az

$$\|\mathbf{u}^n\|_2 \leq e^{\beta t} \|\mathbf{u}^0\|_2$$

egyenlőtlenség, amellyel az egyik irányú következtetést igazoltuk.

Fordítva, ha nem volna igaz a (12.2) egyenlőtlenség, akkor minden $\beta \in \mathbb{R}$ esetén lenne olyan h_K és δ_K, hogy az azokhoz tartozó sémára

$$|\rho(s_k)|^n > e^{\beta\delta_K n_K} = e^{\beta t} \tag{12.3}$$

igaz valamilyen $s_k \in [-\pi, \pi]$ esetén. Ekkor ρ folytonossága miatt ez s_K-nak egy U_K alakú környezetében is igaz. Tekintsünk ekkor egy $v_K \in l_2$ vektort, amelyre $\operatorname{supp} \mathcal{F}(v_K) \subset U_K$. Ekkor a (12.1) formulát és a (12.3)-beli egyenlőtlenséget felhasználva

$$\|Q^n(v_K)\|_2^2 = \|\mathcal{F}(Q^n v_K)\|_2^2 = \|\rho^n(s)\mathcal{F}(v_K)\|_2^2 = \int_{U_K} \rho^{2n}(s) v_K^{2n}(s)\,\mathrm{d}s >$$
$$> e^{2\beta t} \int_{U_K} v_K^{2n}(s)\,\mathrm{d}s = e^{2\beta t}\|v_K\|_2^2,$$

ami ellentmond annak, hogy a Q időlépéshez tartozó séma stabil. □

A fenti állítás élesítését mondja ki a következő lemma.

12.8. Lemma. *Az $u^{n+1} = Qu^n$ séma pontosan akkor exponenciálisan stabil a $\|\cdot\|_2$ normára nézve, ha van olyan $C \in \mathbb{R}$, $\mathbf{h}_0$ és δ_0, hogy minden $\mathbf{h} \le \mathbf{h}_0$ és $\delta \le \delta_0$ esetén minden $s \in [-\pi, \pi]$-re teljesül a következő egyenlőtlenség:*

$$|\rho(s)| \le 1 + C\delta.$$

Bizonyítás. Ha a tételben említett feltétel teljesül, akkor nyilván

$$|\rho(s)| \le 1 + C\delta \le e^{C\delta}$$

miatt (amely *minden* $s \in [-\pi, \pi]$ esetén igaz)

$$\|\mathbf{u}^n\|_2 = \|\mathcal{F}\mathbf{u}^n\|_2 \le e^{C\delta}\|\mathcal{F}\mathbf{u}^{n-1}\|_2 \le \cdots \le e^{Cn\cdot\delta}\|\mathcal{F}\mathbf{u}^0\|_2 = e^{Ct}\|\mathbf{u}^0\|_2,$$

vagyis valóban exponenciálisan stabil a séma.

Fordítva, ha nem teljesül a tételben szereplő egyenlőtlenség, akkor minden C-re van olyan tetszőlegesen kis komponensekből álló (h, δ) pár, hogy valamilyen s-re

$$|\rho(s)| > 1 + C\delta.$$

Ha ez tetszőleges kis δ esetén elérhető, akkor a

$$\lim_{\delta\to 0}(1 + C\delta)^n = \lim_{n\to\infty}(1 + C\frac{t}{n})^n = e^{Ct}.$$

határérték miatt $\rho(s) > e^{\frac{C}{2}t}$, ami az előző állításnak ellentmond. □

12.9. Példa. Az előző példában vizsgált séma *pontosan akkor* stabil, ha $r \leq \frac{1}{2}$.

Azt láttuk, hogy $r \leq \frac{1}{2}$ esetén a stabilitás teljesül. Fordítva, ha $r = \frac{1}{2} + r_0$ állna fenn valamilyen $r_0 > \frac{1}{2}$ esetén, akkor

$$\begin{aligned} |\mathcal{F}(\mathbf{u}^{n+1})(s)| &= |re^{-is} \cdot \mathcal{F}(\mathbf{u}^n)(s) + (1-2r) \cdot \mathcal{F}(\mathbf{u}^n)(s) + re^{is} \cdot \mathcal{F}(\mathbf{u}^n)(s)| \\ &= |\mathcal{F}(\mathbf{u}^n)(s)(2r\cos s + 1 - 2r)|, \end{aligned}$$

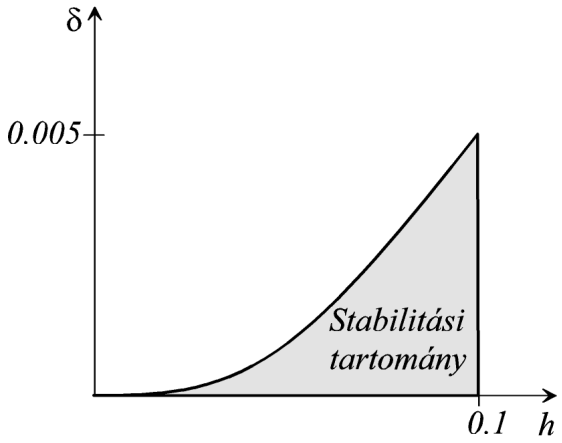
azaz

$$\rho(s) = 2r\cos s + 1 - 2r = -2r_0 + (1 + 2r_0)\cos s.$$

Ha $\cos s = -1$, akkor

$$|\rho(s)| = |-1 - 4r_0| = 1 + 4r_0,$$

amely 1-nél nagyobb és δ-tól független konstans, vagyis a 12.8. lemma miatt ebben az esetben nem lehet stabil a séma. A 11.19. tétel alapján az is nyilvánvaló, hogy a vizsgált sémából kapott numerikus megoldás pontosan akkor lesz konvergens, ha az $r \leq \frac{1}{2}$ feltétel teljesül (12.1. ábra). ♢



12.1. ábra. A stabilitási tartomány ($r \leq 1/2$) szemléltetése az $r = \delta/h^2$ rácsparaméter függvényében.

12.10. Állítás. *Ha az $u^{n+1} = Qu^n$ séma stabil a $\|\cdot\|_{\mathbf{h},2}$ normára nézve, akkor az $u^{n+1} = (Q + b\delta)u^n$ séma is az tetszőleges $b \in \mathbb{R}$ esetén.*

Bizonyítás. A 12.8. lemma alapján az $u^{n+1} = Qu^n$ stabilitása ekvivalens azzal, hogy van olyan $C \in \mathbb{R}$, $\mathbf{h}_0$ és δ_0, hogy minden $\mathbf{h} \leq \mathbf{h}_0$ és $\delta \leq \delta_0$ esetén minden $s \in [-\pi, \pi]$-re teljesül a

$$|\rho(s)| = \frac{\|\mathbf{u}^{n+1}\|_2}{\|\mathbf{u}^n\|_2} \leq 1 + C\delta.$$

egyenlőtlenség. Ekkor viszont a $\mathbf{u}^{n+1} = (Q + b\delta)\mathbf{u}^n$ séma esetén

$$\|\mathbf{u}^{n+1}\|_2 \leq \|\mathbf{u}^n\|_2(1 + C\delta + b\delta),$$

vagyis a 12.8. lemma felhasználásával kapjuk az állítást. □

12.11. Megjegyzés. A fenti állítás alapján ha egy $\partial_t u = L_0 u$ differenciáloperátor diszkretizációjából kapott séma stabil, akkor $\partial_t u = L_0 u + bu$-ből kapott is az, amennyiben b korlátos, és a bu mennyiséget az n-edik időlépésben „természetes módon", azaz bu^n-nel diszkretizáljuk.

12.12. Példa. Az előző példában kapott eredményt is felhasználva kapjuk, hogy a

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x) + bu(t,x), & x \in \mathbb{R},\ t \in \mathbb{R}^+ \\ u(0,x) = u_0(x), & x \in \mathbb{R} \end{cases}$$

feladat megoldására felírt

$$\begin{cases} u_k^0 = u_0(kh), & k = \ldots, -1, 0, 1, \ldots \\ u_k^{n+1} = u_k^n + \frac{\delta\sigma_D}{h^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n) + \delta u_k^n, & k = \ldots, -1, 0, 1, \ldots,\ n = 0, 1, 2, \ldots \end{cases}$$

séma pontosan akkor exponenciálisan stabil, ha $r \leq \frac{1}{2}$.

A 11.19. tétel alapján azt is kapjuk, hogy a séma pontosan akkor konvergens, ha $r \leq \frac{1}{2}$, emellett a konvergencia rendje $(1, 2)$. ♢

12.13. Példa. Vizsgáljuk meg az

$$\begin{cases} \partial_t u(t,x) + a\partial_x u(t,x) = 0, & t \in \mathbb{R}^+, x \in \mathbb{R} \\ u(0,x) = g(x), & x \in \mathbb{R} \end{cases} \tag{12.4}$$

feladat megoldására vonatkozó

$$\frac{u_k^{n+1} - u_k^n}{\delta} + a\frac{u_{k+1}^n - u_k^n}{h} = 0$$

közelítésből kapott

$$\begin{cases} u_k^{n+1} = u_k^n - \frac{\delta a}{h}(u_{k+1}^n - u_k^n) = (1 + R)u_k^n - Ru_{k+1}^n & n \in \mathbb{N},\ k \in \mathbb{Z} \\ u_k^0 = g(kh) & k \in \mathbb{Z} \end{cases} \tag{12.5}$$

séma stabilitását, ahol $R = \frac{\delta a}{h}$!

Mindkét oldal diszkrét idejű Fourier-transzformáltját véve nyerjük, hogy

$$\mathcal{F}(\mathbf{u}^{n+1}) = \mathcal{F}(\mathbf{u}^n) - \mathcal{F}(RD_+\mathbf{u}^n) = \mathcal{F}(\mathbf{u}^n)(1 - R(e^{is} - 1)) = \mathcal{F}(\mathbf{u}^n)(1 + R - Re^{is}),$$

vagyis

$$|\mathcal{F}(\mathbf{u}^{n+1})|^2 = |\mathcal{F}(\mathbf{u}^n)|^2((1+R)^2 - 2R(1+R)\cos s + R^2).$$

Itt tehát

$$|\rho(s)|^2 = (1+R)^2 - 2R(1+R)\cos s + R^2,$$

amelynek kiszámítjuk a maximumhelyét abban az esetben, ha $s \in [-\pi, \pi]$. Nyilván $s = 0, \pm\pi$ esetén lehet szélsőértéke, és ekkor

$$|\rho(\pi)|^2 = (1+R)^2 + 2R(1+R) + R^2 = (1+2R)^2, \quad \text{és} \quad |\rho(0)|^2 = 1.$$

Vagyis $|1+2R| < 1$ szükséges, tehát $-1 \le R \le 0$ a stabilitás szükséges elégséges feltétele, ami elégséges is, hiszen ekkor minden lehetséges s-re $|\rho(0)|^2 \le 1$.

Ez azt jelenti, hogy a csak negatív lehet, valamint

$$-\frac{1}{a} > \frac{h}{\delta} (> 0),$$

vagy az időlépés hosszára átfogalmazott feltétellel

$$(0 <)\delta < -ah. \qquad \diamondsuit$$

12.14. Példa. Vizsgáljuk meg az

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x), \quad x \in \mathbb{R},\ t \in \mathbb{R}^+ \\ u(0,x) = u_0(x), \quad x \in \mathbb{R} \end{cases}$$

feladat megoldására felírt

$$\begin{cases} u_k^0 = u_0(kh), \quad k = \ldots, -1, 0, 1, \ldots \\ u_k^{n+1} = u_k^n + \frac{\delta\sigma_D}{h^2}(u_{k-1}^{n+1} - 2u_k^{n+1} + u_{k+1}^{n+1}), \quad k = \ldots, -1, 0, 1, \ldots,\ n = 0, 1, 2, \ldots \end{cases} \tag{12.6}$$

implicit Euler-séma stabilitásának feltételét!

Az időlépést nyilvánvalóan a következő alakba tudjuk átírni:

$$\mathbf{u}^{n+1} = \mathbf{u}^n + rD_0^2\mathbf{u}^{n+1},$$

amelyre a diszkrét idejű Fourier-transzformáltat alkalmazva nyerjük, hogy

$$\mathcal{F}\mathbf{u}^{n+1} - r\mathcal{F}(D_0^2\mathbf{u}^{n+1}) = \mathcal{F}\mathbf{u}^n.$$

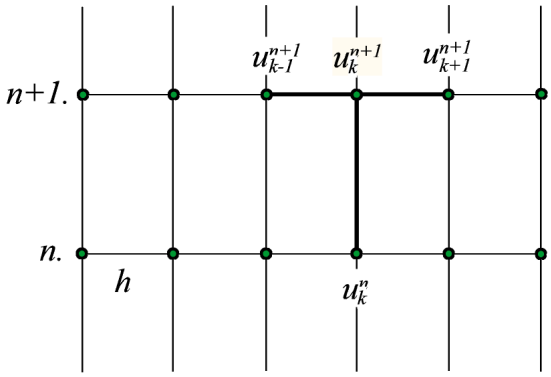
Ezt kifejtve kapjuk, hogy

$$\mathcal{F}\mathbf{u}^{n+1}(s)\left(1+4r\sin^2\frac{s}{2}\right)=\mathcal{F}\mathbf{u}^{n}(s),$$

azaz

$$\rho(s)=\frac{1}{1+4r\sin^2\frac{s}{2}},$$

tehát $0<\rho(s)\le 1$, vagyis a fenti séma feltétel nélkül stabil. ♢



12.2. ábra. Az implicit Euler-séma differenciacsillagja.

12.15. Példa. Vizsgáljuk meg az

$$\begin{cases}\partial_t u(t,x)=\sigma_D\partial_{xx}u(t,x), & x\in\mathbb{R},\ t\in\mathbb{R}^+\\ u(0,x)=u_0(x), & x\in\mathbb{R}\end{cases}$$

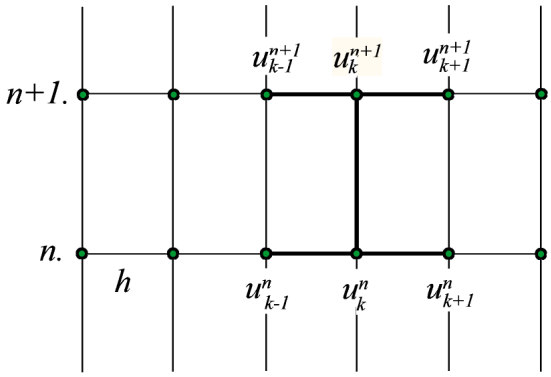
feladat megoldására felírt

$$\begin{cases}u_k^0=u_0(kh), & k=\ldots,-1,0,1,\ldots\\ u_k^{n+1}-\frac{\delta\sigma_D}{2h^2}(u_{k-1}^{n+1}-2u_k^{n+1}+u_{k+1}^{n+1})=u_k^n+\frac{\delta\sigma_D}{2h^2}(u_{k-1}^n-2u_k^n+u_{k+1}^n),\\ k=\ldots,-1,0,1,\ldots,\ n=0,1,2,\ldots\end{cases}\tag{12.7}$$

Crank–Nicolson-séma konzisztenciarendjét és stabilitását (12.3. ábra)!

Ezt nem számoljuk itt ki, mert magasabb dimenzióban részletesen elemezzük. Csak közöljük az eredményt, mely szerint a séma minden változó szerint másodrendben konzisztens, vagyis jobb, mint az explicit vagy az implicit Euler-séma, mert azok az időváltozó szerint csak elsőrendben konzisztensek. Másrészt a séma feltétel nélkül stabil; lásd a 21.21. feladatot! ♢

Ehhez kapcsolódnak a 20.2.5-20.2.6. animációk.



12.3. ábra. A Crank–Nicolson-séma differenciacsillagja.

12.2. A diszkrét idejű Fourier-transzformált több dimenziós esetben

A peremfeltétel nélkül adott sémák vizsgálatának fő eszköze most is a diszkrét idejű Fourier-transzformáció. Az egydimenziós esethez hasonlóan szükségünk lesz a

$$\mathcal{F} : l^d_{2,\mathbf{h}} \to L_2([-\pi,\pi]^d)$$

d-dimenziós Fourier-transzformáltra, amit az $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ és a $\mathbf{k} = (k_1, k_2, \ldots, k_d)$ jelölésekkel az

$$\mathcal{F}u(\alpha_1, \alpha_2, \ldots, \alpha_d) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \sum_{\mathbf{k}\in\mathbb{Z}^d} e^{-i\mathbf{k}\cdot\boldsymbol{\alpha}} u_{\mathbf{k}}$$

hozzárendeléssel definiálunk.

12.2.1. Nevezetes véges differenciákkal kapott mennyiségek Fourier-transzformáltja

Gyakran használjuk a stabilitásvizsgálat során a

$$\begin{aligned}\mathcal{F}(D^2_{+,x}\mathbf{u})(\boldsymbol{\alpha}) &= \sum_{\mathbf{k}\in\mathbb{Z}^d} e^{-i\mathbf{k}\cdot\boldsymbol{\alpha}}(u_{k_1+1,k_2,\ldots,k_d} - u_{\mathbf{k}}) = \\ &= \sum_{\mathbf{k}\in\mathbb{Z}^d} e^{i\alpha_1} e^{-i(k_1+1,k_2,\ldots,k_d)\cdot\boldsymbol{\alpha}} u_{(k_1+1,k_2,\ldots,k_d)} - \sum_{\mathbf{k}\in\mathbb{Z}^d} e^{-i\mathbf{k}\cdot\boldsymbol{\alpha}} u_{\mathbf{k}} = (e^{i\alpha_1} - 1)\mathcal{F}(\mathbf{u})(\boldsymbol{\alpha}),\end{aligned} \tag{12.8}$$

ahol x jelöli az első változót.
Hasonló módon kapható

$$\mathcal{F}(D_{-,x}\mathbf{u})(\boldsymbol{\alpha}) = (1 - e^{-i\alpha_1})\mathcal{F}(\mathbf{u})(\boldsymbol{\alpha}), \tag{12.9}$$

és a

$$\mathcal{F}(D^2_{0,x}\mathbf{u})(\boldsymbol{\alpha}) = -4\sin^2\frac{\alpha_1}{2}\mathcal{F}(\mathbf{u})(\boldsymbol{\alpha}) \tag{12.10}$$

azonosságokat, amelyekhez hasonlót felírhatunk a többi változó szerint vett véges differenciák esetére is.

Hasonlóan az egydimenziós esethez az időben állandó Q időlépéshez tartozó szorzófaktort a

$$\rho(\boldsymbol{\alpha}) = \frac{\mathcal{F}(\mathbf{u}^{n+1})(\boldsymbol{\alpha})}{\mathcal{F}(\mathbf{u}^{n})(\boldsymbol{\alpha})}$$

egyenlőséggel definiáljuk.
A stabilitás és a szorzófaktor kapcsolatára vonatkozólag az egydimenziós esethez hasonló állítások fogalmazhatók meg. Itt csak a legfontosabbat adjuk meg, a 12.7. lemma megfelelőjét, amelynek bizonyítására nem térünk ki, ez az egydimenziós esettel analóg módon történik.

12.16. Lemma. *A Q séma pontosan akkor exponenciálisan stabil, ha van olyan $\mathbf{h}_0, \delta_0$, hogy minden $\mathbf{h} \le \mathbf{h}_0$ és $\delta \le \delta_0$ esetén minden $\boldsymbol{\alpha} \in [-\pi, \pi]^d$-re teljesül a*

$$|\rho(\boldsymbol{\alpha})|^{n+1} \le Ke^{\beta(n+1)\delta} \tag{12.11}$$

egyenlőtlenség.

12.3. Korlátos tartományokon adott differenciáloperátorok diszkretizációjának (exponenciális) stabilitása

A továbbiakban feltesszük, hogy a séma csak $\mathbf{h}$ és δ egy függvényétől függ, továbbá ha ezen függvény állandó, akkor $\mathbf{h}$ növelésével δ is nő. A fentiekben vizsgált sémák is mind ilyen tulajdonságúak voltak, az $r = \frac{\sigma_D \delta}{h^2}$, illetve az $R = \frac{a\delta}{h}$ értéktől függtek, ahol állandó r, illetve R esetén kisebb h-hoz kisebb δ tartozik.

Mivel numerikus megoldást előállító Q időlépés operátornak most tartalmaznia kell a peremfeltételeket is, ez várhatóan összetettebb szerkezetű lesz, mint amilyeneket eddig vizsgáltunk. Az is világos, hogy mivel most véges sok pontban közelítjük a megoldást, Q egy mátrix. Sajnos, a Fourier-transzformációban lényeges, hogy azt „két irányban végtelen” sorozatokon értelmezzük, vagyis ezt a hasznos módszert ebben a szakaszban nem tudjuk közvetlenül alkalmazni. Az előzőekben kapott eredmények mégis hasznosak, amint a következő tétel mutatja. Az egyszerűség kedvéért az állandó együtthatós esettel foglalkozunk.

12.17. Tétel. *Legyen Q_h egy peremfeltételek nélkül adott (például egész $\mathbb{R}^n$-en definiált) differenciáloperátor sémájához tartozó időlépés operátor. Tegyük fel, hogy a séma olyan,*

hogy $u_{\mathbf{k}}^{n+1}$ értéke csak $u_{\mathbf{k}}^{n+1}$ és $u_{\mathbf{k}}^{n}$ rögzített K darab szomszédjától függ, azaz csakis a fentiektől az x_1 irányban legfeljebb k_1 távolságban, az x_2 irányban legfeljebb k_2 távolságban stb. levőn értékektől, ahol $K = 2(k_1 + \cdots + k_d) + 1$. Ha az ennek megfelelő séma nem stabil valamilyen t esetén, akkor bármilyen peremfeltétellel látjuk is el az eredeti feladatot, az abból kapott (a peremfeltételek figyelembevételével felírt) séma sem lehet stabil a t időpontig.

Bizonyítás. Ha az eredeti séma instabil valamilyen $\|\cdot\|$ normában, akkor tetszőleges C-hez van olyan n (amelyre $t = n\delta$) és olyan $\mathbf{u}^0$ vektor, hogy

$$\|\mathbf{u}^n\| = \|Q_{\mathbf{h}}^n(\mathbf{u}^0)\| \geq C\|\mathbf{u}^0\|. \tag{12.12}$$

Sőt, van olyan $\mathbf{u}^0$ is, amely véges sok komponens kivételével nulla és (12.12) teljesül rá. A bizonyításhoz azt kell észrevennünk, hogy egy olyan $\mathbf{u}^0$ -ra, amelynek a perem közelében található értékei nullák, pont ugyanúgy hat a peremfeltételek nélküli, mint a peremfeltételekkel megadott differenciáloperátor diszkretizált verziója is. Elég finom felosztás esetén tudjuk biztosítani, hogy $\mathbf{u}^1$ szintén ilyen legyen, aztán, hogy $\mathbf{u}^2$, ..., és ugyanígy $\mathbf{u}^n$ is.

Pontosabban fogalmazva tekintsünk egy olyan felosztáshoz tartozó (már a peremfeltételeket is tartalmazó) $Q_{\mathbf{h}_0}$ lineáris operátort, amelyre $Q_{\mathbf{h}_0}^n(\mathbf{u}^0) = Q_{\mathbf{h}}^n(\mathbf{u}^0)$.
Ilyen mindig van. Ha ugyanis a nemnulla elemek halmaza olyan, hogy j-edik koordinátájuk maximális különbsége N_j, valamint olyan véges differencia közelítést alkalmazunk, amely a j-edik irányban egy pont k_{1j}-gyel jobbra és k_{2j}-vel balra vett szomszédjaitól függ, akkor akkor a nemnulla elemek halmaza n lépés után olyan, hogy első koordinátájuk különbsége legfeljebb $N_j + nk_{1j} + nk_{2j}$. Vagyis ha egy olyan $Q_{\mathbf{h}_0}$ operátort tekintünk, amely minden irányban a $2 + \max_j N_j + (n+1) \cdot \max_j \{k_{1j} + k_{2j}\}$ felosztáshoz tartozik, abban pedig a fenti $\mathbf{u}^0$ kezdeti vektort tekintjük, amelynek nemnulla elemei a felosztás belsejében találhatók, akkor erre valóban $Q_{\mathbf{h}_0}^n(\mathbf{u}^0) = Q_{\mathbf{h}}^n(\mathbf{u}^0)$.
Ekkor viszont (12.12) szerint a $n\delta$ időpontban az instabilitást jelentő becslést kapjuk. Itt δ a fenti értéknél kisebb lesz, mert a térbeli diszkretizáció finomításával akkor kapjuk ugyanazon együtthatókkal rendelkező operátort, ha δ is csökken. Azaz már t előtt sem igaz semmilyen C-vel a stabilitást jelentő felső becslés. □

12.18. Megjegyzés. Ekkor szükséges feltételt nyerhetünk a stabilitásra a fenti Fourier-transzformációból kapott eredmények alapján. A tétel alkalmazását akkor tárgyaljuk, ha elégséges feltételeket is tudunk adni a stabilitásra. Előfordulhat az is, hogy a tételben szereplő fenti szükséges feltétel nem elégséges egyúttal. ♢

Olyan módszereket akarunk leírni, amelyekkel a peremfeltételekkel ellátott differenciáloperátorok diszkretizációjának exponenciális stabilitására vonatkozó szükséges és elégséges feltételek nyerhetők.
Az exponenciális stabilitás definíciója miatt azt kellene most is biztosítani, hogy létezzen

olyan $\mathbf{h}_0$ és δ_0, hogy valamilyen β esetén minden $\mathbf{h} \leq \mathbf{h}_0, \delta \leq \delta_0$ és $n \in \mathbb{N}$ értékre fennáll, hogy

$$\|Q^n\| \leq e^{\beta n \delta}. \tag{12.13}$$

12.3.1. A spektrálsugár és mátrixnormák kapcsolata

Emlékeztetünk, hogy tetszőleges $A \in \mathbb{R}^{n\times n}$ mátrixra az

$$\sigma(A) := \max\{|\lambda| : \lambda \text{ sajátértéke } A\text{-nak}\}$$

mennyiséget az A mátrix spektrálsugarának nevezzük.

Az alábbiakban felsorolunk a spektrálsugárra vonatkozó néhány fontos azonosságot, amelyet később felhasználunk.

- Fennáll, hogy
$$\sigma(A) \leq \|A\|_2,$$
valamint szimmetrikus mátrixokra egyenlőség áll.

- Teljesül továbbá, hogy
$$\sigma(AA^*) = \|A\|_2^2, \tag{12.14}$$

- valamint a spektrálsugárra igaz a
$$\sigma(A^n) = (\sigma(A))^n \tag{12.15}$$
egyenlőség, ahol a jobb oldalt az egyszerűség kedvéért $\sigma^n(A)$-nel jelöljük.

Az alábbi összefüggés is hasznos lesz.

12.19. Állítás. *Ha $H \in \mathbb{R}^{n\times n}$ invertálható, akkor tetszőleges $S \in \mathbb{R}^{n\times n}$ mátrixra HSH^{-1} és S sajátértékei megegyeznek, és így speciálisan $\sigma(S) = \sigma(HSH^{-1})$ is teljesül.*

12.3.2. Stabilitási feltételek

12.20. Lemma. *Ha Q szimmetrikus, és van olyan $\mathbf{h}_0, \delta_0$ és $C \in \mathbb{R}^+$, hogy minden $\mathbf{h} \leq \mathbf{h}_0$ és $\delta \leq \delta_0$ esetén teljesül az*

$$\sigma(Q) \leq 1 + C\delta$$

egyenlőtlenség, akkor a megfelelő séma stabil.
Másrészt tetszőleges Q mátrixra ez a feltétel szükséges is a stabilitáshoz.

Bizonyítás. Indirekt bizonyítunk. Nyilván teljesül, hogy

$$\|Q^n\|_2 \geq \sigma(Q^n) = (\sigma(Q))^n,$$

vagyis ha $\sigma(Q) \geq 1 + C\delta_j$ állna fenn valamilyen $\delta_j \to 0$ esetén, továbbá a $\|Q^n\|_2 \geq e^{Ct}$ becslés volna érvényes, azaz ha nincs olyan C, amely a lemmában említett tulajdonságnak megfelel, akkor valóban nem lehet exponenciálisan stabil a séma.
Részletesebben kifejtve: Világos, hogy ha δ az időlépés hossza, akkor $n = \left[\frac{t}{\delta}\right]$ lépést véve a t időpont előtt leszünk. Az indirekt állítás szerint minden C-hez és minden $(\delta_0, \mathbf{h}_0)$ párhoz van olyan $\mathbf{h} < \mathbf{h}_0$ és $\delta < \delta_0$, hogy $|\sigma(Q)| > 1 + C\delta$, vagyis

$$\sigma\left(Q^{\left[\frac{t}{\delta}\right]}\right) = (\sigma(Q))^{\left[\frac{t}{\delta}\right]} > (1 + C\delta)^{\left[\frac{t}{\delta}\right]} > \frac{1}{1 + C\delta}(1 + C\delta)^{\frac{t}{\delta}}. \tag{12.16}$$

Tudjuk továbbá, hogy

$$\lim_{\delta \to 0} \frac{1}{1 + C\delta} = 1 \quad \text{és} \quad \lim_{\delta \to 0}(1 + C\delta)^{\frac{t}{\delta}} = e^{Ct},$$

amit a (12.16) egyenlőtlenségre alkalmazva kapjuk, hogy

$$\sup_{t^* < t} \|Q(t^*)\| \geq e^{Ct},$$

ahol $Q(t^*)$ jelöli a t^* időponthoz tartozó lépésmátrixot. Ez ellentmond a stabilitás tényének. Ha Q szimmetrikus, akkor az $1 + \delta C \leq e^{Ct}$ egyenlőtlenség miatt $n\delta < t$ esetén

$$\|Q^n\|_2 = \sigma(Q^n) = (\sigma(Q))^n \leq (1 + C\delta)^n \leq e^{Cn\delta} \leq e^{Ct},$$

ami éppen a t időpontig érvényes exponenciális stabilitást jelenti. □

Mivel többször kell nem szimmetrikus mátrixokkal is számolnunk, az alábbi eredmény is fontos lesz.

12.21. Állítás. *Ha Q olyan, hogy valamilyen H invertálható mátrixszal HQH^{-1} szimmetrikus, akkor a 12.20. lemmában szereplő feltétel a Q-val adott sémára vonatkozólag is szükséges és elégséges.*

Bizonyítás. A bizonyításban a $\|Q^n\|_2$ normát becsüljük felülről:

$$\begin{aligned}\|Q^n\|_2 = \|H^{-1}(HQH^{-1})^n H\|_2 \leq \|H^{-1}\|_2\|H\|_2\|(HQH^{-1})^n\|_2 = \\ = \|H^{-1}\|_2\|H\|_2[\sigma(HQH^{-1})]^n \leq \|H^{-1}\|_2\|H\|_2 e^{Cn\delta},\end{aligned}$$

ami bizonyítja az állítást. □

12.22. Példa. Egyszerűen látható, hogy a fenti $\sigma(Q) \leq 1 + C\delta$ feltétel valóban csak szükséges, általában nem elégséges. Ehhez legyen $a < 0$, és tekintsük a

$$\begin{cases} \partial_t u(t,x) + a\partial_x u(t,x) = 0 \quad t \in \mathbb{R}^+,\ x \in (0,1) \\ u(0,x) = u_0(x), \quad x \in (0,1) \\ u(t,1) = 0 \quad t \in \mathbb{R}^+ \end{cases} \tag{12.17}$$

feladatot! Megjegyezzük, hogy ez korrekt kitűzésű.
Diszkretizáljuk ezt azon a rácson, amely a $(0,1)$ intervallum olyan egyenletes felosztásából adódik, ahol a belső rácspontok indexei $1, 2, \ldots, N$. Tekintsük a differenciáloperátor egy diszkretizációjából kapott

$$u_k^{n+1} = (1+R)u_k^n - Ru_{k+1}^n, \quad k = 1, 2, \ldots, N \tag{12.18}$$

sémát! Ez a feltételek mellett eleve csak akkor lesz kiszámítható, ha $a < 0$, hiszen ekkor tudjuk u_N^{n+1} értékét a peremfeltétel segítségével kifejezni.
Emlékeztetünk, hogy a (12.4) feladat stabilitásvizsgálatából azt kaptuk, hogy a peremfeltételek nélküli (az egész $\mathbb{R}$-en értelmezett) feladathoz tartozó séma stabilitásának $-1 \leq R \leq 0$ szükséges és elégséges feltétele. A 12.17. tétel alapján ez most is szükséges.
Ekkor a séma a $Q =$ `tridiag [0,1+R,-R]` mátrixszal adható meg, azaz $\mathbf{u}^{n+1} = Q\mathbf{u}^n$. Itt a spektrálsugár $1 + R$, ahol

$$|1 + R| \leq 1 \Leftrightarrow 0 > R > -2.$$

A fenti feltétel alapján látjuk, hogy a spektrálsugár vizsgálatából nem kaptunk elégséges feltételt. Azt azonban ez a vizsgálat sem tisztázta, hogy a (12.17) feladat esetére mi lesz egy szükséges és elégséges feltétel. Ezt tesszük majd meg a 12.30. példában. ♢

12.4. A Gersgorin-tétel és alkalmazása a stabilitásvizsgálatban

A fentiekben láttuk, hogy az $\mathbf{u}^{n+1} = Q\mathbf{u}^n$ séma stabilitásának ellenőrzéséhez a Q lépésmátrix sajátértékeit kell megvizsgálnunk. Ehhez felidézzük a Gersgorin - tételt:

12.23. Tétel. *A Q mátrix minden λ sajátértékéhez van olyan j, amelyre*

$$|\lambda - q_{jj}| \leq \sum_{k \neq j} |q_{jk}|.$$

Mivel a spektrálsugár csak szimmetrikus mátrixokra ad normát, és itt éppen a $\|\cdot\|_2$-es normával egyezik meg, úgy tűnhet, hogy a fenti tétel csak olyan sémák vizsgálatakor hasznos, amelyekhez tartozó lépésmátrixok szimmetrikusak. Azonban a módszerrel további esetekben is elégséges feltételt nyerhetünk stabilitásra. Ebben a szakaszban a Q mátrix i-edik sorának j-edik elemét q_{ij}-vel jelöljük.

12.24. Állítás. *Ha egy sémához tartozó Q lépésmátrix szimmetrikus, továbbá minden lehetséges j-re $\sum_i |q_{ij}| \leq 1$, akkor $\|Q\|_2 \leq 1$.*

Bizonyítás. Ha a $\sum_i |q_{ij}| \leq 1$ feltétel teljesül, akkor biztosan minden j-re

$$q_{jj} + \sum_{i \neq j} |q_{ij}| \leq 1.$$

Ha $q_{jj} \geq 0$, akkor az előzőt -1-gyel szorozva kapjuk, hogy

$$q_{jj} - \sum_{i \neq j} |q_{ij}| \geq -q_{jj} - \sum_{i \neq j} |q_{ij}| \geq -1,$$

vagyis az előző sorral összevetve

$$[q_{jj} - \sum_{i \neq j} |q_{ij}|, q_{jj} + \sum_{i \neq j} |q_{ij}|] \subset [-1, 1].$$

Ha pedig $q_{jj} \leq 0$, akkor

$$1 \geq \sum_i |q_{ij}| = -q_{jj} + \sum_{i \neq j} |q_{ij}| \geq q_{jj} + \sum_{i \neq j} |q_{ij}|,$$

amelynek első felét -1-gyel szorozva

$$-1 \leq q_{jj} - \sum_{i \neq j} |q_{ij}|,$$

vagyis az utolsó két formulát összevetve ismét

$$[q_{jj} - \sum_{i \neq j} |q_{ij}|, q_{jj} + \sum_{i \neq j} |q_{ij}|] \subset [-1, 1].$$

Vagyis kaptuk, hogy a Gersgorin-tétel miatt Q összes sajátértéke a $[-1, 1]$ intervallumban van, tehát $\sigma(Q) \leq 1$, így mivel szimmetrikus is, valóban $\|Q\|_2 \leq 1$ is teljesül, ahogy állítottuk. □

12.25. Állítás. *Ha egy sémához tartozó Q lépésmátrix minden i sorában $\sum_j |q_{ij}| \leq 1$, és ugyanígy minden j oszlopában $\sum_i |q_{ij}| \leq 1$, akkor minden $n \in \mathbb{N}$ esetén $\|Q^n\|_2 \leq 1$, vagyis a megfelelő séma $\|\cdot\|_2$ normában exponenciálisan stabil.*

Bizonyítás. Kiszámítjuk a $\|Q\|_2 = \rho(Q^*Q) = \rho(QQ^*)$ mennyiséget. A bizonyításban $Q[i,\cdot]$ jelöli az Q mátrix i-edik sorát (ami egy n hosszú vektor), és $|Q[i,\cdot]|$ azt a vektort, amely az elemenként vett abszolútértékekből áll. Tudjuk, hogy

$$QQ^*[i,j] = Q[i,\cdot]\cdot Q^*[\cdot,j] = Q[i,\cdot]\cdot Q[j,\cdot],$$

vagyis a QQ^* szimmetrikus mátrix i-edik sorában az elemek abszolútértékeinek összege:

$$\sum_j |QQ^*[i,j]| = \sum_j |Q[i,\cdot]\cdot Q[j,\cdot]| \le \sum_j \left|\sum_k Q[i,k]Q[j,k]\right| \le$$
$$\le \sum_k \sum_j |Q[i,k]||Q[j,k]| = \sum_k |Q[i,k]| \sum_j |Q[j,k]| \le \sum_k |Q[i,k]|\cdot 1 \le 1.$$

tehát a 12.24. állítás alapján

$$1 \ge s(QQ^*) = \|Q\|_2^2,$$

amiből kapjuk, hogy

$$\|Q^n\|_2 \le \|Q\|_2^n \le 1,$$

ahogy azt állítottuk. □

12.26. Megjegyzés.

1. Itt fontos volt, hogy a Gersgorin-tételből kapott feltételt, nem pedig egyszerűen a spektrálsugarat vizsgáltuk. Mivel azonban a mátrix nem szimmetrikus, mind a sorok, mind az oszlopok tekintetében meg kell követelnünk a feltételt.

2. A feltételben nem szerepel a mátrix mérete, a becslés attól függetlenül teljesül. A stabilitásvizsgálatban ez a legfontosabb és a legnehezebb részlet is.

3. A 12.25. állítás azért hasznos, mert nem szimmetrikus mátrixokra csupán a Gersgorin tételt alkalmazva azoknak csak a spektrálsugarára kapunk becslést, a stabilitásvizsgálathoz azonban a $\|\cdot\|_2^2$-es normát kell becsülni. ♢

A fenti tételek feltételei bizonyos értelemben nem gyengíthetők. Erre mutat rá az alábbi két példa, ahol

$$A = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 & \dots & 0 \\ \frac{1}{2} & \frac{1}{3} & 0 & 0 & \dots & 0 \\ \vdots & & & & & \vdots \\ \frac{1}{2} & \dots & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & \dots & 0 & 0 & 0 & \frac{1}{3} \end{pmatrix} \qquad B = \begin{pmatrix} \frac{1}{5} & 2 & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{5} & 2 & 0 & \dots & 0 \\ \vdots & & & & & \vdots \\ 0 & \dots & 0 & 0 & \frac{1}{5} & 2 \\ 0 & \dots & 0 & 0 & 0 & \frac{1}{5} \end{pmatrix}$$

Az A mátrix nem szinguláris, a sorokban az elemek abszolútértékeinek összege legfeljebb $\frac{5}{6}$, azonban ha $n \times n$-es méretűről van szó, és $\mathbf{v} = [1, 0, 0, \ldots, 0]^T$, akkor $A\mathbf{v} = \left[\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}\right]^T$, vagyis $\|A\|_2^2 \geq \|A\mathbf{v}\|^2 = \frac{n}{4}$, azaz már $\|A\|_2$ értékére sem lehet n-től független korlátot megadni.
Azonban olyan mátrixok, mint A, nem szoktak lépésmátrixként előfordulni.
Ha a B mátrixot tekintjük, annak spektrálsugara $\frac{1}{5}$, emellett

$$B^2[1,3] = B^2[2,4] = \cdots = 4, \text{ és } B^3[1,4] = B^3[2,5] = \cdots = 8, \text{ stb.},$$

és amíg $k + j \leq n$, addig $B^j[k, k+j] = 2^j$. Azaz könnyen látható, hogy B hatványainak normájára nem adható n-től független felső korlát.

12.27. Megjegyzés. Áramlási feladatok megoldásánál olyan $n \times n$-es lépésmátrixok is előfordulnak, amelyek „majdnem antiszimmetrikusak”, azaz minden $1 \leq j, k \leq n$ index és $j \neq k$ esetén $a_{j,k} = -a_{k,j}$ Mivel itt k-adik oszlop főátlón kívüli elemei épp a k-adik sor elemeinek ellentettjei, ezért a fenti tételt alkalmazva itt egyszerűen a Gersgorin-tétel által adott feltételt elég ellenőrizni a sajátértékek abszolútértékére vonatkozólag. ♢

12.28. Példa. A fentieket alkalmazzuk arra az esetre is, amikor

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x) \quad t \in \mathbb{R}^+, \ x \in (a,b) \\ u(t,a) = u(t,b) = 0 \quad t \in \mathbb{R}^+ \\ u(0,x) = u_0(x) \quad x \in (a,b), \end{cases} \tag{12.19}$$

ahol $a, b \in \mathbb{R}$, $u_0 \in C(a,b)$ adottak, valamint azt a sémát használjuk, amit a 10.5 fejezetben konstruáltunk.

Egyrészt a 12.17. tétel alapján, valamint az $\mathbb{R}$-en adott feladat stabilitására vonatkozó feltétel alapján tudjuk, hogy ahhoz az $r \leq \frac{1}{2}$ feltétel szükséges.

Másrészt könnyen látható, hogy az időlépés operátor mátrix alakja

$$Q = \texttt{tridiag [r, 1-2r, r]},$$

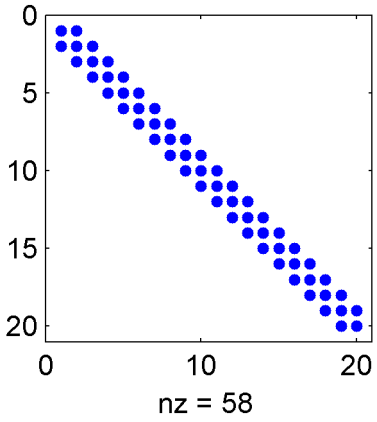
amely nyilván szimmetrikus (12.4. ábra). A spektrálsugárra vonatkozó (Gersgorin-tételből kapott) becslés alapján és $r > 0$ felhasználásával kapjuk, hogy

$$\rho(Q) \in [1 - 2r - 2r, 1 - 2r + 2r] = [1 - 4r, 1], \qquad \diamondsuit$$

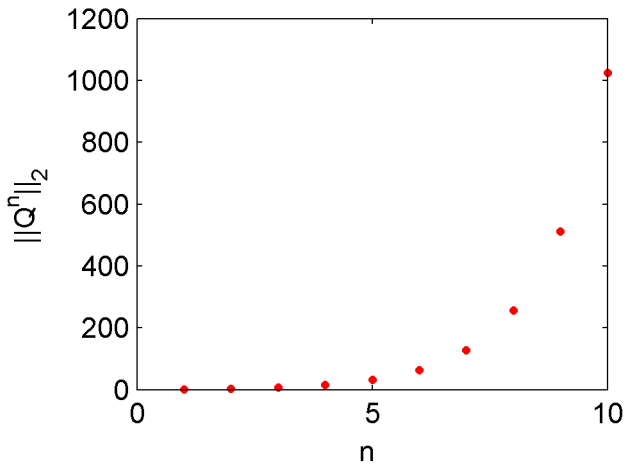
vagyis ha $r \leq \frac{1}{2}$, akkor $\rho(Q) \in [-1, 1]$.

Azaz kaptuk, hogy az $r \leq \frac{1}{2}$ feltétel elégséges is a Q-val adott séma stabilitásához. A 12.5. és a 12.6. ábrák azt az esetet szemléltetik, amikor a stabilitási feltétel nem teljesül. Az első ábra a Q mátrix hatványainak normabeli növekedését mutatja, a másik pedig a numerikus megoldás viselkedését. A 12.7. ábrán a Q mátrix hatványainak normabeli csökkenése látható a stabilitási feltétel teljesülése esetén.

Hasonló példák találhatók a 21.24., a 21.25. és a 21.27. feladatokban.



12.4. ábra. A Q tridiagonális lépésmátrix nemnulla elemeinek elhelyezkedése abban az esetben, ha a mátrix mérete 20×20-as. A mátrix összesen 58 nemnulla elemet tartalmaz.

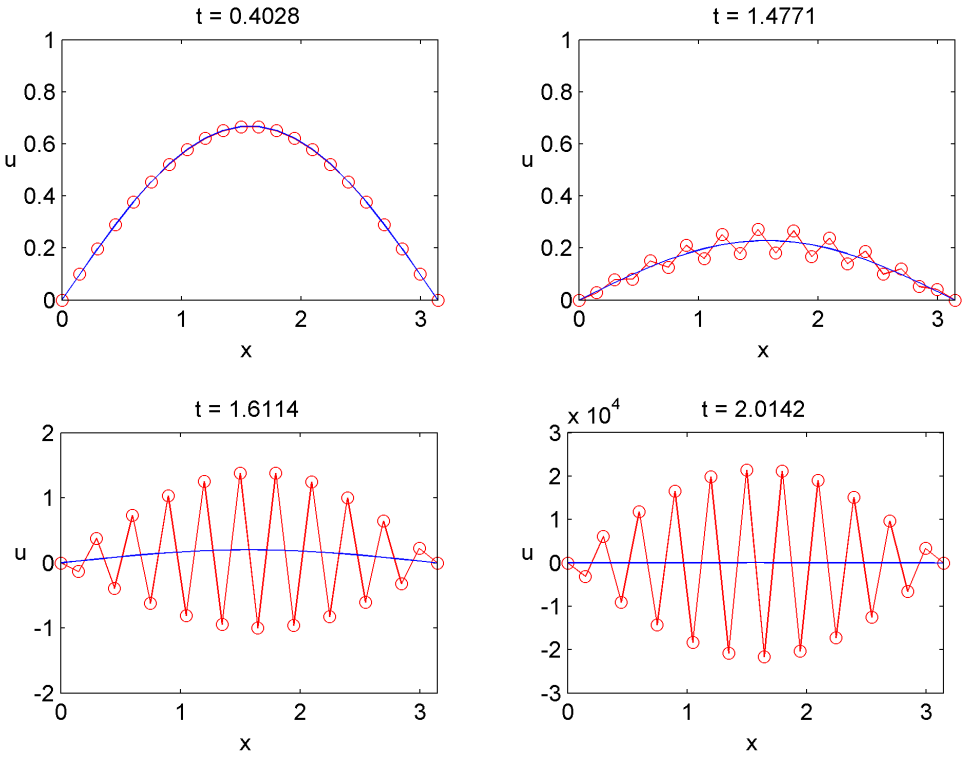


12.5. ábra. A $\|Q^n\|_2$ normák értéke n függvényében, ha $r = \delta/h^2 = 1$ (instabil eset), $n\delta = 1$. Q az explicit Euler-módszer lépésmátrixa.

12.29. Példa. Vizsgáljuk meg a (12.19) feladatra felírt

$$\begin{cases} u_k^0 = u_0(kh), \quad k = \ldots, -1, 0, 1, \ldots \\ u_k^{n+1} = u_k^n + \frac{\delta\sigma_D}{h^2}(u_{k-1}^{n+1} - 2u_k^{n+1} + u_{k+1}^{n+1}), \quad k = 1, 2, \ldots, N,\ n = 0, 1, 2, \ldots \\ u_0^{n+1} = u_{N+1}^{n+1} = 0 \end{cases} \tag{12.20}$$

implicit Euler-séma stabilitásának feltételét!



12.6. ábra. A diffúziós egyenlet pontos (kék) és az explicit Euler-módszerrel nyert numerikus megoldása $h = \pi/21$, $r = 0.6$ választással a 30., 110., 120. és 150. időrétegen ($a = 0$, $b = \pi$, $\sigma_D = 1$, $u_0(x) = \sin x$).

A séma átrendezésével kapjuk, hogy az

$$\tilde{Q}\mathbf{u}^{n+1} = \mathbf{u}^n$$

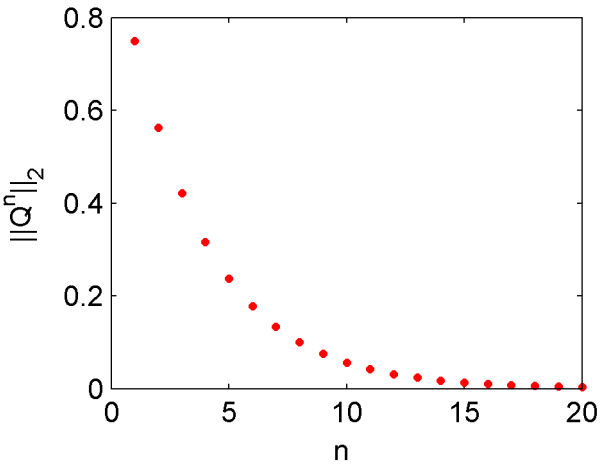
alakú, ahol $\tilde{Q} = \texttt{tridiag [r, 1+2r, r]}$.
Ekkor

$$\mathbf{u}^{n+1} = \tilde{Q}^{-1}\mathbf{u}^n,$$

ahol a $\tilde{Q}^{-1}$ lépésmátrix is szimmetrikus. A Gersgorin-tétel alapján tudjuk, hogy $\tilde{Q}$ sajátértékei az $[1, 1+4r]$ intervallumba esnek, vagyis az inverz sajátértékei a $[0, 1]$ intervallumba. Tehát a (12.20) séma feltétel nélkül stabil. ♢

12.30. Példa. Vizsgáljuk meg a (12.17) feladatra felírt (12.18) séma stabilitását!

Tudjuk egyrészt, hogy a peremfeltételek nélkül adott esetben pontosan akkor stabil a séma, ha $-1 \leq R < 0$, ezért ez a feltétel mindenképp szükséges a most vizsgált esetben is. Másrészt belátjuk, hogy elégséges is. Tudjuk, hogy a lépésmátrix $Q =$



12.7. ábra. A $\|Q^n\|_2$ normák értéke n függvényében, ha $r = \delta/h^2 = 1/4$ (stabil eset), $n\delta = 1$. Q az explicit Euler-módszer lépésmátrixa.

`tridiag [0,1+R,-R]` alakú. A feltétel miatt $|1 + R| = 1 + R$, valamint $| - R| = -R$, tehát a sorokban, illetve oszlopokban levő abszolútérték-összeg

$$0 \leq |1 + R| = 1 + R < 1 \quad \text{vagy} \quad 0 \leq |1 + R| + | - R| = 1,$$

tehát a 12.25. állítás szerint valóban stabil lesz a megfelelő séma. ♢

13. fejezet

Parabolikus egyenletek 1, 2 és 3 dimenzióban

Ebben a fejezetben a

$$\partial_t u(t,\mathbf{x}) = \sigma_{D,x_1}\partial_{11}u(t,\mathbf{x}) + \sigma_{D,x_2}\partial_{22}u(t,\mathbf{x}) + \cdots + \sigma_{D,x_d}\partial_{dd}u(t,\mathbf{x}) + F(t,\mathbf{x}) \qquad (13.1)$$

alakú feladatok numerikus megoldásával foglalkozunk, ezekre vonatkozó sémákat vizsgálunk.
Minden esetben rögzítjük az

$$u_0 = u(0,\cdot)$$

kezdeti feltételt, emellett ha a feladat korlátos tartományon adott, akkor a stabilitásvizsgálathoz homogén peremfeltételeket veszünk. A magasabb dimenziós sémák konzisztenciavizsgálatához alkalmazni fogjuk a következő lemma eredményét, amelyet nem bizonyítunk. Az egyszerűség kedvéért azt a négy állítást fogalmazzuk meg, amelyeknek eredményét ténylegesen felhasználjuk.

13.1. Lemma. *Legyen $u : (0,T)\times\Omega \to \mathbb{R}$ minden változója szerint 4-szer deriválható!*

1. *Ekkor $d = 2$ és $d = 3$ esetén az $\frac{1}{h_x^2}D_{0,x}^2\frac{1}{h_y^2}D_{0,y}^2$ véges differencia $\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)$ rendben közelíti a $\partial_{xx}\partial_{yy}$ negyedrendű parciális deriváltat.*
2. *Az is teljesül, hogy az $\frac{1}{h_x^2}D_{0,x}^2\frac{1}{h_y^2}D_{0,y}^2\frac{1}{\delta}D_{+,t}$ véges differencia $\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)+\mathcal{O}(\delta)$ rendben közelíti a $\partial_{xx}\partial_{yy}\partial_t$ ötödrendű parciális deriváltat.*
3. *Hasonlóan, $d=3$ esetén az $\frac{1}{h_x^2}D_{0,x}^2\frac{1}{h_y^2}D_{0,y}^2\frac{1}{h_z^2}D_{0,z}^2$ véges differencia $\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)+\mathcal{O}(h_z^2)$ rendben közelíti a $\partial_{xx}\partial_{yy}\partial_{zz}$ hatodrendű parciális deriváltat.*
4. *Szintén hasonlóan adódik az is, hogy $d=3$ esetén az $\frac{1}{h_x^2}D_{0,x}^2\frac{1}{h_y^2}D_{0,y}^2\frac{1}{h_z^2}D_{0,z}^2\frac{1}{\delta}D_{0,t}$ véges differencia $\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)+\mathcal{O}(h_z^2)\mathcal{O}(\delta^2)$ rendben közelíti a $\partial_{xx}\partial_{yy}\partial_{zz}\partial_\delta$ hetedrendű parciális deriváltat.*

5. *A hiperbolikus feladatok vizsgálatához használjuk még, hogy az* $\frac{1}{2h_x}D_{0,x}\frac{1}{2h_y}D_{0,y}\frac{1}{\delta}D_{+,t}$ *véges differencia* $\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)+\mathcal{O}(\delta)$ *rendben közelíti a* $\partial_{xx}\partial_{yy}\partial_t$ *ötödrendű parciális deriváltat.*

13.1. Az egydimenziós esetre kapott eredmények összefoglalása

Az alábbi táblázatban foglaljuk össze röviden, hogy milyen eredményeket kaptunk az egydimenziós eset elemzésekor.

13.1. táblázat. Három különböző séma tulajdonságai az egy dimenziós parabolikus feladat megoldásának numerikus közelítésére (peremfeltétel nélküli eset)

módszer	komplexitás	stabilitás feltétele	rend időben	rend térben
explicit Euler	explicit	$r \leq \frac{1}{2}$	1	2
implicit Euler	implicit	-	1	2
Crank–Nicolson	implicit	-	2	2

13.2. A kétdimenziós esetre vonatkozó sémák vizsgálata

Ebben a szakaszban a (13.1) formulában szereplő egyenlet

$$\partial_t u(t,x,y) = \sigma_{D,x}\partial_{xx}u(t,x,y) + \sigma_{D,y}\partial_{yy}u(t,x,y) \tag{13.2}$$

kétdimenziós verziójával foglalkozunk. Egyszerűen igazolható, hogy a következő

$$u_{j,k}^{n+1} = u_{j,k}^n + \delta\left(\frac{\sigma_{D,x}}{h_x^2}D_{0,x}^2u_{j,k}^n + \frac{\sigma_{D,y}}{h_y^2}D_{0,y}^2u_{j,k}^n\right) = u_{j,k}^n + \left(r_xD_{0,x}^2u_{j,k}^n + r_yD_{0,y}^2u_{j,k}^n\right) \tag{13.3}$$

explicit séma, ahol az $r_x = \delta\frac{\sigma_{D,x}}{h_x^2}$ és az $r_y = \delta\frac{\sigma_{D,y}}{h_y^2}$ jelöléseket alkalmaztuk, a (13.2) egyenlettel konzisztens. A konzisztencia rendje az x és y változók szerint 2, t szerint pedig 1.

13.2. Állítás. *A (13.3) séma stabilitásának szükséges és elégséges feltétele, hogy* $r_x + r_y \leq \frac{1}{2}$ *legyen. Hasonlóan, ha a megfelelő feladatban homogén Dirichlet-peremfeltételt alkalmazunk, és ennek megfelelően a sémát a peremfeltételek megadásával egészítjük ki, a fenti feltétel akkor is szükséges és elégséges.*

Bizonyítás. Fourier-transzformációt alkalmazunk. A (12.10) formula szerint

$$\begin{aligned}(\mathcal{F}\mathbf{u}^{n+1})(\alpha_1,\alpha_2) &= (\mathcal{F}\mathbf{u}^n)(\alpha_1,\alpha_2) + \delta\mathcal{F}(D^2_{0,x}u^n_{j,k} + D^2_{0,y}u^n_{j,k})(\alpha_1,\alpha_2)\\ &= (1 - 4r_x\sin^2\frac{\alpha_1}{2} - 4r_y\sin^2\frac{\alpha_2}{2}), \mathcal{F}\mathbf{u}^n(\alpha_1,\alpha_2),\end{aligned}$$

ahol a szorzófaktor biztosan legfeljebb 1, és pontosan akkor lesz minden (α_1,α_2)-ra legalább -1, ha $r_x + r_y \leq \frac{1}{2}$. Ezzel igazoltuk az állítást a peremfeltételt nem tartalmazó esetre.

Az állítás második felének igazolásához látnunk kell, milyen szerkezetű a homogén Dirichlet-peremfeltételhez tartozó lépésmátrix. Legyen egy téglalap felosztásában n_x belső osztópont x irányban, és n_y darab y irányban. Az ismeretleneket tartalmazó vektor ekkor

$$\mathbf{u} = [u_{1,1}, u_{2,1}, \ldots, u_{n_x,1}, u_{1,2}, u_{2,2}, \ldots, u_{n_x,2}, \ldots, \ldots, , u_{1,n_y}, u_{2,n_y}, \ldots, u_{n_x,n_y}]$$

alakú. Világos, hogy a (13.3) formula alapján a keresett mátrix főátlójába minden esetben $1 - 2r_x - 2r_y$ kerül. Másrészt az $\mathbf{u}$ vektorban vele szomszédos elemek r_x-szeresét adjuk hozzá, amennyiben nem az első, az n_x-edik, az $n_x + 1$-edik, $2n_x$-edik, stb. elemekről van szó, mert ekkor a homogén peremfeltételek felhasználásával csak a jobb, illetve a bal oldali szomszédot adjuk hozzá. Vagyis az eddigiek figyelembevételével a lépésmátrix azon részét határoztuk meg, amely $\texttt{diag}(B)$ alakú (blokkdiagonális) mátrix, ahol $B = \texttt{tridiag}[r_x, 1 - 2r_x - 2r_y, r_x]$ egy ilyen $n_x \times n_x$-es blokk.
A homogén peremfeltételek felhasználásával kapjuk, hogy $k = 2, 3, \ldots, n_x - 1$ esetén

$$u^{n+1}_{k,1} = u^n_{k,1} + r_x(u^n_{k-1,1} - 2u^n_{k,1} + u^n_{k+1,1}) + r_y(0 - 2u^n_{k,1} + u^n_{k,2}),$$

valamint

$$u^{n+1}_{1,1} = u^n_{1,1} + r_x(0 - 2u^n_{1,1} + u^n_{2,1}) + r_y(0 - 2u^n_{1,1} + u^n_{1,2}),$$

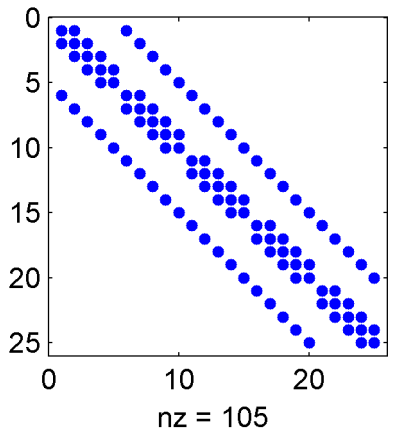
és

$$u^{n+1}_{nx,1} = u^n_{n_x,1} + r_x(u^n_{n_x-1,1} - 2u^n_{n_x,1} + 0) + r_y(0 - 2u^n_{n_x,1} + u^n_{n_x,2}).$$

Vagyis a következő időlépésben szereplő $\mathbf{u}^{n+1}$ vektor első n_x darab elemét úgy kapjuk meg, hogy a fenti $\texttt{diag}(B)$ mátrixszal való szorzáson kívül a vektor n_x-szel jobbra levő elemeinek r_y-szorosát adjuk hozzá. Sőt, ha vannak ilyenek az $\mathbf{u}^n$ vektorban, akkor mind a n_x-szel jobbra levő, mind az n_x-szel balra levő elemek r_y-szorosát hozzá kell adni. Vagyis a $\texttt{diag}(B)$ mátrixhoz hozzá kell adni két olyan mellékátlót, amelyek a főátlótól n_x-szel jobbra, és ugyanennyivel balra vannak. Összességében a lépésmátrix az alábbi alakú $n_x \times n_y$-os méretű *szimmetrikus* mátrix:

$$\begin{pmatrix} B & r_yI & 0 & 0 & \ldots & 0\\ r_yI & B & r_yI & 0 & \ldots & 0\\ 0 & r_yI & B & r_yI & \ldots & 0\\ \vdots & & & & & \vdots\\ 0 & \ldots & 0 & r_yI & B & r_yI\\ 0 & \ldots & 0 & 0 & r_yI & B \end{pmatrix},$$

ahol az összes blokk $n_x \times n_x$-es, a blokkok száma pedig n_y (13.1. ábra).



13.1. ábra. A nemnulla elemek elhelyezkedése a lépésmátrixban $n_x = n_y = 5$ esetén.

Itt minden sorban a főátlóban $1 - 2r_x - 2r_y$ áll, a mellette levő elemek abszolút értékeinek összege pedig legfeljebb $2r_x + 2r_y$, azaz a Gersgorin-tétel értelmében a $1 - 4r_x - 4r_y \geq -1$ feltételnek kell teljesülnie, azaz $\frac{1}{2} \geq r_x + r_y$ valóban elégséges feltétele a stabilitásnak, amivel a fenti állítást beláttuk. □

Hasonló példa szerepel a 21.26. feladatban, lásd továbbá a 20.2.7-20.2.9. animációkat!

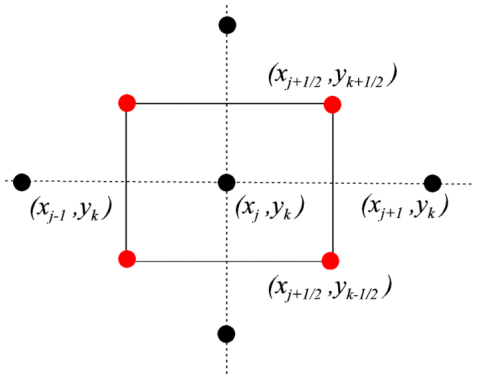
13.2.1. Egy Crank–Nicolson-típusú séma a kétdimenziós esetre

A következő levezetésben egy véges térfogatrészre vonatkozó megmaradási törvényt használunk, amely szerint egy térfogatrészen az anyagmennyiség össz megváltozása egyenlő a beáramlott össz anyagmennyiséggel. Ezt természetesen teljesíti az egyenlet pontos megoldása. Tudjuk még, hogy a beáramló anyagmennyiség (fluxusa) a Fick-törvény szerint a gradiens ellentettjével arányos. Mindezt a felosztás egy $(x_j - \frac{h_x}{2}, x_j + \frac{h_x}{2}) \times (y_k - \frac{h_y}{2}, y_k + \frac{h_y}{2}) = I_{j,k}$ részére (13.2. ábra) alkalmazunk tetszőleges t időpontban:

$$\partial_t \int_{y_k - \frac{h_y}{2}}^{y_k + \frac{h_y}{2}} \int_{x_j - \frac{h_x}{2}}^{x_j + \frac{h_x}{2}} u(t,x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\partial I_{j,k}} \boldsymbol{\nu} \cdot \nabla u(t,x,y)\,\mathrm{d}x\,\mathrm{d}y. \tag{13.4}$$

A formulák egyszerűsítése érdekében használjuk továbbá az

$$x_j \pm \frac{h_x}{2} = x_{j\pm\frac{1}{2}} \quad \text{és} \quad y_k \pm \frac{h_y}{2} = y_{k\pm\frac{1}{2}}$$



13.2. ábra. A levezetésben használt véges térfogatrész.

jelöléseket. (13.4) mindkét oldalát t_n és t_{n+1} között integrálva a bal oldalon a Newton-Leibniz-formulát alkalmazva, a jobb oldalon pedig a Neumann-féle peremfeltételt kifejtve nyerjük, hogy

$$\begin{aligned}
&\int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t_{n+1},x,y)\,\mathrm{d}x\,\mathrm{d}y - \int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t_n,x,y)\,\mathrm{d}x\,\mathrm{d}y = \\
&= -\int_{t_n}^{t_{n+1}} \int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \boldsymbol{\nu}\cdot\nabla u(t,x_{j-\frac{1}{2}},y)\,\mathrm{d}y\,\mathrm{d}t + \int_{t_n}^{t_{n+1}} \int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \boldsymbol{\nu}\cdot\nabla u(t,x_{j-\frac{1}{2}},y)\,\mathrm{d}y\,\mathrm{d}t - \\
&- \int_{t_n}^{t_{n+1}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \boldsymbol{\nu}\cdot\nabla u(t,x,y_{k-\frac{1}{2}})\,\mathrm{d}x\,\mathrm{d}t + \int_{t_n}^{t_{n+1}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \boldsymbol{\nu}\cdot\nabla u(t,x,y_{k+\frac{1}{2}})\,\mathrm{d}x\,\mathrm{d}t = \\
&= -\int_{t_n}^{t_{n+1}} \int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \partial_x u(t,x_{j-\frac{1}{2}},y)\,\mathrm{d}y\,\mathrm{d}t + \int_{t_n}^{t_{n+1}} \int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \partial_x u(t,x_{j+\frac{1}{2}},y)\,\mathrm{d}y\,\mathrm{d}t - \\
&- \int_{t_n}^{t_{n+1}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \partial_y u(t,x,y_{k-\frac{1}{2}})\,\mathrm{d}x\,\mathrm{d}t + \int_{t_n}^{t_{n+1}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \partial_y u(t,x,y_{k+\frac{1}{2}})\,\mathrm{d}x\,\mathrm{d}t.
\end{aligned} \tag{13.5}$$

A továbbiakban alkalmazott közelítésekhez megjegyezzük, hogy a kétdimenziós középpont-szabály rendje változónként 2, azaz

$$\begin{aligned}
&\int_{a_1}^{b_1} \int_{a_2}^{b_2} g(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= (b_1-a_1)(b_2-a_2) g\left(\frac{a_1+b_1}{2}, \frac{a_2+b_2}{2}\right) + (b_1-a_1)(b_2-a_2)\cdot\mathcal{O}((b_1-a_1)^2+(b_2-a_2)^2),
\end{aligned} \tag{13.6}$$

és ugyanilyen rendű a trapéz-szabály is. Ennek megfelelően (13.5) bal oldalán két dimenzióban a középpont-szabályt alkalmazva:

$$\int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t_{n+1}, x, y)\,\mathrm{d}x\,\mathrm{d}y = h_x h_y (u(t_{n+1}, x_j, y_k) + \mathcal{O}(h_x^2 + h_y^2)), \tag{13.7}$$

és hasonló eredményt kapunk a másik tagra is.
Teljesül továbbá, hogy

$$\begin{aligned} &\frac{1}{h_x}(\partial_x u(t, x_{j+\frac{1}{2}}, y) - \partial_x u(t, x_{j-\frac{1}{2}}, y)) = \partial_{xx} u(t, x_j, y) + \mathcal{O}(h_x^2) = \\ =&\frac{u(t, x_{j-1}, y) - 2u(t, x_j, y) + u(t, x_{j+1}, y)}{h_x^2} + \mathcal{O}(h_x^2) = D_{0,x}^2 u(t, x_j, y) + \mathcal{O}(h_x^2). \end{aligned} \tag{13.8}$$

A (13.8) formula alapján (13.5) utolsó előtti sorának tagjait a következőképpen adhatjuk meg:

$$\begin{aligned} &-\int_{t_n}^{t_{n+1}} \int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \partial_x u(t, x_{j-\frac{1}{2}}, y)\,\mathrm{d}y\,\mathrm{d}t + \int_{t_n}^{t_{n+1}} \int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \partial_x u(t, x_{j+\frac{1}{2}}, y)\,\mathrm{d}y\,\mathrm{d}t = \\ &= \int_{t_n}^{t_{n+1}} \int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} h_x (D_{0,x}^2 u(t, x_j, y) + \mathcal{O}(h_x^2))\,\mathrm{d}y\,\mathrm{d}t = \\ &= \int_{t_n}^{t_{n+1}} h_x h_y (D_{0,x}^2 u(t, x_j, y_k) + \mathcal{O}(h_x^2) + \mathcal{O}(h_y^2))\,\mathrm{d}t, \end{aligned}$$

amit a trapéz-szabállyal közelítve nyerjük a

$$\begin{aligned} &-\int_{t_n}^{t_{n+1}} \int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \partial_x u(t, x_{j-\frac{1}{2}}, y)\,\mathrm{d}y\,\mathrm{d}t + \int_{t_n}^{t_{n+1}} \int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \partial_x u(t, x_{j+\frac{1}{2}}, y)\,\mathrm{d}y\,\mathrm{d}t = \\ &= h_x h_y \frac{\delta}{2} (D_{0,x}^2 u(t_{n+1}, x_j, y_k) + D_{0,x}^2 u(t_n, x_j, y_k) + \mathcal{O}(h_x^2) + \mathcal{O}(h_y^2) + \mathcal{O}(\delta^2)) \end{aligned} \tag{13.9}$$

közelítést.
Hasonlóan kapjuk az

$$\begin{aligned} &-\int_{t_n}^{t_{n+1}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \partial_y u(t, x, y_{k-\frac{1}{2}})\,\mathrm{d}x\,\mathrm{d}t + \int_{t_n}^{t_{n+1}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \partial_y u(t, x, y_{k+\frac{1}{2}})\,\mathrm{d}x\,\mathrm{d}t \\ &= h_x h_y \frac{\delta}{2} (D_{0,y}^2 u(t_{n+1}, x_j, y_k) + D_{0,y}^2 u(t_n, x_j, y_k) + \mathcal{O}(h_x^2) + \mathcal{O}(h_y^2) + \mathcal{O}(\delta^2)) \end{aligned} \tag{13.10}$$

egyenlőséget is. Ezután (13.5) bal oldalára a (13.7), jobb oldalára pedig a (13.9) és a (13.10) közelítéseket alkalmazva kapjuk, hogy

$$\begin{aligned} &h_x h_y(u(t_{n+1},x_j,y_k)+\mathcal{O}(h_x^2+h_y^2))-h_x h_y(u(t_n,x_j,y_k)+\mathcal{O}(h_x^2+h_y^2))= \\ &=h_x h_y\frac{\delta}{2}(D_{0,x}^2u(t_{n+1},x_j,y_k)+D_{0,x}^2u(t_n,x_j,y_k)+\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)+\mathcal{O}(\delta^2))+ \\ &+h_x h_y\frac{\delta}{2}(D_{0,y}^2u(t_{n+1},x_j,y_k)+D_{0,y}^2u(t_n,x_j,y_k)+\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)+\mathcal{O}(\delta^2)), \end{aligned} \tag{13.11}$$

vagyis mindkét oldalt h_xh_y-nal osztva, valamint a hibatagokat elhagyva kapjuk az

$$\mathbf{u}^{n+1}-\left(\frac{r_x}{2}D_{0,x}^2\mathbf{u}^{n+1}+\frac{r_y}{2}D_{0,y}^2\mathbf{u}^{n+1}\right)=\mathbf{u}^n+\left(\frac{r_x}{2}D_{0,x}^2\mathbf{u}^n+\frac{r_y}{2}D_{0,y}^2\mathbf{u}^n\right) \tag{13.12}$$

Crank–Nicolson-típusú sémát.

13.3. Lemma. *A (13.2) egyenletre vonatkozó (13.12) séma minden változó szerint másodrendben konzisztens, valamint feltétel nélkül stabil.*

Bizonyítás. A konzisztenciára vonatkozó állítás, sajnos, nem látszik a (13.11) alakból, mert annak első sorában (azaz a formula bal oldalán) levő hibatag nem tartalmazza a δ paramétert. Ezért közvetlen bizonyítást adunk erre; csak a pontonkénti konzisztenciát igazoljuk.
Ehhez a sémát az

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\delta}=\frac{1}{2}\left(\frac{1}{h_x^2}D_{0,x}^2(\mathbf{u}^n+\mathbf{u}^{n+1})+\frac{1}{h_y^2}D_{0,y}^2(\mathbf{u}^n+\mathbf{u}^{n+1})\right) \tag{13.13}$$

alakba írjuk. Tudjuk, hogy a bal oldalba a pontos megoldást helyettesítve

$$\frac{u((n+1)\delta,x_j,y_k)-u(n\delta,x_j,y_k)}{\delta}=\partial_t u((n+\frac{1}{2})\delta,x_j,y_k)+\mathcal{O}(\delta^2),$$

továbbá (13.13) jobb oldalán a $\frac{1}{h_x^2}D_{0,x}^2(\mathbf{u}^n+\mathbf{u}^{n+1})$ tagba helyettesítve az a következő alakba írható:

$$\partial_{xx}u((n+1)\delta,x_j,y_k)+\partial_{xx}u((n+1)\delta,x_j,y_k)+\mathcal{O}(\delta^2).$$

Felhasználva a

$$\partial_{xx}u((n+\frac{1}{2})\delta,x_j,y_k)=\frac{1}{2}(\partial_{xx}u((n+1)\delta,x_j,y_k)+\partial_{xx}u(n\delta,x_j,y_k))+\mathcal{O}(\delta^2)$$

formulát, majd ugyanezt az y változó szerint is alkalmazva összességében a (13.13) sémába a pontos megoldást helyettesítve nyerjük, hogy

$$\begin{aligned} &\partial_t u((n+\frac{1}{2})\delta,x_j,y_k)+\mathcal{O}(\delta^2)=\partial_{xx}u((n+\frac{1}{2})\delta,x_j,y_k)+\mathcal{O}(h_x^2)+\mathcal{O}(\delta^2)+ \\ &+\partial_{yy}u((n+\frac{1}{2})\delta,x_j,y_k)+\mathcal{O}(h_y^2)+\mathcal{O}(\delta^2) \end{aligned} \tag{13.14}$$

Mivel u a pontos megoldás, így a (13.13) séma hibatagja

$$\mathcal{O}(h_x^2) + \mathcal{O}(h_y^2) + \mathcal{O}(\delta^2),$$

amelyet δ-val kell szorozni, hogy az eredeti (13.12) sémát kapjuk, így az pontonként valóban minden változó szerint másodrendben konzisztens. A stabilitás igazolásához a (13.12) formulában szereplő egyenlőség mindkét oldalára Fourier-transzformációt alkalmazva kapjuk, hogy

$$\begin{aligned}&\mathcal{F}\left(\mathbf{u}^{n+1} - \left(\frac{r_x}{2} D_{0,x}^2 \mathbf{u}^{n+1} + \frac{r_y}{2} D_{0,y}^2 \mathbf{u}^{n+1}\right)\right)(\alpha_1, \alpha_2) = \\ &= \mathcal{F}(\mathbf{u}^{n+1})(\alpha_1, \alpha_2)\left(1 + \frac{r_x}{2} \cdot 4\sin^2\frac{\alpha_1}{2} + \frac{r_y}{2} \cdot 4\sin^2\frac{\alpha_2}{2}\right),\end{aligned}$$

valamint

$$\begin{aligned}&\mathcal{F}\left(\mathbf{u}^{n} - \left(\frac{r_x}{2} D_{0,x}^2 \mathbf{u}^{n+1} + \frac{r_y}{2} D_{0,y}^2 \mathbf{u}^{n+1}\right)\right)(\alpha_1, \alpha_2) = \\ &= \mathcal{F}(\mathbf{u}^{n})(\alpha_1, \alpha_2)\left(1 - \frac{r_x}{2} \cdot 4\sin^2\frac{\alpha_1}{2} - \frac{r_y}{2} \cdot 4\sin^2\frac{\alpha_2}{2}\right),\end{aligned}$$

azaz a szorzófaktor a következő alakú lesz:

$$\rho(\alpha_1, \alpha_2) = \frac{1 - \frac{r_x}{2} \cdot 4\sin^2\frac{\alpha_1}{2} - \frac{r_y}{2} \cdot 4\sin^2\frac{\alpha_2}{2}}{1 + \frac{r_x}{2} \cdot 4\sin^2\frac{\alpha_1}{2} + \frac{r_y}{2} \cdot 4\sin^2\frac{\alpha_2}{2}}.$$

Itt az $a_x = \frac{r_x}{2} \cdot 4\sin^2\frac{\alpha_1}{2}$ és $a_y = \frac{r_y}{2} \cdot 4\sin^2\frac{\alpha_2}{2}$. jelölések bevezetésével, valamint az $a_x, a_y \geq 0$ egyenlőtlenség felhasználásával nyerjük, hogy

$$-(1 + a_x + a_y) \leq 1 - a_x - a_y \leq 1 + a_x + a_y,$$

azaz

$$|\rho(\alpha_1, \alpha_2)| = \left|\frac{1 - a_x - a_y}{1 + a_x + a_y}\right| \leq 1,$$

amiből (12.11) következik a (13.12) séma stabilitása. □

A (13.12) sémával kapcsolatban lásd a 20.2.10. animációt!

Figyeljük meg, hogy a sémában szereplő időlépés végrehajtásához egy olyan mátrixot kell invertálnunk, amelyikben egy nemnulla főátlón kívül még 4 nemnulla elem van, és ezek nincsenek mind közel a főátlóhoz. Ezért azzal próbálkozunk, hogy olyan sémát találjunk, amely ugyanígy feltétel nélkül stabil, ugyanakkor az abban szereplő egyenletrendszer megoldása jóval kevesebb műveletet igényel. Felmerül természetesen a kérdés, hogy szükséges-e implicit sémát alkalmazni. A válasz, sajnos, igenlő, amit a numerikus függési tartomány vizsgálatakor adunk meg részletesebben.

13.2.2. Váltakozó irányban implicit (ADI) típusú sémák

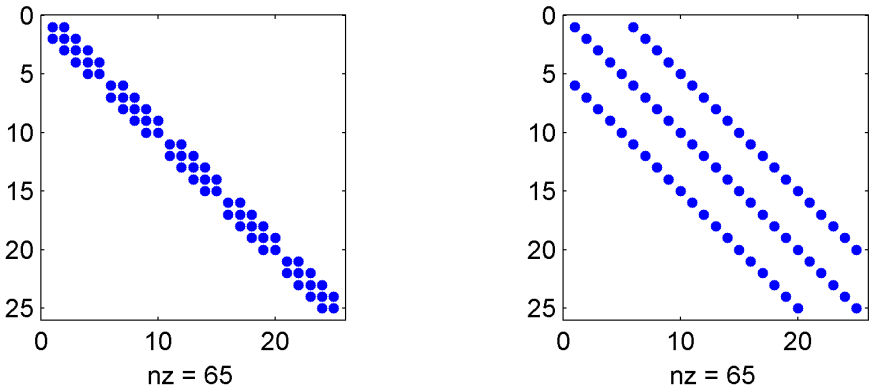
A fenti kívánalomnak megfelelő sémát konstruálhatunk az alábbi elv alapján:

- Először egy fél lépést teszünk úgy, hogy az egyik változó (legyen ez x) szerint explicit a másik szerint pedig implicit sémát írunk fel.
- Az előző fél lépés eredményét használva most az y változó szerint explicit, majd az x szerint implicit sémát alkalmazunk.

A fenti elv szerint kapott sémát az

$$\begin{cases} \mathbf{u}^{n+\frac{1}{2}} - \frac{r_y}{2} D_{0,y}^2 \mathbf{u}^{n+\frac{1}{2}} = \mathbf{u}^n + \frac{r_x}{2} D_{0,x}^2 \mathbf{u}^n \\ \mathbf{u}^{n+1} - \frac{r_x}{2} D_{0,x}^2 \mathbf{u}^{n+1} = \mathbf{u}^{n+\frac{1}{2}} + \frac{r_y}{2} D_{0,y}^2 \mathbf{u}^{n+\frac{1}{2}} \end{cases} \tag{13.15}$$

alakba írhatjuk, és Peaceman–Rachford-sémának nevezzük (a sémában szereplő mátrixok alakjához lásd a 13.3. ábrát).



13.3. ábra. A Peaceman–Rachford-séma bal oldalán szereplő $I - \frac{r_x}{2} D_{0,x}^2$ és $I - \frac{r_y}{2} D_{0,y}^2$ mátrixok nemnulla elemeinek elhelyezkedése.

13.4. Lemma. *A (13.2) egyenletre vonatkozó (13.15) séma feltétel nélkül stabil.*

Bizonyítás. A (13.15) sémában szereplő egyenlőség mindkét oldalára Fourier-transzformációt alkalmazva kapjuk, hogy

$$\mathcal{F}\mathbf{u}^{n+\frac{1}{2}}(\alpha_1, \alpha_2) \left(1 + \frac{r_y}{2} \cdot 4 \sin^2 \frac{\alpha_2}{2}\right) = \mathcal{F}(\mathbf{u}^n)(\alpha_1, \alpha_2) \left(1 - \frac{r_x}{2} \cdot 4 \sin^2 \frac{\alpha_1}{2}\right),$$

valamint

$$\mathcal{F}\mathbf{u}^{n+1}(\alpha_1,\alpha_2)\left(1+\frac{r_x}{2}\cdot 4\sin^2\frac{\alpha_1}{2}\right)=\mathcal{F}\mathbf{u}^{n+\frac{1}{2}}(\alpha_1,\alpha_2)\left(1-\frac{r_y}{2}\cdot 4\sin^2\frac{\alpha_2}{2}\right),$$

vagyis összevetve ezeket adódik, hogy

$$\begin{aligned}\rho(\alpha_1,\alpha_2)&=\frac{\mathcal{F}\mathbf{u}^{n+1}(\alpha_1,\alpha_2)}{\mathcal{F}\mathbf{u}^{n}(\alpha_1,\alpha_2)}=\\&=\frac{(1-2r_y\cdot\sin^2\frac{\alpha_2}{2})(1-2r_x\cdot\sin^2\frac{\alpha_1}{2})}{(1+2r_y\cdot\sin^2\frac{\alpha_2}{2})(1+2r_x\cdot\sin^2\frac{\alpha_1}{2})}.\end{aligned}$$

A 13.3. lemma jelöléseivel tehát

$$\rho(\alpha_1,\alpha_2)=\frac{(1-a_x)(1-a_y)}{(1+a_x)(1+a_y)}$$

és $0\le a_x,a_y$ miatt ismét $|\rho(\alpha_1,\alpha_2)|\le 1$, tehát a (13.15) séma valóban feltétel nélkül stabil. □

A stabilitás bizonyítása egyszerű volt, a módszer pedig szinte ugyanaz, mint amit korábban használtunk. Érdekes, hogy a konzisztencia bizonyítása egyáltalán nem nyilvánvaló. Az alábbi bizonyításban szereplő elemzésnek magasabb dimenzióban is hasznát vehetjük.

13.5. Lemma. *A (13.15) séma $\mathcal{O}(\delta^2)+\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)$ rendben konzisztens a (13.2) egyenlettel.*

Bizonyítás. A bizonyításhoz a sémát először az

$$\begin{cases}\left(I-\frac{r_y}{2}D_{0,y}^2\right)\mathbf{u}^{n+\frac{1}{2}}=\left(I+\frac{r_x}{2}D_{0,x}^2\right)\mathbf{u}^{n}\\ \left(I-\frac{r_x}{2}D_{0,x}^2\right)\mathbf{u}^{n+1}=\left(I+\frac{r_y}{2}D_{0,y}^2\right)\mathbf{u}^{n+\frac{1}{2}}\end{cases}\tag{13.16}$$

alakba írjuk. Az első sort balról $I+\frac{r_y}{2}D_{0,y}^2$-vel szorozva nyerjük, hogy

$$\left(I+\frac{r_y}{2}D_{0,y}^2\right)\left(I-\frac{r_y}{2}D_{0,y}^2\right)\mathbf{u}^{n+\frac{1}{2}}=\left(I+\frac{r_y}{2}D_{0,y}^2\right)\left(I+\frac{r_x}{2}D_{0,x}^2\right)\mathbf{u}^{n},$$

ahol az első két komponens felcserélésével (ez itt megtehető), majd (13.16) második azonosságának alkalmazásával kapjuk, hogy

$$\begin{aligned}\left(I+\frac{r_y}{2}D_{0,y}^2\right)\left(I+\frac{r_x}{2}D_{0,x}^2\right)\mathbf{u}^{n}&=\left(I-\frac{r_y}{2}D_{0,y}^2\right)\left(I+\frac{r_y}{2}D_{0,y}^2\right)\mathbf{u}^{n+\frac{1}{2}}=\\&=\left(I-\frac{r_y}{2}D_{0,y}^2\right)\left(I-\frac{r_x}{2}D_{0,x}^2\right)\mathbf{u}^{n+1},\end{aligned}\tag{13.17}$$

amely tehát a (13.16) séma következménye. Ugyanakkor az

$$\left(I + \frac{r_y}{2} D^2_{0,y}\right)\left(I + \frac{r_x}{2} D^2_{0,x}\right)\mathbf{u}^n = \left(I - \frac{r_y}{2} D^2_{0,y}\right)\left(I - \frac{r_x}{2} D^2_{0,x}\right)\mathbf{u}^{n+1}, \tag{13.18}$$

séma ekvivalens is a (13.16) sémával, mert az

$$\mathbf{u}^{n+\frac{1}{2}} := \left(I - \frac{r_y}{2} D^2_{0,y}\right)^{-1}\left(I + \frac{r_x}{2} D^2_{0,x}\right)\mathbf{u}^n$$

egyenlőséggel definiált $\mathbf{u}^{n+\frac{1}{2}}$ vektorra teljesül (13.16) első sora, emellett ezt az egyenlőséget balról $\left(I + \frac{r_y}{2} D^2_{0,y}\right)$-nal szorozva kapjuk a (13.17) egyenlőségben szereplő első egyenlőséget, valamint (13.18) alapján a benne szereplő második egyenlőség is teljesül, vagyis abból nyerjük a (13.16) séma második sorát is. Kifejtve a kapott egyenlőség jobb és bal oldalát nyerjük azt is, hogy

$$\begin{aligned}&\left(I + \frac{r_y}{2} D^2_{0,y} + \frac{r_x}{2} D^2_{0,x} + \frac{r_x r_y}{4} D^2_{0,y} D^2_{0,x}\right)\mathbf{u}^n = \\ &= \left(I - \frac{r_y}{2} D^2_{0,y} - \frac{r_x}{2} D^2_{0,x} + \frac{r_x r_y}{4} D^2_{0,y} D^2_{0,x}\right)\mathbf{u}^{n+1},\end{aligned}$$

ahol az utolsó tagokat elhagyva pontosan a Crank–Nicolson-sémát kapnánk. Úgy is mondhatjuk, ennyivel perturbáltuk azt, hogy a fenti (13.15) sémát nyerjük. Itt a két oldal utolsó tagjába a pontos megoldást helyettesítve, és azokat egymásból kivonva kapjuk, majd a 13.1. lemma 2. pontjában szereplő eredményt is felhasználva kapjuk, hogy

$$\begin{aligned}&\frac{r_x r_y}{4} D^2_{0,y} D^2_{0,x}[\mathbf{u}((n+1)\delta, \cdot) - \mathbf{u}(n\delta, \cdot)] = \frac{\delta^3}{4}\frac{1}{h_y^2} D^2_{0,y}\frac{1}{h_x^2} D^2_{0,x}\frac{1}{\delta} D_+ \mathbf{u}(n\delta, \cdot) = \\ &= \delta^3(\partial_{xxyy}\partial_t \mathbf{u}(n\delta, \cdot) + \mathcal{O}(h_y^2) + \mathcal{O}(h_x^2))\end{aligned}$$

vagyis a (13.17)-ban, azaz a (13.16)-ben felírt séma is ugyanúgy $\mathcal{O}(\delta^2) + \mathcal{O}(h_x^2) + \mathcal{O}(h_y^2)$ rendben konzisztens a (13.2) egyenlettel. □

A séma (13.17) faktorizált alakjából egy másik, az eredetivel rokon sémát is felírhatunk:
Legyen ezúttal $u^{n+\frac{1}{2}} := \left(I - \frac{r_x}{2} D^2_{0,x}\right)\mathbf{u}^{n+1}$, azaz a (13.17) egyenlőségből

$$\begin{cases}\left(I - \frac{r_y}{2} D^2_{0,y}\right)\mathbf{u}^{n+\frac{1}{2}} = \left(I + \frac{r_y}{2} D^2_{0,y}\right)\left(I + \frac{r_x}{2} D^2_{0,x}\right)\mathbf{u}^n \\ \left(I - \frac{r_x}{2} D^2_{0,x}\right)\mathbf{u}^{n+1} = \mathbf{u}^{n+\frac{1}{2}},\end{cases}$$

amelyet D'Yakonov-sémának is neveznek. Nyilvánvalóan ez ugyanolyan rendben konzisztens, mint a Peaceman–Rachford-séma, és ugyanúgy stabil is, hiszen a hozzá tartozó lépésoperátor ugyanaz, mint a Peaceman–Rachford-sémához tartozó.

13.3. Parabolikus egyenletek 3 dimenzióban

A (13.1) egyenlet háromdimenziós verziójára vonatkozó sémák keresésekor kézenfekvőnek tűnik az olyan ADI típusú séma, amellyel 1/3 lépést teszünk 3-szor, mindig csak egy-egy változó szerint implicit formulát alkalmazva, azaz a

$$\begin{cases} \mathbf{u}^{n+\frac{1}{3}} - \frac{r_z}{3} D^2_{0,h_z} \mathbf{u}^{n+\frac{1}{3}} = \mathbf{u}^n + \frac{r_x}{3} D^2_{0,h_x} \mathbf{u}^n + \frac{r_y}{3} D^2_{0,h_y} \mathbf{u}^n \\ \mathbf{u}^{n+\frac{2}{3}} - \frac{r_y}{3} D^2_{0,h_y} \mathbf{u}^{n+\frac{2}{3}} = \mathbf{u}^{n+\frac{1}{3}} + \frac{r_x}{3} D^2_{0,h_x} \mathbf{u}^{n+\frac{1}{3}} + \frac{r_z}{3} D^2_{0,h_z} \mathbf{u}^{n+\frac{1}{3}} \\ \mathbf{u}^{n+1} - \frac{r_x}{3} D^2_{0,h_x} \mathbf{u}^{n+1} = \mathbf{u}^{n+\frac{2}{3}} + \frac{r_y}{3} D^2_{0,h_y} \mathbf{u}^{n+\frac{2}{3}} + \frac{r_z}{3} D^2_{0,h_z} \mathbf{u}^{n+\frac{2}{3}} \end{cases} \tag{13.19}$$

sémát írhatjuk fel. Kiderül azonban, hogy ez nem rendelkezik a kívánt stabilitási tulajdonsággal, sőt a konzisztenciarendje is alacsonyabb, mint a (13.2) egyenletre vonatkozó (13.15) sémának: idő szerint csak $\mathcal{O}(\delta)$ rendben konzisztens. Csak az előbbi állítást bizonyítjuk:

13.6. Állítás. *A (13.19) séma feltételesen stabil.*

Bizonyítás. Mindhárom egyenletre diszkrét idejű Fourier-transzformációt alkalmazva kapjuk, hogy

$$\begin{cases} \mathcal{F}\mathbf{u}^{n+\frac{1}{3}}(\boldsymbol{\alpha})(1 + \frac{4}{3} r_z \sin^2 \frac{\alpha_3}{2}) = \mathcal{F}\mathbf{u}^n(\boldsymbol{\alpha})(1 - \frac{4}{3} r_x \sin^2 \frac{\alpha_1}{2} - \frac{4}{3} r_y \sin^2 \frac{\alpha_2}{2}) \\ \mathcal{F}\mathbf{u}^{n+\frac{2}{3}}(\boldsymbol{\alpha})(1 + \frac{4}{3} r_y \sin^2 \frac{\alpha_2}{2}) = \mathcal{F}\mathbf{u}^{n+\frac{1}{3}}(\boldsymbol{\alpha})(1 - \frac{4}{3} r_z \sin^2 \frac{\alpha_3}{2} - \frac{4}{3} r_x \sin^2 \frac{\alpha_1}{2}) \\ \mathcal{F}\mathbf{u}^{n+1}(\boldsymbol{\alpha})(1 + \frac{4}{3} r_x \sin^2 \frac{\alpha_1}{2}) = \mathcal{F}\mathbf{u}^{n+\frac{2}{3}}(\boldsymbol{\alpha})(1 - \frac{4}{3} r_y \sin^2 \frac{\alpha_2}{2} - \frac{4}{3} r_z \sin^2 \frac{\alpha_3}{2}), \end{cases}$$

amelyeket összeszorozva, majd átrendezve nyerjük az

$$\frac{\mathcal{F}\mathbf{u}^{n+1}(\boldsymbol{\alpha})}{\mathcal{F}\mathbf{u}^n(\boldsymbol{\alpha})} = \\ = \frac{(1 - \frac{4}{3} r_z \sin^2 \frac{\alpha_3}{2} - \frac{4}{3} r_x \sin^2 \frac{\alpha_1}{2})(1 - \frac{4}{3} r_y \sin^2 \frac{\alpha_2}{2} - \frac{4}{3} r_z \sin^2 \frac{\alpha_3}{2})(1 - \frac{4}{3} r_x \sin^2 \frac{\alpha_1}{2} - \frac{4}{3} r_y \sin^2 \frac{\alpha_2}{2})}{(1 + \frac{4}{3} r_x \sin^2 \frac{\alpha_1}{2})(1 + \frac{4}{3} r_y \sin^2 \frac{\alpha_2}{2})(1 + \frac{4}{3} r_z \sin^2 \frac{\alpha_3}{2})}$$

egyenlőséget. Ha itt például $r_x = r_y = r_z = 3$, akkor

$$\rho(\pi, \pi, \pi) = \frac{\mathcal{F}\mathbf{u}^{n+1}\left(\frac{\pi}{2}\right)}{\mathcal{F}\mathbf{u}^n\left(\frac{\pi}{2}\right)} = -\frac{7^3}{5^3},$$

vagyis a séma ebben az esetben biztosan instabil. □

Tovább fogunk keresni, olyan sémát szeretnénk, amely minden változó szerint másodrendben konzisztens, valamint feltétel nélkül stabil is.

13.3.1. Sémák faktorizációja

Háromdimenziós ADI típusú sémák előállításánál nagyon bonyolult formulákat kapnánk, ha a konzisztenciarend kiszámításához mindig egyszerűen a Taylor-sorfejtéseket alkalmaznánk. Azt sem tudjuk előre, hogy milyen sémával próbálkozzunk, ezért a levezetésekhez egy egységes, kevés számolást igénylő eljárást volna jó használni.

Az alapötlet ugyanaz, mint a kétdimenziós esetben. Egy ismert konzisztens sémából indulunk ki, és megváltoztatjuk úgy, hogy a konzisztenciarend megmaradjon, ugyanakkor a kapott sémában szereplő közelítést egyszerűen és hatékonyan ki tudjuk számolni. Ehhez egyszerűen invertálható operátorok szorzatára bontjuk.

Ennek megfelelően a (13.1) egyenlet háromdimenziós verziójára vonatkozó ismert Crank–Nicolson-sémából indulunk ki (13.4. ábra):

$$\mathbf{u}^{n+1} - (\frac{r_x}{2}D^2_{0,x} + \frac{r_y}{2}D^2_{0,y} + \frac{r_z}{2}D^2_{0,z})\mathbf{u}^{n+1} = \mathbf{u}^n + (\frac{r_x}{2}D^2_{0,x} + \frac{r_y}{2}D^2_{0,y} + \frac{r_z}{2}D^2_{0,z})\mathbf{u}^n. \quad (13.20)$$

Erre vonatkozóan használjuk a következő lemma eredményét.

13.7. Lemma. *A (13.20) séma $\mathcal{O}(\delta^2) + \mathcal{O}(h_x^2) + \mathcal{O}(h_y^2)$ rendben konzisztens a (13.1) egyenlet háromdimenziós verziójával, valamint feltétel nélkül stabil.*

A bizonyítás analóg a kétdimenziós esetre vonatkozó hasonló állítás bizonyításával, a konzisztencia igazolása összetett. Itt nem térünk ki ezekre. Tudjuk, hogy

$$I - (\frac{r_x}{2}D^2_{0,x} + \frac{r_y}{2}D^2_{0,y} + \frac{r_z}{2}D^2_{0,z}) = (I - \frac{r_x}{2}D^2_{0,x})(I - \frac{r_y}{2}D^2_{0,y})(I - \frac{r_z}{2}D^2_{0,z}) -$$
$$- \frac{r_x}{2}\frac{r_y}{2}D^2_{0,x}D^2_{0,y} - \frac{r_y}{2}\frac{r_z}{2}D^2_{0,y}D^2_{0,z} - \frac{r_z}{2}\frac{r_x}{2}D^2_{0,z}D^2_{0,x} + \frac{r_x}{2}\frac{r_y}{2}\frac{r_z}{2}D^2_{0,x}D^2_{0,y}D^2_{0,z},$$

és hasonlóan

$$I + (\frac{r_x}{2}D^2_{0,x} + \frac{r_y}{2}D^2_{0,y} + \frac{r_z}{2}D^2_{0,z}) = (I + \frac{r_x}{2}D^2_{0,x})(I + \frac{r_y}{2}D^2_{0,y})(I + \frac{r_z}{2}D^2_{0,z}) -$$
$$- \frac{r_x}{2}\frac{r_y}{2}D^2_{0,x}D^2_{0,y} - \frac{r_y}{2}\frac{r_z}{2}D^2_{0,y}D^2_{0,z} - \frac{r_z}{2}\frac{r_x}{2}D^2_{0,z}D^2_{0,x} - \frac{r_x}{2}\frac{r_y}{2}\frac{r_z}{2}D^2_{0,x}D^2_{0,y}D^2_{0,z}.$$

Ezeket a (13.20) formula bal és jobb oldalába helyettesítve, majd az egészet rendezve, és a 13.7. lemmában szereplő hibatagokat beírva nyerjük, hogy

$$\begin{aligned}
&(I - \frac{r_x}{2}D^2_{0,x})(I - \frac{r_y}{2}D^2_{0,y})(I - \frac{r_z}{2}D^2_{0,z})\mathbf{u}^{n+1} = \\
&= (I + \frac{r_x}{2}D^2_{0,x})(I + \frac{r_y}{2}D^2_{0,y})(I + \frac{r_z}{2}D^2_{0,z})\mathbf{u}^n + \\
&+ \left(\frac{r_x}{2}\frac{r_y}{2}D^2_{0,x}D^2_{0,y} + \frac{r_y}{2}\frac{r_z}{2}D^2_{0,y}D^2_{0,z} + \frac{r_z}{2}\frac{r_x}{2}D^2_{0,z}D^2_{0,x}\right)(\mathbf{u}^{n+1} - \mathbf{u}^n) - \\
&- \frac{r_x}{2}\frac{r_y}{2}\frac{r_z}{2}D^2_{0,x}D^2_{0,y}D^2_{0,z}(\mathbf{u}^{n+1} + \mathbf{u}^n) + \delta(\mathcal{O}(\delta^2) + \mathcal{O}(h_x^2) + \mathcal{O}(h_y^2)).
\end{aligned} \quad (13.21)$$

A jobb oldal hibatag előtti utolsó tagja

$$\frac{1}{8}\delta^3\frac{1}{h_x^2}D_{0,x}^2\frac{1}{h_y^2}D_{0,y}^2\frac{1}{h_z^2}D_{0,z}^2(\mathbf{u}^{n+1}+\mathbf{u}^n)$$

alakba írható. A 13.1. lemma 3. pontja szerint ez pontonként $\mathcal{O}(\delta^4)$ nagyságrendű. Hasonlóan, a jobb oldalon az előtte levő sor első tagja

$$\frac{r_x}{2}\frac{r_y}{2}D_{0,x}^2D_{0,y}^2(\mathbf{u}^{n+1}-\mathbf{u}^n)=\frac{1}{4}\delta^3\frac{1}{h_y^2}D_{0,y}^2\frac{1}{h_x^2}D_{0,x}^2\frac{1}{\delta}D_{+,\delta}\mathbf{u}^n$$

alakba írható. A 13.1. lemma 2. pontja szerint ez pontonként $\mathcal{O}(\delta^3)$ nagyságrendű. Ugyanilyen nagyságrendi becslés teljesül (13.21) többi tagjára is.

Innen kapjuk a következő lemma eredményét:

13.8. Lemma. *Az alábbi*

$$(I-\frac{r_x}{2}D_{0,x}^2)(I-\frac{r_y}{2}D_{0,y}^2)(I-\frac{r_z}{2}D_{0,z}^2)\mathbf{u}^{n+1}=(I+\frac{r_x}{2}D_{0,x}^2)(I+\frac{r_y}{2}D_{0,y}^2)(I+\frac{r_z}{2}D_{0,z}^2)\mathbf{u}^n, \tag{13.22}$$

faktorizált alakú séma $\mathcal{O}(\delta^2)+\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)+\mathcal{O}(h_z^2)$ rendben konzisztens a (13.1) egyenlet háromdimenziós verziójával, valamint a belőle kapott lépésoperátor feltétel nélkül stabil.

Bizonyítás. A (13.21) formula, valamint a jobb oldalon a hibatag előtti négy tagra vonatkozó nagyságrendi becslés alapján kapjuk, hogy

$$(I-\frac{r_x}{2}D_{0,x}^2)(I-\frac{r_y}{2}D_{0,y}^2)(I-\frac{r_z}{2}D_{0,z}^2)\mathbf{u}^{n+1}=(I+\frac{r_x}{2}D_{0,x}^2)(I+\frac{r_y}{2}D_{0,y}^2)(I+\frac{r_z}{2}D_{0,z}^2)\mathbf{u}^n+$$
$$+\delta\mathcal{O}(\delta^2)+\delta(\mathcal{O}(\delta^2)+\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)++\mathcal{O}(h_z^2)),$$

ami éppen a lemmában szereplő nagyságrend. □

Két konstrukció a faktorizált séma alapján

A faktorizált séma segítségével különböző eljárásokat adhatunk u^{n+1} kiszámítására. A formulák egyszerűsítése érdekében a véges differenciákra bevezetjük az

$$A_x:=\frac{r_x}{2}D_{0,x}^2,\ A_y:=\frac{r_y}{2}D_{0,y}^2,\ A_z:=\frac{r_z}{2}D_{0,z}^2$$

jelöléseket.

Az első konstrukcióhoz először (13.22) mindkét oldalából az

$$(I+A_xA_y+A_yA_z+A_zA_x)\,\mathbf{u}^n$$

mennyiséget kivonva, majd az

$$(A_x + A_y + A_z + A_x A_y A_z)\,\mathbf{u}^n$$

mennyiséget hozzáadva kapjuk, hogy

$$(I - A_x)(I - A_y)(I - A_z)(\mathbf{u}^{n+1} - \mathbf{u}^n) = 2(A_x + A_y + A_z)\mathbf{u}^n + 2A_x A_y A_z \mathbf{u}^n.$$

Itt az $A_x A_y A_z \mathbf{u}^n$ tag nagyságrendje $\mathcal{O}(\delta^3)$, tehát a séma konzisztenciarendje változatlan marad, ha ezt elhagyjuk. Tehát a továbbiakban a

$$(I - A_x)(I - A_y)(I - A_z)(\mathbf{u}^{n+1} - \mathbf{u}^n) = 2(A_x + A_y + A_z)\mathbf{u}^n.$$

sémával foglalkozunk. Az

$$\mathbf{u}^{n+\frac{2}{3}} := (I - A_z)(\mathbf{u}^{n+1} - \mathbf{u}^n) \quad \text{és} \quad \mathbf{u}^{n+\frac{1}{3}} := (I - A_y)\mathbf{u}^{n+\frac{2}{3}} = (I - A_y)(I - A_z)(\mathbf{u}^{n+1} - \mathbf{u}^n)$$

jelölések bevezetésével a következő sémát kapjuk:

$$\begin{cases} (I - A_x)\mathbf{u}^{n+\frac{1}{3}} = 2(A_x + A_y + A_z)\mathbf{u}^n \\ (I - A_y)\mathbf{u}^{n+\frac{2}{3}} = \mathbf{u}^{n+\frac{1}{3}} \\ (I - A_z)(\mathbf{u}^{n+1} - \mathbf{u}^n) = \mathbf{u}^{n+\frac{2}{3}} \\ \mathbf{u}^{n+1} = \mathbf{u}^n + \mathbf{u}^{n+1} - \mathbf{u}^n, \end{cases} \tag{13.23}$$

ahol az utolsó két lépést akár össze is vonhatjuk.
A (13.23) sémát Douglas–Gunn-sémának nevezzük (13.5., 13.6., 13.7. ábrák). A másik eljárás ahhoz hasonlít, ahogy a D'Yakonov-sémát konstruáltuk. Ekkor a (13.22) faktorizált sémában az

$$\mathbf{u}^{n+\frac{2}{3}} := (I - A_z)\mathbf{u}^{n+1} \quad \text{és} \quad \mathbf{u}^{n+\frac{1}{3}} := (I - A_y)\mathbf{u}^{n+\frac{2}{3}} = (I - A_y)(I - A_z)\mathbf{u}^{n+1}$$

jelölések bevezetésével a

$$\begin{cases} (I - A_x)\mathbf{u}^{n+\frac{1}{3}} = (I + A_x)(I + A_y)(I + A_z)\mathbf{u}^n \\ (I - A_y)\mathbf{u}^{n+\frac{2}{3}} = \mathbf{u}^{n+\frac{1}{3}} \\ (I - A_z)\mathbf{u}^{n+1} = \mathbf{u}^{n+\frac{2}{3}} \end{cases} \tag{13.24}$$

sémát nyerjük, amelyet háromdimenziós D'Yakonov-sémának nevezhetünk. Mindkét fenti séma alkalmazásakor csakis olyan mátrix invertálására van szükség, ahol 3 nemnulla átló van, azaz minden sor legfeljebb 3 nemnulla elemet tartalmaz. Emiatt, habár az eljárás összetettebbnek tűnik, mint a Crank–Nicolson-séma, mégis annál hatékonyabb módszert kapunk.
A következőkben igazoljuk, hogy a fenti módszerek stabilitási tulajdonsága is megfelelő.

13.9. Lemma. *Mind a (13.23), mind a (13.24) képletben megadott séma feltétel nélkül stabil.*

Bizonyítás. A (13.23) séma egyes lépéseiben mindkét oldal Fourier-transzformáltját véve (az utolsó egyenlőséget közben átalakítva) kapjuk, hogy

$$\begin{cases} \left(1+2r_x\sin^2\frac{\alpha_1}{2}\right)\mathcal{F}\mathbf{u}^{n+\frac{1}{3}}(\boldsymbol{\alpha}) = -4\left(r_x\sin^2\frac{\alpha_1}{2}+r_y\sin^2\frac{\alpha_2}{2}+r_z\sin^2\frac{\alpha_3}{2}\right)\mathcal{F}\mathbf{u}^n(\boldsymbol{\alpha}) \\ \left(1+2r_y\sin^2\frac{\alpha_2}{2}\right)\mathcal{F}\mathbf{u}^{n+\frac{2}{3}}(\boldsymbol{\alpha}) = \mathcal{F}\mathbf{u}^{n+\frac{1}{3}}(\boldsymbol{\alpha}) \\ \left(1+2r_z\sin^2\frac{\alpha_3}{2}\right)\mathcal{F}(\mathbf{u}^{n+1}-\mathbf{u}^n)(\boldsymbol{\alpha}) = \mathcal{F}\mathbf{u}^{n+\frac{2}{3}}(\boldsymbol{\alpha}) \\ \mathcal{F}\mathbf{u}^{n+1}(\boldsymbol{\alpha})-\mathcal{F}\mathbf{u}^n(\boldsymbol{\alpha}) = \mathcal{F}(\mathbf{u}^{n+1}-\mathbf{u}^n)(\boldsymbol{\alpha}). \end{cases} \tag{13.25}$$

Ezeket összeszorozva, egyszerűsítve, majd rendezve nyerjük, hogy

$$\frac{\mathcal{F}\mathbf{u}^{n+1}(\boldsymbol{\alpha})-\mathcal{F}\mathbf{u}^n(\boldsymbol{\alpha})}{\mathcal{F}\mathbf{u}^n(\boldsymbol{\alpha})} = -4\frac{\left(r_x\sin^2\frac{\alpha_1}{2}+r_y\sin^2\frac{\alpha_2}{2}+r_z\sin^2\frac{\alpha_3}{2}\right)}{\left(1+2r_x\sin^2\frac{\alpha_1}{2}\right)\left(1+2r_y\sin^2\frac{\alpha_2}{2}\right)\left(1+2r_z\sin^2\frac{\alpha_3}{2}\right)},$$

azaz bevezetve az

$$A = 2r_x\sin^2\frac{\alpha_1}{2},\ B = 2r_y\sin^2\frac{\alpha_2}{2},\ C = 2r_z\sin^2\frac{\alpha_3}{2}$$

jelöléseket kapjuk, hogy

$$\rho(\boldsymbol{\alpha}) = \frac{\mathcal{F}\mathbf{u}^{n+1}(\boldsymbol{\alpha})}{\mathcal{F}\mathbf{u}^n(\boldsymbol{\alpha})} = 1-2\frac{A+B+C}{(1+A)(1+B)(1+C)}.$$

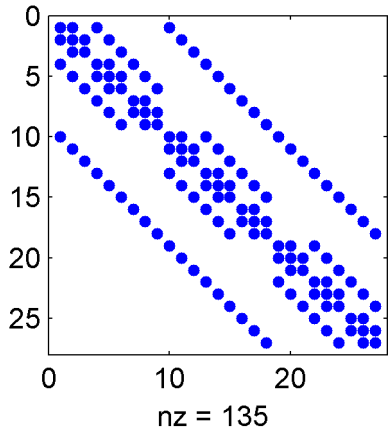
A feltétel nélküli stabilitáshoz azt kell tehát igazolnunk, hogy tetszőleges pozitív A, B és C esetén

$$0 \le \frac{A+B+C}{(1+A)(1+B)(1+C)} \le 1.$$

Ez $A \le 1+A$, $B \le 1+B$ és $C \le 1+C$ miatt nyilvánvaló, így igazoltuk a (13.23) séma feltétel nélküli stabilitását. A második állítást nem bizonyítjuk, ugyanis az a 13.4. lemma bizonyításának nyilvánvaló módosításával kapható. □

13.4. Forrástagok beépítése a faktorizált sémákba

Ebben a szakaszban olyan parabolikus feladatokat vizsgálunk, ahol a jobb oldalon egy f forrástag is szerepel. Akárhogy is építjük be ezeket a faktorizált sémákba, azok stabilitása nyilván nem változik. Amit a következő sémák konstrukciója során ellenőrizni kell, az a megfelelő eljárások konzisztenciája.



13.4. ábra. A háromdimenziós feladat Crank–Nicolson-sémájának bal oldalán álló $I - \frac{r_x}{2}D^2_{0,x} - \frac{r_y}{2}D^2_{0,y} - \frac{r_z}{2}D^2_{0,z}$ mátrix nemnulla elemek elhelyezkedése ($n_x = n_y = n_z = 3$).

13.4.1. A kétdimenziós eset

A Crank–Nicolson-séma elemzésekor kapott (13.14) formulában a hibatagok másodrendűek maradnak, ha a forrástagot tartalmazó esetben a jobb oldalra

$$f((n+\frac{1}{2})\delta, x_j, y_k) \quad \text{vagy} \quad \frac{1}{2}(f(n\delta, x_j, y_k) + f((n+1)\delta, x_j, y_k))$$

kerül. Ennek megfelelően teljesül a következő lemma:

13.10. Lemma. *A kétdimenziós f forrástagot is tartalmazó (13.2) feladattal minden változó szerint másodrendben konzisztensek a következő sémák:*

$$\mathbf{u}^{n+1} - (\frac{r_x}{2}D^2_{0,x} + \frac{r_y}{2}D^2_{0,y})\mathbf{u}^{n+1} = \mathbf{u}^n + (\frac{r_x}{2}D^2_{0,x} + \frac{r_y}{2}D^2_{0,y})\mathbf{u}^n + \delta\mathbf{f}^{n+\frac{1}{2}},$$

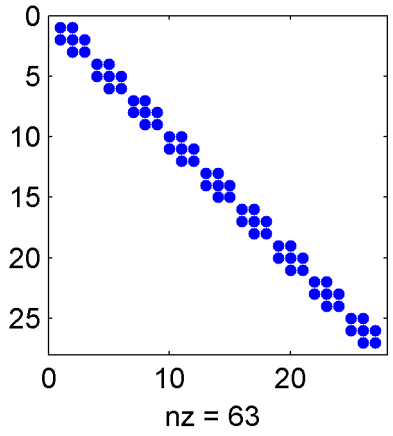
valamint

$$\mathbf{u}^{n+1} - (\frac{r_x}{2}D^2_{0,x} + \frac{r_y}{2}D^2_{0,y})\mathbf{u}^{n+1} = \mathbf{u}^n + (\frac{r_x}{2}D^2_{0,x} + \frac{r_y}{2}D^2_{0,y})\mathbf{u}^n + \delta\frac{\mathbf{f}^n + \mathbf{f}^{n+1}}{2}. \qquad (13.26)$$

Ennek megfelelően a következőket állíthatjuk.

13.11. Állítás. *A (13.15)-beli Peaceman–Rachford-séma módosításával kapott*

$$\begin{cases} \mathbf{u}^{n+\frac{1}{2}} - \frac{r_y}{2}D^2_{0,y}\mathbf{u}^{n+\frac{1}{2}} = \mathbf{u}^n + \frac{r_x}{2}D^2_{0,x}\mathbf{u}^n + \frac{\delta}{2}\mathbf{f}^{n+\frac{1}{2}} \\ \mathbf{u}^{n+1} - \frac{r_x}{2}D^2_{0,x}\mathbf{u}^{n+1} = \mathbf{u}^{n+\frac{1}{2}} + \frac{r_y}{2}D^2_{0,y}\mathbf{u}^{n+\frac{1}{2}} + \frac{\delta}{2}\mathbf{f}^{n+\frac{1}{2}}, \end{cases}$$



13.5. ábra. A háromdimenziós feladat Douglas–Gunn-sémájának bal oldalán álló $I - \frac{r_x}{2} D^2_{0,x}$ mátrix nemnulla elemeinek elhelyezkedése ($n_x = n_y = n_z = 3$).

valamint a

$$\begin{cases} \mathbf{u}^{n+\frac{1}{2}} - \frac{r_y}{2} D^2_{0,y} \mathbf{u}^{n+\frac{1}{2}} = \mathbf{u}^n + \frac{r_x}{2} D^2_{0,x} \mathbf{u}^n + \frac{\delta}{2} \mathbf{f}^n \\ \mathbf{u}^{n+1} - \frac{r_x}{2} D^2_{0,x} \mathbf{u}^{n+1} = \mathbf{u}^{n+\frac{1}{2}} + \frac{r_y}{2} D^2_{0,y} \mathbf{u}^{n+\frac{1}{2}} + \frac{\delta}{2} \mathbf{f}^{n+1} \end{cases} \tag{13.27}$$

faktorizált sémák másodrendben konzisztensek.

Bizonyítás. A második sémára vonatkozó állítást igazoljuk csak, az első hasonlóan történik, sőt egyszerűbb is.
A bizonyításra térve szorozzuk meg (13.27) első sorát $I + \frac{r_y}{2} D^2_{0,y}$-tel, majd cseréljük fel a bal oldalon a komponenseket! Ekkor

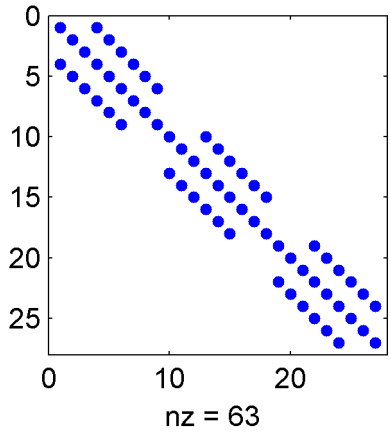
$$\left(I - \frac{r_y}{2} D^2_{0,y}\right) \left(I + \frac{r_y}{2} D^2_{0,y}\right) \mathbf{u}^{n+\frac{1}{2}} = \left(I + \frac{r_y}{2} D^2_{0,y}\right) \left(I + \frac{r_x}{2} D^2_{0,x}\right) \mathbf{u}^n + \left(I + \frac{r_y}{2} D^2_{0,y}\right) \frac{\delta}{2} \mathbf{f}^n,$$

azaz (13.27) második sora miatt

$$\left(I + \frac{r_y}{2} D^2_{0,y}\right) \left(I + \frac{r_x}{2} D^2_{0,x}\right) \mathbf{u}^n + \left(I + \frac{r_y}{2} D^2_{0,y}\right) \frac{\delta}{2} \mathbf{f}^n = \left(I - \frac{r_y}{2} D^2_{0,y}\right) \left(I - \frac{r_x}{2} D^2_{0,y}\right) \mathbf{u}^{n+1} - \\ - \frac{\delta}{2} \left(I - \frac{r_y}{2} D^2_{0,y}\right) \mathbf{f}^{n+1},$$

tehát

$$\left(I + \frac{r_y}{2} D^2_{0,y}\right) \left(I + \frac{r_x}{2} D^2_{0,x}\right) \mathbf{u}^n + \delta \frac{\mathbf{f}^n + \mathbf{f}^{n+1}}{2} = \left(I - \frac{r_y}{2} D^2_{0,y}\right) \left(I - \frac{r_x}{2} D^2_{0,y}\right) \mathbf{u}^{n+1} - \\ + \frac{\delta}{2} \frac{r_y}{2} D^2_{0,y} (\mathbf{f}^{n+1} - \mathbf{f}^n).$$



13.6. ábra. A háromdimenziós feladat Douglas–Gunn-sémájának bal oldalán álló $I - \frac{r_y}{2} D^2_{0,y}$ mátrix nemnulla elemeinek elhelyezkedése ($n_x = n_y = n_z = 3$).

Ez a másodrendben konzisztens (13.26) sémától a $\frac{\delta}{2}\frac{r_y}{2} D^2_{0,y}(\mathbf{f}^{n+1} - \mathbf{f}^n)$ tagban különbözik. Azonban ez $\frac{\delta^3}{4}\frac{1}{h_y^2} D^2_{0,y} \frac{1}{\delta} D_0 \mathbf{f}^{n+\frac{1}{2}}$ alakra írható át, amely δ^3 nagyságrendű, vagyis a fenti konzisztenciarenden nem változtat. □

A forrástag egy másik lehetséges beépítését tárgyalja a 21.13. feladat.

A fenti gondolatmenetnél egyszerűbben, közvetlenül kapjuk a D'Yakonov-séma módosítására vonatkozó alábbi eredményt.

13.12. Állítás. *Mind az*

$$\begin{cases} \left(I - \frac{r_y}{2} D^2_{0,y}\right) \mathbf{u}^{n+\frac{1}{2}} = \left(I + \frac{r_y}{2} D^2_{0,y}\right) \left(I + \frac{r_x}{2} D^2_{0,x}\right) \mathbf{u}^n + \delta \mathbf{f}^{n+\frac{1}{2}} \\ \left(I - \frac{r_x}{2} D^2_{0,x}\right) \mathbf{u}^{n+1} = \mathbf{u}^{n+\frac{1}{2}}, \end{cases}$$

mind az

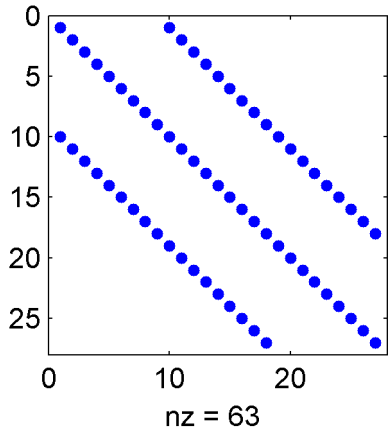
$$\begin{cases} \left(I - \frac{r_y}{2} D^2_{0,y}\right) \mathbf{u}^{n+\frac{1}{2}} = \left(I + \frac{r_y}{2} D^2_{0,y}\right) \left(I + \frac{r_x}{2} D^2_{0,x}\right) \mathbf{u}^n + \frac{\delta}{2}(\mathbf{f}^n + \mathbf{f}^{n+1}) \\ \left(I - \frac{r_x}{2} D^2_{0,x}\right) \mathbf{u}^{n+1} = \mathbf{u}^{n+\frac{1}{2}} \end{cases}$$

séma másodrendben konzisztens a (13.2) egyenlet f forrástaggal ellátott verziójával.

13.4.2. A háromdimenziós eset

A három dimenzióban felírt és f forrástagot is tartalmazó (13.1) feladattal minden változó szerint másodrendben konzisztensek a következő sémák:

$$\mathbf{u}^{n+1} - (\frac{r_x}{2} D^2_{0,x} + \frac{r_y}{2} D^2_{0,y} + \frac{r_z}{2} D^2_{0,z})\mathbf{u}^{n+1} = \mathbf{u}^n + (\frac{r_x}{2} D^2_{0,x} + \frac{r_y}{2} D^2_{0,y} + \frac{r_z}{2} D^2_{0,z})\mathbf{u}^n + \delta \mathbf{f}^{n+\frac{1}{2}},$$



13.7. ábra. A háromdimenziós feladat Douglas–Gunn-sémájának bal oldalán álló $I - \frac{r_z}{2} D^2_{0,z}$ mátrix nemnulla elemeinek elhelyezkedése ($n_x = n_y = n_z = 3$).

valamint

$$\mathbf{u}^{n+1} - (\frac{r_x}{2} D^2_{0,x} + \frac{r_y}{2} D^2_{0,y} + \frac{r_z}{2} D^2_{0,z}) \mathbf{u}^{n+1} = \mathbf{u}^n + (\frac{r_x}{2} D^2_{0,x} + \frac{r_y}{2} D^2_{0,y} + \frac{r_z}{2} D^2_{0,z}) \mathbf{u}^n + \delta \frac{\mathbf{f}^n + \mathbf{f}^{n+1}}{2}.$$

Innen az előző szakaszban ismertetett módszerrel kapjuk a következő állítást.

13.13. Állítás. *A Douglas–Gunn-séma forrástagot is tartalmazó*

$$\begin{cases} \left(I - \frac{r_x}{2} D^2_{0,x}\right) \mathbf{u}^{n+\frac{1}{3}} = \left(\frac{r_x}{2} D^2_{0,x} + \frac{r_y}{2} D^2_{0,y} + \frac{r_z}{2} D^2_{0,z}\right) \mathbf{u}^n + \frac{1}{2}(\mathbf{f}^n + \mathbf{f}^{n+1}) \\ \left(I - \frac{r_y}{2} D^2_{0,y}\right) \mathbf{u}^{n+\frac{2}{3}} = \mathbf{u}^{n+\frac{1}{3}} \\ \left(I - \frac{r_z}{2} D^2_{0,z}\right) (\mathbf{u}^{n+1} - \mathbf{u}^n) = \mathbf{u}^{n+\frac{2}{3}} \\ \mathbf{u}^{n+1} = \mathbf{u}^n + \mathbf{u}^{n+1} - \mathbf{u}^n \end{cases}$$

verziója másodrendben konzisztens a forrástagot is tartalmazó (13.1) *egyenlet háromdimenziós verziójával.*

További példák találhatók a [29] könyvben.

14. fejezet

Elsőrendű hiperbolikus egyenletek

Ebben a fejezetben

$$\partial_t u(t,\mathbf{x}) + \mathbf{a}\cdot\partial_{\mathbf{x}} u(t,\mathbf{x}) = f(t,\mathbf{x})$$

alakú elsőrendű hiperbolikus egyenleteket vizsgálunk egy, majd több dimenzióban, ahol $\mathbf{a}\in\mathbb{R}^d$ és $\mathbb{R}\times\mathbb{R}^n\to\mathbb{R}$ típusú f függvény adottak. A felépítés és a sémák vizsgálatára vonatkozó módszerek hasonlóak a parabolikus egyenletek esetéhez. Az erre vonatkozó feladatokat minden esetben kezdeti feltétellel látjuk el, továbbá $\mathbf{a}$ értékétől függően, illetve a kapcsolódó modell alapján különböző peremfeltételeket alkalmazhatunk. A megfelelő advekciós modellekben $\mathbf{a}$ az áramlás (ismert) sebessége.

A stabilitásvizsgálat során elegendő ennek homogén verziójával, az $f=0$ esettel foglalkozni. Ennek pontos megoldása gyakran kiszámítható a következő összefüggés alapján:

$$u(t,\mathbf{x}) = u(0,\mathbf{x}-t\mathbf{a});$$

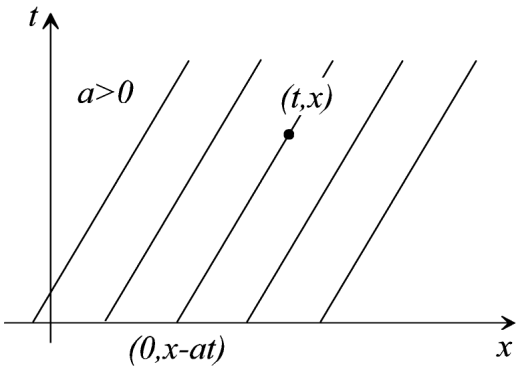
például abban az esetben, ha a megfelelő feladat $\mathbb{R}^d\times[0,T]$-on adott. Ezzel szoros kapcsolatban van az elsőrendű egyenletekre vonatkozó karakterisztikák elmélete, amely szerint a homogén feladat megoldása a karakterisztikák (14.1. és 14.2. ábrák) mentén állandó.

Ennek felhasználásával egyéb esetekben is viszonylag egyszerű a megfelelő feladatokra megoldóképletet adni. Ezért a következő vizsgálatok gyakorlati haszna nem az ilyen egyenletek közvetlen megoldása, hanem általánosítása azokban az esetekben, amikor az áramlást, hullámterjedést összetettebb modellek és egyenletek segítségével vizsgáljuk.

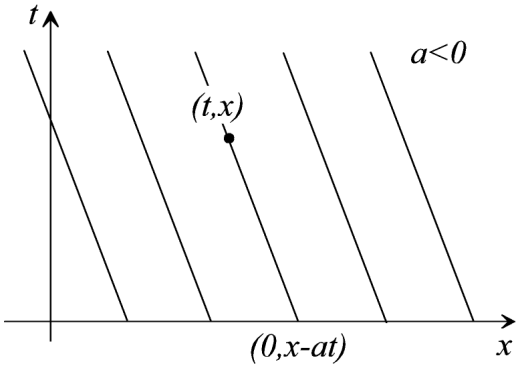
14.1. Hiperbolikus egyenletek 1 dimenzióban

Először az egész $\mathbb{R}$-en adott

$$\partial_t u(t,x) + a\partial_x u(t,x) = 0$$



14.1. ábra. Az egydimenziós advekciós egyenlet néhány karakterisztikája $a > 0$ esetén.



14.2. ábra. Az egydimenziós advekciós egyenlet néhány karakterisztikája $a < 0$ esetén.

problémával foglalkozunk, ahol $a \in \mathbb{R}$ adott. Az alábbi

$$\begin{cases} \partial_t u(t,x) + a\partial_x u(t,x) = 0, & t \in \mathbb{R}^+,\ x \in \mathbb{R} \\ u(0,x) = g(x), & x \in \mathbb{R} \end{cases} \tag{14.1}$$

feladatra vonatkozó vizsgálatot már végeztünk, ahol minden esetben a $R = a\frac{\delta}{h}$ jelöléssel élve a

$$u_k^{n+1} = (1+R)u_k^n - Ru_{k+1}^n \tag{14.2}$$

sémát alkalmaztuk. A 12.13. példa következményeként kaptuk, hogy ez csakis akkor lehet stabil, ha $a \in \mathbb{R}^-$. Továbbá ekkor a stabilitás feltétele mind a peremérték nélküli, mind az intervallum bal oldalán rögzített peremfeltétel mellett $-1 \leq R \leq 0$; ez utóbbit a 12.30. példában írtuk le.

Nem térünk ki a következő állítás bizonyítására, mert az a fenti eredményekhez használt gondolatmenet minimális változtatásával adódik; lásd ezzel kapcsolatban a 21.18. feladatot!

14.1. Állítás. *A (14.1) feladat megoldását közelítő*

$$u_k^{n+1} = u_k^n - RD_- u_k^n = (1-R)u_k^n + Ru_{k-1}^n \tag{14.3}$$

séma stabilitásának szükséges feltétele, hogy $a > 0$ legyen. Továbbá $a > 0$ esetén a séma pontosan akkor stabil, ha $0 < R \leq 1$.

Egy ilyen módszernek felel meg a 20.2.13. animáció.

Ha $a < 0$, akkor a (14.2) sémát *upwind* sémának nevezzük, mert a diszkretizáció során használt másik alappontot az áramlással ellentétes irányban vesszük fel, a (14.3) sémát pedig *downwind* sémának.

A továbbiakban egy olyan sémát vizsgálunk, amely első ránézésre jobbnak tűnik, mert mindenképpen az „áramlással szemben" is diszkretizálunk.

14.2. Állítás. *A (14.1) feladat megoldását közelítő*

$$u_k^{n+1} = u_k^n - RD_0 u_k^n = u_k^n - \frac{R}{2}(u_{k+1}^n - u_{k-1}^n) \tag{14.4}$$

séma semmilyen $R < 0$ érték esetén sem stabil.

Ez az eredmény kissé meglepő, hiszen itt a fenti előnyön kívül a konzisztenciarend is jobb, mint az upwind séma esetében. Erre a problémára vonatkozik a 21.19. feladat. Szeretnénk olyan sémát is, amely az idő- és a térváltozók szerint egyaránt másodrendű, emellett stabil is (esetleg valamilyen feltétellel).

Ennek konstrukciójához megjegyezzük, hogy ha $u \in C^2([0,T]\times\Omega)$, akkor a $\partial_t u(t,x) = -a\partial_x u(t,x)$ egyenlőség alapján

$$\partial_{tt} u(t,x) = -\partial_t a \partial_x u(t,x) = -a\partial_x \partial_t u(t,x) = a^2 \partial_x \partial_x u(t,x) = a^2 \partial_{xx} u(t,x). \tag{14.5}$$

A pontos megoldás δ változó szerinti Taylor-sorfejtését, a rá vonatkozó egyenletet, a fenti a (14.5) egyenlőséget, majd az első és a második deriváltak másodrendű közelítéseit

használva kapjuk, hogy

$$
\begin{aligned}
&u((n+1)\delta, kh) = u(n\delta, kh) + \delta\partial_t u(n\delta, kh) + \frac{\delta^2}{2}\partial_{tt} u(n\delta, kh) + \mathcal{O}(\delta^3)\\
&= u(n\delta, kh) - a\delta\partial_x u(n\delta, kh) + a^2\frac{\delta^2}{2}\partial_{xx} u(n\delta, kh) + \mathcal{O}(\delta^3)\\
&= u(n\delta, kh) - a\delta\left(\frac{u(n\delta, (k+1)h) - u(n\delta, (k-1)h)}{2h} + \mathcal{O}(h^2)\right)\\
&+ a^2\frac{\delta^2}{2}\left(\frac{u(n\delta, (k+1)h) - 2u(n\delta, kh) + u(n\delta, (k-1)h}{h^2} + \mathcal{O}(h^2)\right) + \mathcal{O}(\delta^3)\\
&= u_k^n - \frac{R}{2}(u(n\delta, (k+1)h) - u(n\delta, (k-1)h)) + \delta\mathcal{O}(h^2) +\\
&+ \frac{R^2}{2}(u(n\delta, (k+1)h) - 2u(n\delta, kh) + u(n\delta, (k-1)h)) + \delta^2\mathcal{O}(h^2) + \mathcal{O}(\delta^3)\\
&= \left(\frac{R}{2} + \frac{R^2}{2}\right)u(n\delta, (k-1)h) + (1-R^2)u(n\delta, kh) +\\
&+ (\frac{R^2}{2} - \frac{R}{2})u(n\delta, (k+1)h) + \delta\mathcal{O}(\delta h^2) + \delta^2\mathcal{O}(h^2) + \mathcal{O}(\delta^3).
\end{aligned}
$$

A levezetés nyilvánvaló következményeként kapjuk a következő állítást.

14.3. Állítás. *Az alábbi*

$$u_k^{n+1} = (\frac{R}{2} + \frac{R^2}{2})u_{k-1}^n + (1-R^2)u_k^n + (\frac{R^2}{2} - \frac{R}{2})u_{k+1}^n \tag{14.6}$$

Lax–Wendroff-séma *mindkét változója szerint másodrendben konzisztens a (14.1) egyenlettel.*

A (14.6) séma stabilitására vonatkozik a következő eredmény.

14.4. Állítás. *A (14.1) feladat megoldására vonatkozó* (14.6) *Lax–Wendroff-séma pontosan akkor stabil, ha* $|R| \leq 1$.

Bizonyítás. A (14.6) sémát a számolásokhoz az

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \frac{R}{2}D_0\mathbf{u}^n + \frac{R^2}{2}D_0^2\mathbf{u}^n$$

alakba írjuk. Itt mindkét oldal diszkrét idejű Fourier-transzformáltját véve kapjuk, hogy

$$\mathcal{F}\mathbf{u}^{n+1}(s) = \mathcal{F}\mathbf{u}^n(s)\left(1 - \frac{R}{2}\cdot 2i\sin s + \frac{R^2}{2}\cdot 2(\cos s - 1)\right),$$

azaz

$$\begin{aligned}|\rho|^2(s) &= (R\sin s)^2 + (1 + R^2(\cos s - 1))^2 = \\ &= R^2\sin^2 s + 1 + 2R^2(\cos s - 1) + R^4(\cos s - 1)^2 = \\ &= 1 + R^2(\sin^2 s + 2\cos s - 2 + R^2(\cos^2 s + 1 - 2\cos s)) = \\ &= 1 + R^2(-\cos^2 s + 2\cos s - 1 + R^2(\cos^2 s + 1 - 2\cos s)) = \\ &= 1 + R^2(\cos s - 1)^2(R^2 - 1).\end{aligned}$$

Ez pontosan akkor lesz legfeljebb 1 minden s értékre, ha $R^2 - 1 \le 0$, azaz ha $|R| \le 1$, ahogy a tételben állítottuk. □

Ehhez kapcsolódik a 20.2.14. és a 20.2.16. animáció.

14.5. Megjegyzés.

1. A Lax–Wendroff-séma leginkább előnyös oldala az, hogy a feladatban szereplő a együttható előjelétől függetlenül használható. Ez különösen fontos akkor, ha olyan feladatok egyes lépéseiben, részfeladataiban alkalmazzuk, ahol a megfelelő fizikai problémában az áramlás maga az ismeretlen, így irányát nem ismerjük, és az természetes módon változhat is.

2. A Lax–Wendroff-séma két dimenziós kiterjesztését tárgyalja a 21.12. feladat.

3. A 14.1 feladatra vonatkozó további érdekes sémát tartalmaz a 21.20. feladat. ♢

14.2. Implicit sémák vizsgálata

A 14.2. állítás negatív eredménye után megpróbáljuk az ott felírt (14.4) sémát javítani.

14.6. Állítás. *A (14.1) feladat megoldásához tartozó*

$$u_k^{n+1} = u_k^n - RD_0u_k^{n+1} = u_k^n + \frac{R}{2}(u_{k+1}^{n+1} - u_{k-1}^{n+1}) \tag{14.7}$$

séma feltétel nélkül stabil.

Bizonyítás. A séma két oldalának diszkrét idejű Fourier-transzformációjával kapjuk, hogy

$$\mathcal{F}\mathbf{u}^{n+1}(s) = \mathcal{F}\mathbf{u}^n(s) - 2Ri\cdot\sin s\cdot\mathcal{F}\mathbf{u}^{n+1}(s),$$

vagyis

$$\rho(s) = \frac{\mathcal{F}\mathbf{u}^{n+1}(s)}{\mathcal{F}\mathbf{u}^n(s)} = \frac{1}{1 + 2Ri\cdot\sin s}, \tag{14.8}$$

ahol $|1 + 2Ri \cdot \sin s| \geq 1$, azaz reciprokára a

$$|\rho(s)| = \frac{1}{|1 + 2Ri \cdot \sin s|} \leq 1$$

egyenlőtlenség teljesül. Ebből pedig valóban következik, hogy a fenti (14.7) séma stabil.□

Hasonlóan vizsgálhatjuk a Crank–Nicolson típusú módosítást is.

14.7. Állítás. *A (14.1) feladat megoldását közelítő*

$$u_k^{n+1} + \frac{R}{4} D_0 u_k^{n+1} = u_k^n - \frac{R}{4} D_0 u_k^n \tag{14.9}$$

séma feltétel nélkül stabil.

Bizonyítás. A séma két oldalának diszkrét idejű Fourier-transzformációjával kapjuk, hogy

$$\mathcal{F}\mathbf{u}^{n+1}(s)(1 + i \cdot \frac{R}{2} \sin s) = \mathcal{F}\mathbf{u}^n(s)(1 - i \cdot \frac{R}{2} \sin s),$$

vagyis

$$\rho(s) = \frac{\mathcal{F}\mathbf{u}^{n+1}(s)}{\mathcal{F}\mathbf{u}^n(s)} = \frac{1 - i \cdot \frac{R}{2} \sin s}{1 + i \cdot \frac{R}{2} \sin s}, \tag{14.10}$$

ami két olyan komplex szám hányadosa, amelyek egymás konjugáltjai, vagyis minden s-re $|\rho(s)| = 1$. Ebből pedig valóban következik, hogy a fenti séma stabil. □

14.8. Megjegyzés. A gyakorlatban problémát okozhat, ha $|\rho(s)| = 1$ minden s-re. A mátrixokkal való műveletek során ugyanis felléphet olyan kerekítési hiba, ami miatt minden lépésben nem 1-gyel, hanem egy 1-nél kicsit nagyobb értékkel szorzódik a Fourier-transzformált. Ekkor a gyakorlatban nem kapunk olyan konvergenciát, amire az itt tárgyalt elmélet alapján számítunk. ♢

14.2.1. Kétoldali implicit sémák korlátos tartományokon

Egydimenziós hiperbolikus feladatokra vonatkozó kétoldali sémák vizsgálatakor azzal a problémával szembesülünk, hogy egy korrekt kitűzésű feladat esetén nem adott peremfeltétel a vizsgált tartomány (intervallum) mindkét oldalán.

Ezért ha csak az egyik oldalon adott a peremfeltétel, akkor meg kell változtatni a sémát a másik oldalon, hogy a perem mellett levő rácspontokban a közelítés kiszámítható legyen. Ezt felfoghatjuk úgy is, hogy a jobb oldalon mesterségesen valamilyen peremfeltételt konstruálunk, ezért az ezt leíró egyenlőségeket az irodalomban *numerikus peremfeltétel*nek is nevezik.

Ezt az összetett problémát nem tárgyaljuk általánosan, a stabilitásra vonatkozó feltételek bizonyítása helyett egy konkrét példán szemléltetjük a megfelelő eljárást.

14.9. Példa. Legyen $\Omega = (0,1)$ és a pozitív! Ekkor adott folytonos $g : (0,1) \to \mathbb{R}$ és $u_0 : (0,T) \to \mathbb{R}$ esetén a

$$\begin{cases} \partial_t u(t,x) + a\partial_x u(t,x) = 0, & t \in (0,T),\ x \in (0,1) \\ u(0,x) = g(x), & x \in (0,1) \\ u(t,0) = u_0(t), & t \in (0,T) \end{cases} \tag{14.11}$$

korrekt kitűzésű feladat numerikus megoldását vizsgáljuk. Ezt T-ig az $(x_1, x_2, \ldots, x_N)$ belső pontokban számítjuk ki, ahol

$$x_0 = 0,\ x_1 = \frac{1}{N+1},\ \ldots,\ x_N = \frac{Nh}{N+1},\ x_{N+1} = 1,$$

azaz $\Omega_h = \{x_0, x_1, \ldots, x_{N+1}\}$. Ekkor a Lax–Wendroff-sémában (14.6) alapján, valamint a peremfeltétel figyelembevételével

$$\begin{aligned} u_1^{n+1} &= \left(\frac{R}{2} + \frac{R^2}{2}\right) u_0^n + (1-R^2)u_1^n + \left(\frac{R^2}{2} - \frac{R}{2}\right) u_2^n = \\ &= \left(\frac{R}{2} + \frac{R^2}{2}\right) u_0(n\delta) + (1-R^2)u_1^n + \left(\frac{R^2}{2} - \frac{R}{2}\right) u_2^n, \end{aligned}$$

amely az előző időlépésben kapott értékekből kiszámítható, ugyanakkor

$$u_N^{n+1} = (\frac{R}{2} + \frac{R^2}{2})u_{N-1}^n + (1-R^2)u_N^n + (\frac{R^2}{2} - \frac{R}{2})u_{N+1}^n,$$

ahol u_{N+1}^n nincs megadva. Emiatt az intervallum jobb szélén az alábbi közelítést alkalmazzuk:

$$u_N^{n+1} = Ru_{N-1}^n + (1-R)u_N^n. \tag{14.12}$$

A (14.12) képletben szereplő egyenlőség azonban hatással lehet a módszer stabilitására.

Hasonló mondható akkor is, ha implicit módszert tekintünk. Bizonyítás nélkül említjük meg, hogy a (14.12) képlettel definiált esetben a séma stabil, azonban ha (14.12) helyett az

$$u_N^{n+1} = u_{N-1}^{n+1}.$$

egyenlőséget használjuk, akkor a megfelelő séma nemcsak alacsony rendben lesz konzisztens, hanem instabillá is válik. ♢

Az ehhez kapcsolódó elmélet részletei a [28] könyvben találhatók.

14.2.2. Kétoldali implicit sémák periodikus peremfeltétellel

Az előző szakaszban tárgyalt példával ellentétben a kétoldali sémák természetes módon használhatók abban az esetben, ha a (14.11) feladatban a peremfeltételt periodikusra cseréljük, azaz a

$$\begin{cases} \partial_t u(t,x) + a\partial_x u(t,x) = 0, \quad t \in (0,T),\ x \in (0,1) \\ u(0,x) = g(x), \quad x \in (0,1) \\ u(t,0) = u(t,1), \quad t \in (0,T) \end{cases} \tag{14.13}$$

feladat numerikus megoldását végezzük el így. A megfelelő lépésmátrixok felírását két példán szemléltetjük.

14.10. Példa. A (14.13) feladat megoldása implicit Lax–Wendroff-séma segítségével. Ekkor az időlépést a

$$B_+ \mathbf{u}^{n+1} = B_0 \mathbf{u}^n$$

képlettel adjuk meg, ahol az egyes mátrixok a következők:

$$B_0 = \begin{pmatrix} 1-\frac{R^2}{2} & \frac{R^2}{4}-\frac{R}{4} & 0 & 0 & 0 & \frac{R^2}{4}+\frac{R}{4} \\ \frac{R^2}{4}+\frac{R}{4} & 1-\frac{R^2}{2} & \frac{R^2}{4}-\frac{R}{4} & 0 & 0 & 0 \\ 0 & \frac{R^2}{4}+\frac{R}{4} & 1-\frac{R^2}{2} & \frac{R^2}{4}-\frac{R}{4} & 0 & 0 \\ 0 & 0 & \frac{R^2}{4}+\frac{R}{4} & 1-\frac{R^2}{2} & \frac{R^2}{4}-\frac{R}{4} & 0 \\ 0 & 0 & 0 & \frac{R^2}{4}+\frac{R}{4} & 1-\frac{R^2}{2} & \frac{R^2}{4}-\frac{R}{4} \\ \frac{R^2}{4}-\frac{R}{4} & 0 & 0 & 0 & \frac{R^2}{4}+\frac{R}{4} & 1-\frac{R^2}{2} \end{pmatrix},$$

továbbá

$$B_+ = \begin{pmatrix} 1+\frac{R^2}{2} & -\frac{R^2}{4}+\frac{R}{4} & 0 & 0 & 0 & -\frac{R^2}{4}-\frac{R}{4} \\ -\frac{R^2}{4}-\frac{R}{4} & 1+\frac{R^2}{2} & -\frac{R^2}{4}+\frac{R}{4} & 0 & 0 & 0 \\ 0 & -\frac{R^2}{4}-\frac{R}{4} & 1+\frac{R^2}{2} & -\frac{R^2}{4}+\frac{R}{4} & 0 & 0 \\ 0 & 0 & -\frac{R^2}{4}-\frac{R}{4} & 1+\frac{R^2}{2} & -\frac{R^2}{4}+\frac{R}{4} & 0 \\ 0 & 0 & 0 & -\frac{R^2}{4}-\frac{R}{4} & 1+\frac{R^2}{2} & -\frac{R^2}{4}+\frac{R}{4} \\ -\frac{R^2}{4}+\frac{R}{4} & 0 & 0 & 0 & -\frac{R^2}{4}-\frac{R}{4} & 1+\frac{R^2}{2} \end{pmatrix},$$

ahol a mátrix első, illetve utolsó sorában figyelembe vettük a periodikus peremfeltételből adódó $u_0 = u_N$ és az $u_1 = u_{N+1}$ egyenlőséget. ♢

14.11. Példa. A (14.13) feladat megoldása a (14.9) formulának megfelelő implicit Crank–Nicolson-séma segítségével.
Ekkor az időlépést a

$$A_+ \mathbf{u}^{n+1} = A_0 \mathbf{u}^n \tag{14.14}$$

képlettel adjuk meg, ahol az egyes mátrixok a következők:

$$A_0 = \begin{pmatrix} 1 & \frac{R}{4} & 0 & 0 & 0 & -\frac{R}{4} \\ -\frac{R}{4} & 1 & \frac{R}{4} & 0 & 0 & 0 \\ 0 & -\frac{R}{4} & 1 & \frac{R}{4} & 0 & 0 \\ 0 & 0 & -\frac{R}{4} & 1 & \frac{R}{4} & 0 \\ 0 & 0 & 0 & -\frac{R}{4} & 1 & \frac{R}{4} \\ \frac{R}{4} & 0 & 0 & 0 & -\frac{R}{4} & 1 \end{pmatrix},$$

továbbá

$$A_+ = \begin{pmatrix} 1 & -\frac{R}{4} & 0 & 0 & 0 & \frac{R}{4} \\ \frac{R}{4} & 1 & -\frac{R}{4} & 0 & 0 & 0 \\ 0 & \frac{R}{4} & 1 & -\frac{R}{4} & 0 & 0 \\ 0 & 0 & \frac{R}{4} & 1 & -\frac{R}{4} & 0 \\ 0 & 0 & 0 & \frac{R}{4} & 1 & -\frac{R}{4} \\ -\frac{R}{4} & 0 & 0 & 0 & \frac{R}{4} & 1 \end{pmatrix},$$

ahol a mátrixok első és utolsó sorában most is figyelembe vettük a periodikus peremfeltételből adódó $u_0 = u_N$ és az $u_1 = u_{N+1}$ egyenlőséget. ♢

Természetes módon merül fel a kérdés, hogy a fenti peremfeltétellel ellátott problémák esetében is megmarad-e a vizsgált implicit sémák stabilitása. Az utóbbi példa esetében adunk erre választ.

14.12. Állítás. *A (14.14) formulával megadott séma feltétel nélkül stabil.*

Bizonyítás. Legyen

$$Q = \begin{pmatrix} 0 & -\frac{R}{4} & 0 & 0 & 0 & \frac{R}{4} \\ \frac{R}{4} & 0 & -\frac{R}{4} & 0 & 0 & 0 \\ 0 & \frac{R}{4} & 0 & -\frac{R}{4} & 0 & 0 \\ 0 & 0 & \frac{R}{4} & 1 & -\frac{R}{4} & 0 \\ 0 & 0 & 0 & \frac{R}{4} & 0 & -\frac{R}{4} \\ -\frac{R}{4} & 0 & 0 & 0 & \frac{R}{4} & 0 \end{pmatrix},$$

amellyel a (14.14) séma az

$$(I+Q)\mathbf{u}^{n+1} = (I-Q)\mathbf{u}^n$$

alakba írható, vagyis a megfelelő lépésmátrix $(I+Q)^{-1}(I-Q)$. Ekkor a (12.14) azonosság, valamint a $Q^* = -Q$ egyenlőség alapján kapjuk, hogy

$$\begin{aligned} &\|(I+Q)^{-1}(I-Q)\|_2^2 = \sigma((I+Q)^{-1}(I-Q)[(I+Q)^{-1}(I-Q)]^*) = \\ &= \sigma((I+Q)^{-1}(I-Q)(I-Q)^*[(I+Q)^{-1}]^*) = \sigma((I+Q)^{-1}(I-Q)(I+Q)[(I+Q)^{-1}]^*) = \\ &= \sigma((I+Q)^{-1}(I+Q)(I-Q)[(I+Q)^{-1}]^*) = \sigma((I-Q)[(I+Q)^{-1}]^*) = \\ &= \sigma((I+Q)^{-1}(I-Q)^*) = \sigma((I+Q)^{-1}(I+Q)) = 1, \end{aligned}$$

amivel a bizonyítandó állítást beláttuk. □

14.13. Megjegyzés. A 14.7. állításhoz hasonlóan itt is azt kaptuk, hogy a lépésmátrix l_2-normája minden esetben 1. ♢

14.2.3. A Sherman–Morrison-algoritmus

Egy eljárást mutatunk a 14.10. és a 14.11. példákban szereplő implicit sémák hatékony numerikus kezelésére. Mindkét esetben olyan mátrixot kell invertálnunk, amelyben három nemnulla átlón kívül csak 1-1 nemnulla elem található. Erre vonatkozólag az alábbi általános lemma eredményét alkalmazzuk.

14.14. Lemma. *Legyen $B \in \mathbb{R}^{n\times n}$ invertálható mátrix, továbbá $A = B - \mathbf{w}\mathbf{z}^T$ valamilyen $\mathbf{w}, \mathbf{z} \in \mathbb{R}^n$ vektorokkal! Ekkor*

$$A^{-1} = B^{-1} + \frac{1}{1 - \mathbf{z}^T B^{-1}\mathbf{w}} B^{-1}\mathbf{w}\mathbf{z}^T B^{-1}.$$

Bizonyítás. A bizonyításhoz egyszerűen az $A^{-1}A$ szorzatot számoljuk ki, ahol A^{-1} helyébe a fenti formulát helyettesítjük. Nyilvánvaló átalakításokkal kapjuk, hogy

$$\begin{aligned}
A^{-1}A &= \left(B^{-1} + \frac{1}{1 - \mathbf{z}^T B^{-1}\mathbf{w}} B^{-1}\mathbf{w}\mathbf{z}^T B^{-1}\right) \cdot (B - \mathbf{w}\mathbf{z}^T) = \\
&= I - B^{-1}\mathbf{w}\mathbf{z}^T + \frac{1}{1 - \mathbf{z}^T B^{-1}\mathbf{w}} B^{-1}\mathbf{w}\mathbf{z}^T B^{-1} \cdot (B - \mathbf{w}\mathbf{z}^T) = \\
&= I - B^{-1}\mathbf{w}\mathbf{z}^T + \frac{1}{1 - \mathbf{z}^T B^{-1}\mathbf{w}} B^{-1}\mathbf{w}\mathbf{z}^T - \frac{1}{1 - \mathbf{z}^T B^{-1}\mathbf{w}} B^{-1}\mathbf{w}\mathbf{z}^T B^{-1}\mathbf{w}\mathbf{z}^T) = \\
&= I + B^{-1}\mathbf{w}\mathbf{z}^T \left(-I + \frac{1}{1 - \mathbf{z}^T B^{-1}\mathbf{w}} I - \frac{\mathbf{z}^T B^{-1}\mathbf{w}}{1 - \mathbf{z}^T B^{-1}\mathbf{w}} I\right) = \\
&= I + B^{-1}\mathbf{w}\mathbf{z}^T I \left(-1 + \frac{1 - \mathbf{z}^T B^{-1}\mathbf{w}}{1 - \mathbf{z}^T B^{-1}\mathbf{w}}\right) = I,
\end{aligned}$$

ami bizonyítja a lemmát. □

14.15. Példa. A 14.11. példában felírt séma esetén

$$A_+ = A_{+,0} - \mathbf{w}\mathbf{z}^T,$$

ahol

$$A_{+,0} = \begin{pmatrix}
1 - \frac{R}{4} & -\frac{R}{4} & 0 & 0 & 0 & 0 \\
\frac{R}{4} & 1 & -\frac{R}{4} & 0 & 0 & 0 \\
0 & \frac{R}{4} & 1 & -\frac{R}{4} & 0 & 0 \\
0 & 0 & \frac{R}{4} & 1 & -\frac{R}{4} & 0 \\
0 & 0 & 0 & \frac{R}{4} & 1 & -\frac{R}{4} \\
0 & 0 & 0 & 0 & \frac{R}{4} & 1 + \frac{R}{4}
\end{pmatrix},$$

továbbá

$$\mathbf{w} = (-1\ 0\ \ldots\ 0\ 1)^T \quad \text{és} \quad \mathbf{z}^T = \left(\frac{R}{4}\ 0\ \ldots\ 0\ \frac{R}{4}\right). \qquad \diamondsuit$$

Így a 14.14. lemma alapján visszavezettük a (14.14) rendszer megoldását olyan rendszer megoldására, ahol elegendő a tridiagonális $A_{+,0}$ mátrix inverzét kiszámolni.

14.16. Megjegyzés. A gyakorlatban általában még mátrixok inverzét sem számoljuk ki, ugyanis egy lineáris rendszer megoldásához erre nincs szükség. $\diamondsuit$

A fenti eredmények alapján az egyes sémák tulajdonságait a következő táblázatban foglaljuk össze.

14.1. táblázat. Három különböző séma tulajdonságai az egydimenziós (14.1) hiperbolikus feladat megoldásának numerikus közelítésére.

módszer	komplexitás	stabilitás feltétele	rend időben	rend térben
upwind explicit Euler	explicit	$R \geq -1$	1	1
downwind explicit Euler	explicit	instabil	1	1
centrális explicit Euler	explicit	instabil	1	2
Lax–Wendroff	explicit	$R \geq -1$	2	2
centrális implicit Euler	implicit	-	1	2
centrális Crank–Nicolson	implicit	-	2	2

14.3. Hiperbolikus egyenletek magasabb dimenzióban: ADI sémák

A két dimenzióban adott

$$\partial_t u(t,x,y) + a\,\partial_x u(t,x,y) + b\,\partial_y u(t,x,y) = 0 \tag{14.15}$$

hiperbolikus egyenlet numerikus megoldását vizsgáljuk, ahol az

$$R_x = a\frac{\delta}{h_x} \quad \text{és} \quad R_y = b\frac{\delta}{h_y}$$

jelöléseket alkalmazzuk majd. A standard Crank–Nicolson-sémát fogjuk felírni, és abból approximatív faktorizációval egy ADI típusú sémát konstruálunk.

A két dimenzióban adott (14.15) egyenletre vonatkozó Crank–Nicolson-séma a következő

$$\left(1+\frac{R_x}{4}D_{0,x}+\frac{R_y}{4}D_{0,y}\right)u_{j,k}^{n+1}=\left(1-\frac{R_x}{4}D_{0,x}-\frac{R_y}{4}D_{0,y}\right)u_{j,k}^{n}, \tag{14.16}$$

amelyről nyilvánvaló (de kissé összetett) számolással látszik, hogy $\mathcal{O}(\delta^2)+\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)$ rendben konzisztens a (14.15) egyenlettel. Átalakítva a (14.16) egyenletet kapjuk, hogy

$$\begin{aligned}&\left(I+\frac{R_x}{4}D_{0,x}\right)\left(I+\frac{R_y}{4}D_{0,y}\right)\mathbf{u}^{n+1}-\frac{R_x}{4}D_{0,x}\frac{R_y}{4}D_{0,y}\mathbf{u}^{n+1}=\\&=\left(I-\frac{R_x}{4}D_{0,x}\right)+\left(I-\frac{R_y}{4}D_{0,y}\right)\mathbf{u}^{n}-\frac{R_x}{4}D_{0,x}\frac{R_y}{4}D_{0,y}\mathbf{u}^{n},\end{aligned}$$

azaz

$$\begin{aligned}&\left(I+\frac{R_x}{4}D_{0,x}\right)\left(I+\frac{R_y}{4}D_{0,y}\right)\mathbf{u}^{n+1}=\\&=\left(I-\frac{R_x}{4}D_{0,x}\right)\left(I-\frac{R_y}{4}D_{0,y}\right)\mathbf{u}^{n}+\frac{R_x}{4}D_{0,x}\frac{R_y}{4}D_{0,y}(\mathbf{u}^{n+1}-\mathbf{u}^{n}).\end{aligned} \tag{14.17}$$

Fennáll továbbá, hogy

$$\frac{R_xR_y}{16}D_{0,x}D_{0,y}(\mathbf{u}^{n+1}-\mathbf{u}^{n})=\delta^3\frac{ab}{16h_xh_y}D_{0,x}D_{0,y}\frac{1}{\delta}D_{+,t}\mathbf{u}^{n},$$

vagyis a lemma 5. pontja alapján kapjuk, hogy

$$\delta^3\frac{ab}{16h_xh_y}D_{0,x}D_{0,y}\frac{1}{\delta}D_{+,t}\mathbf{u}^{n}=\delta^3(\frac{R_xR_y}{4}\partial_x\partial_y\partial_t u(n\delta,x_j,y_k)+\mathcal{O}(h_x^2)+\mathcal{O}(h_y^2)),$$

tehát a (14.16) és így a (14.17) séma minden változója szerint másodrendben konzisztens. Emiatt a (14.17) sémából nyert

$$\left(I+\frac{R_x}{4}D_{0,x}\right)\left(I+\frac{R_y}{4}D_{0,y}\right)\mathbf{u}^{n+1}=\left(I-\frac{R_x}{4}D_{0,x}\right)\left(I-\frac{R_y}{4}D_{0,y}\right)\mathbf{u}^{n}$$

illetve a vele ekvivalens

$$\begin{cases}\left(I+\frac{R_x}{4}D_{0,x}\right)\mathbf{u}^{n+\frac{1}{2}}=\left(I-\frac{R_x}{4}D_{0,x}\right)\left(I-\frac{R_y}{4}D_{0,y}\right)\mathbf{u}^{n}\\\left(I+\frac{R_y}{4}D_{0,y}\right)\mathbf{u}^{n+1}=\mathbf{u}^{n+\frac{1}{2}}\end{cases}$$

Beam–Warming-séma másodrendben konzisztens.

14.17. Állítás. *A Beam–Warming-séma feltétel nélkül stabil.*

Bizonyítás. A séma mindkét oldalára diszkrét idejű Fourier-transzformációt alkalmazva kapjuk, hogy

$$\begin{cases} \mathcal{F}\mathbf{u}^{n+\frac{1}{2}}(\alpha,\beta)\left(1+i\cdot\frac{R_x}{2}\sin\alpha\right)=\mathcal{F}\mathbf{u}^{n}(\alpha,\beta)\left(1-i\cdot\frac{R_x}{2}\sin\alpha\right)\left(1-i\cdot\frac{R_y}{2}\sin\beta\right) \\ \mathcal{F}\mathbf{u}^{n+\frac{1}{2}}(\alpha,\beta)\left(1+i\cdot\frac{R_y}{2}\sin\beta\right)=\mathcal{F}\mathbf{u}^{n+\frac{1}{2}}(\alpha,\beta), \end{cases}$$

amelyeket összeszorozva, majd egyszerűsítve nyerjük a

$$\rho(\alpha,\beta)=\frac{\mathcal{F}\mathbf{u}^{n+1}(\alpha,\beta)}{\mathcal{F}\mathbf{u}^{n}(\alpha,\beta)}=\frac{\left(1-i\cdot\frac{R_x}{2}\sin\alpha\right)\left(1-i\cdot\frac{R_y}{2}\sin\beta\right)}{\left(1+i\cdot\frac{R_x}{2}\sin\alpha\right)\left(1+i\cdot\frac{R_y}{2}\sin\beta\right)}$$

egyenlőséget, ahol a jobb oldal egy komplex számnak és konjugáltjának hányadosa, vagyis abszolút értéke 1. Ezzel az állítást beláttuk. □

Két dimenziós hiperbolikus probléma numerikus megoldására vonatkozó kézenfekvő módszer lehet a Lax–Wendroff-séma általánosítása. Ezt tárgyalja a 21.12. feladat.

Hiperbolikus egyenletekre vonatkozó sémák kvalitatív tulajdonságait elemzik a [28] és az [1] könyvek.

15. fejezet

A függési tartományok vizsgálata

Láttuk az előző fejezetekben, hogy a parabolikus és a hiperbolikus egyenletekre konstruált összes egylépéses explicit séma csak valamilyen feltétel mellett volt stabil. Azt vizsgáljuk meg a következőkben, hogy vajon ez törvényszerű, vagy ügyesebb konstrukcióval nyerhetnénk akár feltétel nélkül konvergens explicit sémákat is.

A módszer az lesz, hogy összehasonlítjuk azt a tartományt, amelytől a közelítés során, illetve az eredeti feladatban egy rögzített helyen kapott megoldás függ. Ezt az általános

$$\begin{cases} \partial_t u(t,\mathbf{x}) = Lu(t,\mathbf{x}) \quad t \in (0,T),\ \mathbf{x} \in \Omega \\ u(0,\mathbf{x}) = u_0(\mathbf{x}) \quad \mathbf{x} \in \Omega \end{cases} \tag{15.1}$$

típusú probléma esetén vizsgáljuk, ahol u_0 adott, és feltesszük, hogy a probléma korrekt kitűzésű.

15.1. Definíció. Egy $(t,\mathbf{x})$ ponthoz tartozó analitikus függési tartománynak nevezzük és $\Omega_{t,\mathbf{x}}$-szel jelöljük azt az Ω-beli halmazt, amelyre $u(t,\mathbf{x})$ értéke $u_0|_{\Omega_{t,\mathbf{x}}}$-től ténylegesen függ. Ezt úgy értjük, hogy $\Omega_{t,\mathbf{x}}$ azon legszűkebb halmaz, amelyre igaz, hogy

$$u_0|_{\Omega_{t,\mathbf{x}}} = v_0|_{\Omega_{t,\mathbf{x}}} \Rightarrow u(t,\mathbf{x}) = v(t,\mathbf{x}). \qquad \diamondsuit$$

A következőkben a függési tartomány numerikus megfelelőjét definiáljuk.

15.2. Definíció. Jelölje $H_{\mathbf{h},\delta} \subset \Omega_{\mathbf{h}}$ azt a legszűkebb halmazt, amelyre $u^n_{\mathbf{k}}$ függ $u_0|_{\Omega_h}$-tól a $\mathbf{h}$ és δ paraméterekkel rendelkező felosztás mellett. Ekkor az

$$\overline{\bigcup_{\delta,\mathbf{h}} H_{\mathbf{h},\delta}}$$

halmazt az $(n\delta, \mathbf{k} \otimes \mathbf{h})$ ponthoz tartozó numerikus függési tartománynak nevezzük, ahol azon felosztásparaméterekre vesszük az uniót, amelyeket a numerikus módszerben használhatunk. $\diamondsuit$

15.3. Megjegyzés. A $H_{\mathbf{h},\delta}$ halmaz jelölése az egyszerűség kedvéért nem tartalmazza az n-től és k-tól való függést. ♢

A konvergenciára vonatkozó szükséges feltételt valamilyen értelemben homogén feladatokra igazoljuk.

15.4. Tétel. *Tegyük fel, hogy az eredeti (15.1) feladat olyan, hogy $u_0 = 0$ esetén $u(t, \mathbf{x}) = 0$, továbbá azonosan nulla kezdeti feltétel mellett a numerikus megoldás is nulla. Ekkor a következő értelemben vett*

$$\lim_{\substack{n\delta=t,\ \mathbf{k}\otimes\mathbf{h}=\mathbf{x} \\ \delta\to 0,\ \mathbf{h}\to 0}} u_{\mathbf{k}}^n = u(t, \mathbf{x})$$

konvergencia szükséges feltétele az, hogy a $(t, \mathbf{x}) = (n\delta, \mathbf{k} \otimes \mathbf{h})$ ponthoz tartozó analitikus függési tartomány része az ehhez a ponthoz tartozó numerikus függési tartománynak.

Bizonyítás. Tegyük fel, hogy a tartalmazás nem teljesül! Ekkor van olyan $\mathbf{x}_0$ pontja az analitikus függési tartománynak, amely a numerikusban nincsen benn. Legyen továbbá a kezdeti feltétel a numerikus függési tartományon nulla! Válasszuk úgy az $\mathbf{x}_0$-beli kezdeti feltételt, hogy az ebből kapott megoldás a $(t, \mathbf{x})$ helyen nem nulla. Ez nyilván megtehető, mivel $\mathbf{x}_0$ a 15.1. definíció értelmében az analitikus függési tartomány része. Emellett a numerikus megoldásra tetszőleges közelítés esetén a feltételből kapjuk, hogy $u_{\mathbf{k}}^n = 0$, vagyis a konvergencia nem teljesül. □

A fenti szükséges feltételt gyakran Courant–Friedrichs–Lévy-feltételnek (CFL-feltételnek) is nevezik.

15.5. Megjegyzés.

1. A tétel haszna az, hogy akár nemlineáris egyenletek esetében is alkalmazható.
2. A tétel feltételei nem szigorúak. Ha ugyanis a sémára vonatkozó feltétel nem teljesülne, akkor lineáris feladatok esetén az nem lehetne stabil, így konvergens sem. ♢

Megvizsgáljuk néhány példán a fenti fogalmakat és a CFL-feltétel alkalmazását.

15.6. Példa. Először a (14.1) feladatra vonatkozó, $a < 0$ esetén alkalmazott

$$u_k^{n+1} = u_k^n - aR(u_{k+1}^n - u_k^n)$$

sémát vizsgáljuk, amelyet egy egyenletes felosztású rácson diszkretizálunk.

Tudjuk, hogy a diszkretizációtól függetlenül a (t, x) ponthoz tartozó analitikus függési tartomány $x - at$, mert az egyenlet pontos megoldása $u(t, x) = u(0, x - at)$, lásd a 14.2. ábrát is!

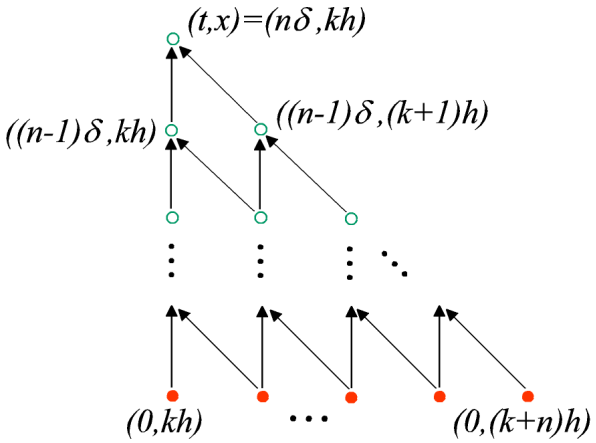
A numerikus függési tartomány kiszámításához először a $H_{\mathbf{h},\delta}$ halmazt határozzuk meg. Feltesszük, hogy $n\delta = t$. A séma szerint u_k^n az előző időlépésben vett u_k^{n-1} és u_{k+1}^{n-1} értékektől függ. Ezek pedig az előző időlépésben vett u_k^{n-2}, u_{k+1}^{n-2}, és u_{k+1}^{n-2}, u_{k+2}^{n-2} értékektől függnek. Ezt folytatva kapjuk, hogy u_k^n a kezdetben adott értékek közül a

$$u_k^0, u_{k+1}^0, u_{k+2}^0, \ldots, u_{k+n}^0$$

értékektől függ, ahol az első és az utolsó alappont különbsége nh; lásd a 15.1. ábrát! Mivel az R paraméterre vonatkozó feltételt akarunk levezetni, mindenhol ennek függvényében írjuk fel az egyes mennyiségeket. Így

$$nh = \frac{T}{\delta} \cdot h = t \cdot \frac{a}{R},$$

ami rögzített R esetén a felosztástól független. Vagyis a $H_{h,\delta}$ halmaz olyan osztópontokból áll, amelyek közül a bal szélső x, a jobb szélső (a felosztás h és δ paramétereitől függetlenül) $x + t \cdot \frac{a}{R}$, az osztópontok távolsága pedig h. Ha ezen pontok unióját vesszük minden lehetséges $h > 0$ esetén, akkor az így kapott halmaz lezártja az $\left[x, x + t \cdot \frac{a}{R}\right]$ intervallum. A $H_{h,\delta}$ halmazt mutatja egy lehetséges paraméterpár esetén a 15.1. ábra.



15.1. ábra. A (t, x) ponthoz tartozó $H_{\delta,h}$ halmaz a 15.6. példában a numerikus függési halmaz meghatározásához. Ezeket piros pontok, a közelítések során használt rácspontokat zöld körök jelölik.

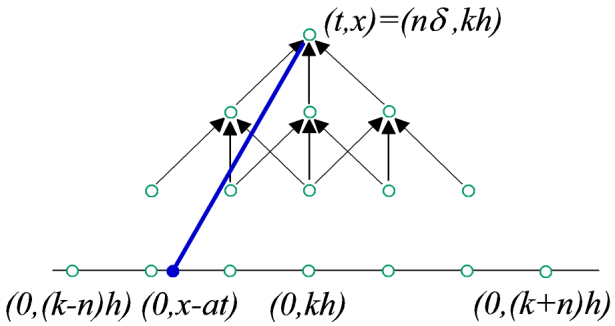
A CFL-feltétel teljesülésének vizsgálatához tehát azt kell ellenőriznünk, hogy $x - at \leq x + t \cdot \frac{a}{R}$ milyen esetben teljesül. Egyszerű számolással - figyelembe véve az $a < 0$ és $R < 0$ relációt - adódik, hogy

$$-at \leq t \cdot \frac{a}{R} \Leftrightarrow -a \leq \frac{a}{R} \Leftrightarrow 1 \leq -\frac{1}{R} \Leftrightarrow R \geq -1,$$

ami pontosan ugyanaz a feltétel, amelyet a 12.13. példában kaptunk. ♢

15.7. Példa. A (14.1) feladatra vonatkozó Lax–Wendroff-séma esetére vizsgáljuk a CFL-feltételt.

Az analitikus függési tartomány megegyezik az előző példában szereplővel, hiszen ez csak a feladattól függ. Az előző példa leírását követve kapjuk, hogy a numerikus függési tartomány (lásd még a 15.2. ábrát is)



15.2. ábra. A (t, x) ponthoz tartozó $H_{\delta,h}$ halmaz a 15.7. példában olyan esetben, amikor a CFL-feltétel teljesül. A vastag kék vonal a karakterisztika $a > 0$ esetén, a kék pont pedig az analitikus függési tartomány.

$$\left[x - t \cdot \frac{a}{R}, x + t \cdot \frac{a}{R}\right].$$

Ekkor a CFL-feltétel azt jelenti, hogy

$$x - t \cdot \frac{a}{R} \leq x - at \leq x + t \cdot \frac{a}{R},$$

ami ekvivalens a

$$-\frac{a}{R} \leq -a \leq \frac{a}{R},$$

egyenlőtlenséggel. Itt $a \leq 0$ esetén a fenti levezetés a $-1 \leq R < 0$ feltételt adja. Hasonlóan $a \geq 0$ esetén az $1 \geq R > 0$ feltételt kapjuk, vagyis összességében annak kell teljesülnie, hogy $|R| \leq 1$. Ez pontosan a 14.4. állításban szereplő feltétel. ♢

15.8. Megjegyzés. A fenti két példában a konvergenciára vonatkozó szükséges és elégséges feltételt kaptunk. ♢

15.9. Példa. Az egydimenziós parabolikus

$$\begin{cases} \partial_t u(t,x) + \sigma_D \partial_{xx} u(t,x) = 0, \quad t \in \mathbb{R}^+, \ x \in \mathbb{R} \\ u(0,x) = g(x), \quad x \in \mathbb{R} \end{cases} \tag{15.2}$$

feladat megoldására felírt

$$u_k^{n+1} = u_k^n + r(u_{k-1}^n - 2u_k^n + u_{k+1}^n)$$

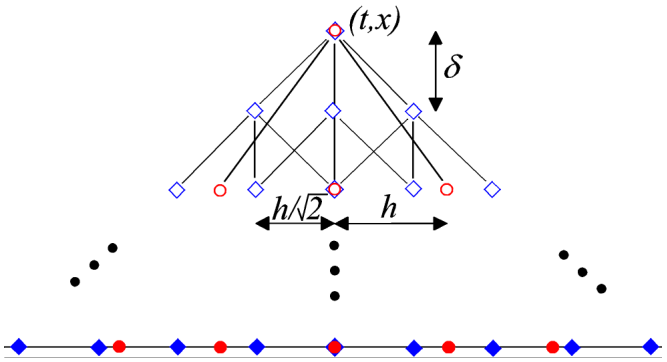
sémát vizsgáljuk, amelyet egy egyenletes felosztású rácson diszkretizálunk.

Tudjuk, hogy a diszkretizációtól függetlenül a (t, x) ponthoz tartozó analitikus függési tartomány $\mathbb{R}$, mert az egyenlet pontos megoldása

$$u(t,x) = \frac{1}{\sqrt{4\pi t}} \int_{\mathbb{R}} e^{-\frac{(x-y)^2}{4t}} g(x-y)\,\mathrm{d}y. \tag{15.3}$$

Ha most egy $\mathbb{R}$-nél szűkebb H halmaz volna az analitikus függési tartomány, akkor az azonosan nulla kezdeti feltételből indított megoldás és az a megoldás, amelyet egy folytonos, H-ban nulla, H^c-ben nemnegatív, de nem nulla kezdeti feltételből indítunk, azonos kellene, hogy legyen, de ez nem állhat fenn, mert az utóbbi a (15.3) formula miatt valahol pozitív lesz.

A numerikus függési tartomány kiszámításához először a $H_{\mathbf{h},\delta}$ halmazt határozzuk meg. Feltesszük, hogy $n\delta = t$, valamint az $r = \frac{\delta}{h^2}$ paraméter rögzített, éppen erre vonatkozó feltételt keresünk.



15.3. ábra. A (t, x) ponthoz tartozó $H_{\delta,h}$ halmaz elemei (teli piros körök) és az ennek meghatározásához szükséges pontok (üres piros körök), valamint a $H_{\delta/2,h/\sqrt{2}}$ halmaz elemei (teli kék rombusz) és az ennek meghatározásához szükséges pontok (üres kék rombusz). Itt $r = \delta/h = (\delta/2)/(h/\sqrt{2})^2$.

A séma szerint u_k^n az előző időlépésben vett u_{k-1}^{n-1}, u_k^{n-1} és u_{k+1}^{n-1} értékektől függ. Ezek pedig az előző időlépésben vett $u_{k-2}^{n-2}, u_{k-1}^{n-2}, u_k^{n-2}, u_{k+1}^{n-2}, u_{k+2}^{n-2}$ értékektől függnek. Ezt folytatva kapjuk, hogy u_k^n a kezdetben adott értékek közül a

$$u_{k-n}^0, u_{k-n+1}^0, u_k^0, \ldots, u_{k+n}^0$$

értékektől függ, ahol az első és az utolsó alappont különbsége $2nh$. Definíció szerint $H_{\mathbf{h},\delta}$ a fenti pontokból áll; lásd a 15.3 ábrát is! Az alappontok távolsága nullához tart a felosztás

finomításával, továbbá

$$nh = \frac{t}{\delta} \cdot h = \frac{t}{r} \cdot \frac{1}{h},$$

ami rögzített r esetén a felosztás fenti finomításával végtelenhez tart. Így a numerikus függési tartomány is $\mathbb{R}$.
Itt sajnos, semmilyen feltételt nem kaptunk az r mennyiségre vonatkozólag. ♢

15.10. Megjegyzés. A fenti levezetés mintájára kapjuk, hogy ha a (15.2) feladat diszkretizációjában $\frac{\delta}{h}$ állandó, akkor a numerikus függési tartomány korlátos. Azaz még még a $\frac{\delta}{h}$ állandóságára vonatkozó feltétellel sem lehet konvergens a séma, tehát feltétel nélküli konvergenciát semmiképpen nem kaphatunk. ♢

16. fejezet

Lineáris PDE rendszerek

Az egydimenziós esetet vizsgáljuk. Azaz legyen az ismeretlen $\mathbf{v}$ függvény $\mathbb{R}\times\mathbb{R}\to\mathbb{R}^l$ típusú, amelyre vonatkozólag a

$$\partial_t\begin{pmatrix}v_1(t,x)\\v_2(t,x)\\\vdots\\v_l(t,x)\end{pmatrix}=A_0\begin{pmatrix}v_1(t,x)\\v_2(t,x)\\\vdots\\v_l(t,x)\end{pmatrix}+A_1\partial_x\begin{pmatrix}v_1(t,x)\\v_2(t,x)\\\vdots\\v_l(t,x)\end{pmatrix}+\cdots+A_J\partial_x^J\begin{pmatrix}v_1(t,x)\\v_2(t,x)\\\vdots\\v_l(t,x)\end{pmatrix}\tag{16.1}$$

PDE rendszer adott, ahol $A_0,A_1,\ldots,A_J$ rögzített mátrixok. Ezt tömören a

$$\partial_t\mathbf{v}=\sum_{j=0}^{J}A_j\partial^j\mathbf{v}$$

alakba írhatjuk, amely még abban az esetben is használható, ha $\mathbf{v}:\mathbb{R}\times\mathbb{R}^n\to\mathbb{R}^l$ alakú; ekkor j a megfelelő parciális deriválthoz tartozó multiindex. A korábbiakhoz hasonlóan használni fogjuk a

$$\mathbf{v}_k^n\approx(v_1,v_2,\ldots,v_l)^T(t_n,x_k),$$

és az ennek segítségével felírt

$$\mathbf{v}^n=(\mathbf{v}_1^n,\mathbf{v}_2^n,\ldots,\mathbf{v}_l^n)^T,$$

jelöléseket. Ezzel a numerikus megoldásra vonatkozó egylépéses véges differencia séma a

$$Q_1\mathbf{v}^{n+1}=Q_2\mathbf{v}^n+\delta\mathbf{F}^n\tag{16.2}$$

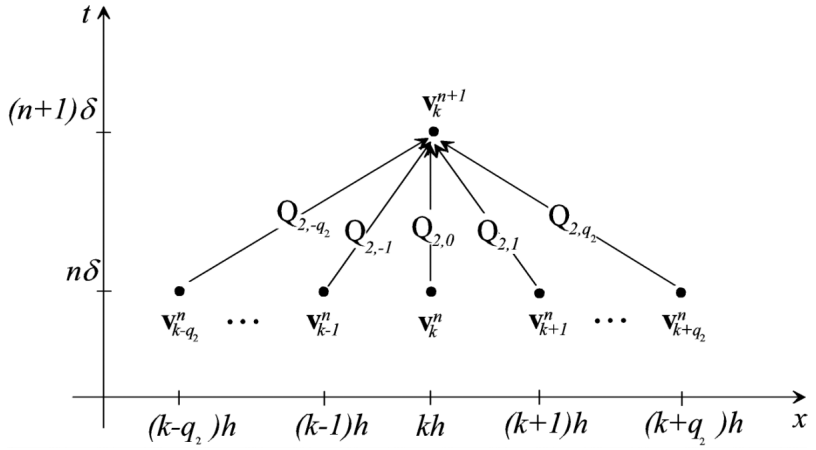
alakba írható, ahol valamilyen $q_1,q_2\in\mathbb{N}$ esetén adott $Q_{1,-q_1},Q_{1,-q_1+1},\ldots,Q_{1,q_1}\in\mathbb{R}^{l\times l}$ és $Q_{1,-q_2},Q_{1,-q_2+1},\ldots,Q_{1,q_2}\in\mathbb{R}^{l\times l}$ mátrixokkal

$$[Q_1\mathbf{v}^{n+1}]_k=Q_{1,-q_1}\mathbf{v}_{k-q_1}^{n+1}+Q_{1,-q_1+1}\mathbf{v}_{k-q_1+1}^{n+1}+\cdots+Q_{1,q_1}\mathbf{v}_{k+q_1}^{n+1},$$

valamint hasonlóan

$$[Q_2\mathbf{v}^n]_k = Q_{2,-q_2}\mathbf{v}^n_{k-q_1} + Q_{2,-q_2+1}\mathbf{v}^n_{k-q_2+1} + \cdots + Q_{2,q_2}\mathbf{v}^n_{k-q_2}.$$

Nyilvánvalóan, ha $q_1 = 0$ és $Q_{1,0} = I$, akkor a séma explicit; ekkor az időlépést a 16.1 ábrán szemléltetjük.



16.1. ábra. Rendszerekre vonatkozó explicit sémák szemléltetése. Az egyes helyekhez tartozó közelítő vektorok az $n\delta$ időpontban és az időlépést megadó mátrixok, melyekből a $\mathbf{v}_k^{n+1}$ vektort meghatározzuk.

A stabilitásvizsgálat alapvető eszköze itt is a diszkrét idejű Fourier-transzformáció lesz. Ezt a vektorokra komponensenként alkalmazzuk, vagyis az

$$\mathcal{F}\mathbf{v}^n(s) := [\mathcal{F}\mathbf{v}_1^n(s), \mathcal{F}\mathbf{v}_2^n(s), \ldots, \mathcal{F}\mathbf{v}_l^n(s)]^T \tag{16.3}$$

egyenlőséggel definiáljuk. A stabilitásvizsgálathoz ismét lényeges lesz ismerni, hogy a fenti Fourier-transzformált hogyan változik lépésről lépésre.

16.1. Definíció. Azt a $\boldsymbol{\rho} : [-\pi, \pi] \to \mathbb{R}^{l\times l}$ mátrixfüggvényt, amellyel

$$\mathcal{F}\mathbf{v}^{n+1}(s) = \boldsymbol{\rho}(s)\mathcal{F}\mathbf{v}^n(s)$$

teljesül, a megfelelő sémához tartozó szorzómátrixnak nevezzük. ♢

Teljes indukcióval egyszerűen látható az is, hogy

$$\mathcal{F}\mathbf{v}^n(s) = \boldsymbol{\rho}^n(s)\mathcal{F}\mathbf{v}^0(s). \tag{16.4}$$

16.2. Megjegyzés. Az itt tárgyalt állítások és a felhasznált módszerek hasonlítanak a lépésmátrixok elemezése során látottakhoz. Fontos különbség azonban, hogy itt a szorzómátrix komplex értékű lehet. ♢

A stabilitásvizsgálathoz szükségünk lesz valamilyen normára is. A $\mathbf{v}^n$ -re vonatkozó „természetes" norma itt mindenképp olyan, ahol $\mathbf{v}^n$ komponenseinek $\|\cdot\|_2$-es normáját használjuk, hiszen a számítások során a végeredmény egyes komponenseinek ilyen normában vett eltérését vizsgáljuk. Ennek megfelelően definiáljuk a

$$\|\!|\mathbf{v}^n|\!\| = \sqrt{\sum_{j=1}^{l} \|\mathbf{v}_j^n\|^2}$$

normát. Ezzel és a (16.3) formulával nyilvánvalóan

$$\begin{aligned}\int_{-\pi}^{\pi} \|\mathcal{F}\mathbf{v}^n(s)\|_2^2 \,\mathrm{d}s = \int_{-\pi}^{\pi} \sum_{j=1}^{l} \left|\mathcal{F}\mathbf{v}_j^n(s)\right|^2 \,\mathrm{d}s = \sum_{j=1}^{l} \int_{-\pi}^{\pi} \left|\mathcal{F}\mathbf{v}_j^n(s)\right|^2 \,\mathrm{d}s = \\ = \sum_{j=1}^{l} \left|\mathbf{v}_j^n\right|_2^2 = |\!|\!|\mathbf{v}^n|\!|\!|^2.\end{aligned} \tag{16.5}$$

Az egyes lépésekben vett $|\!|\!|\cdot|\!|\!|$ normák változását becsülhetjük a következő lemma segítségével.

16.3. Lemma. *A fenti (16.2) típusú séma homogén verziójához tartozó $\boldsymbol{\rho}(s)$ szorzómátrixra és az n-edik lépésben kapott $\mathbf{v}^n$ közelítés normájára teljesül a következő becslés:*

$$|\!|\!|\mathbf{v}^n|\!|\!| \le \max_{s\in[-\pi,\pi]} \|\boldsymbol{\rho}(s)\|_2 \cdot |\!|\!|\mathbf{v}^0|\!|\!|.$$

Bizonyítás. A (16.5) formula, valamint a (16.4) egyenlőség felhasználásával kapjuk, hogy

$$\begin{aligned}|\!|\!|\mathbf{v}^n|\!|\!|^2 &= \int_{-\pi}^{\pi} \|\mathcal{F}\mathbf{v}^n(s)\|_2^2 \,\mathrm{d}s = \int_{-\pi}^{\pi} \|\boldsymbol{\rho}^n(s)\mathcal{F}\mathbf{v}^0(s)\|_2^2 \,\mathrm{d}s \le \\ &\le \int_{-\pi}^{\pi} \max_{s\in[-\pi,\pi]} \|\boldsymbol{\rho}^n(s)\|_2^2 \|\mathcal{F}\mathbf{v}^0(s)\|_2^2 \,\mathrm{d}s = \max_{s\in[-\pi,\pi]} \|\boldsymbol{\rho}^n(s)\|_2^2 \int_{-\pi}^{\pi} \|\mathcal{F}\mathbf{v}^0(s)\|_2^2 \,\mathrm{d}s = \\ &= \max_{s\in[-\pi,\pi]} \|\boldsymbol{\rho}^n(s)\|_2^2 \cdot |\!|\!|\mathbf{v}^0|\!|\!|^2,\end{aligned} \tag{16.6}$$

ahogy azt a lemmában állítottuk. □

A 12.8. lemma bizonyításával analóg módon kapjuk, hogy teljesül a következő állítás.

16.4. Állítás. *A (16.2) séma homogén verziója pontosan akkor exponenciálisan stabil a $\|\!|\cdot|\!\|$ normára nézve, ha van olyan $C \in \mathbb{R}$, $\mathbf{h}_0$ és δ_0, hogy minden $\mathbf{h} \leq \mathbf{h}_0$ és $\delta \leq \delta_0$ esetén minden $s \in [-\pi, \pi]$-re teljesül a következő egyenlőtlenség:*

$$\|\boldsymbol{\rho}(s)\|_2 \leq 1 + C\delta.$$

A fenti állítás feltételének azt a gyengítését, ahol a szorzómátrix $\|\cdot\|_2$ normája helyett annak spektrálsugara szerepel, (ebben az esetben is) Neumann-feltételnek nevezzük.

16.5. Állítás. *A Neumann-feltétel teljesülése szükséges a stabilitáshoz.*

Bizonyítás. Az $\sigma(\boldsymbol{\rho}(s)) \leq \|\boldsymbol{\rho}(s)\|_2$ egyenlőtlenség és az előző állítás miatt ha a séma stabil, akkor van olyan $C \in \mathbb{R}$, $\mathbf{h}_0$ és δ_0, hogy minden $\mathbf{h} \leq \mathbf{h}_0$ és $\delta \leq \delta_0$ esetén minden $s \in [-\pi, \pi]$-re teljesül a következő egyenlőtlenség:

$$\sigma(\boldsymbol{\rho}(s)) \leq \|\boldsymbol{\rho}(s)\|_2 \leq 1 + C\delta,$$

azaz teljesül a Neumann-feltétel. □

A Neumann-feltétel azonban nem elégséges a megfelelő séma stabilitásához, amint a következő példában rámutatunk.

16.6. Példa. Vizsgáljuk meg a

$$\begin{cases} \partial_t v_1(t,x) = \sigma_D \partial_{xx} v_2(t,x) + f_1(t,x) \\ \partial_t v_2(t,x) = f_2(t,x) \end{cases}$$

probléma homogén verziójához tartozó

$$\begin{cases} v_{k,1}^{n+1} = v_{k,1}^n + r(v_{k-1,2}^n - 2v_{k,2}^n + v_{k+1,2}^n) \\ v_{k,2}^{n+1} = v_{k,2}^n \end{cases} \tag{16.7}$$

séma stabilitását!

Ez a (16.2) alakba írható, ahol a bal oldalon

$$q_1 = 0, \quad Q_{1,0} = I,$$

a jobb oldalon pedig

$$q_2 = 1, \quad Q_{2,-1} = Q_{2,1} = \begin{pmatrix} 0 & r \\ 0 & 0 \end{pmatrix} \quad \text{és} \quad Q_{2,0} = \begin{pmatrix} 1 & -2r \\ 0 & 1 \end{pmatrix}.$$

Ekkor valóban

$$I \begin{pmatrix} v_{k,1}^{n+1} \\ v_{k,2}^{n+1} \end{pmatrix} = Q_{2,-1} \begin{pmatrix} v_{k-1,1}^{n} \\ v_{k-1,2}^{n+1} \end{pmatrix} + Q_{2,0} \begin{pmatrix} v_{k,1}^{n} \\ v_{k,2}^{n+1} \end{pmatrix} + Q_{2,1} \begin{pmatrix} v_{k+1,1}^{n} \\ v_{k+1,2}^{n+1} \end{pmatrix}.$$

A (16.7) formulában szereplő egyenlőségek mindkét oldalának diszkrét idejű Fourier-transzformáltját véve kapjuk, hogy

$$\begin{cases} \mathcal{F}\mathbf{v}_1^{n+1}(s) = \mathcal{F}\mathbf{v}_1^n(s) + \mathcal{F}\mathbf{v}_2^n(s) \cdot -4r\sin^2\frac{s}{2} \\ \mathcal{F}\mathbf{v}_2^{n+1}(s) = \mathcal{F}\mathbf{v}_2^n(s) \end{cases}$$

azaz

$$\begin{pmatrix} \mathcal{F}\mathbf{v}_1^{n+1}(s) & \mathcal{F}\mathbf{v}_1^{n+1}(s) \end{pmatrix}^T = \begin{pmatrix} 1 & -4r\sin^2\frac{s}{2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathcal{F}\mathbf{v}_1^n(s) & \mathcal{F}\mathbf{v}_1^n(s) \end{pmatrix}^T,$$

tehát

$$\boldsymbol{\rho}(s) = \begin{pmatrix} 1 & -4r\sin^2\frac{s}{2} \\ 0 & 1 \end{pmatrix}.$$

A Neumann-feltétel teljesül, hiszen ennek a mátrixnak mindkét sajátértéke 1. Ugyanakkor teljes indukcióval egyszerűen igazolható, hogy

$$\boldsymbol{\rho}^n(s) = \begin{pmatrix} 1 & -4nr\sin^2\frac{s}{2} \\ 0 & 1 \end{pmatrix},$$

vagyis mivel a 2×2-es mátrixok normái ekvivalensek, valamilyen $C \in \mathbb{R}^+$ konstanssal

$$\|\boldsymbol{\rho}^n(\pi)\|_2 \geq C \cdot 4n.$$

Ekkor a (16.4) formula felhasználásával, valamint a 12.7. lemma bizonyítása során alkalmazott gondolatmenet felhasználásával kapjuk, hogy alkalmas $\mathbf{v}^0$ vektorra

$$\|\!|v^n|\!\| \geq C \cdot 3n \cdot \|\!|v^0|\!\|,$$

ami a megfelelő módszer instabilitását jelenti. ♢

A továbbiakban olyan feltételeket adunk meg, amelyek a Neumann-feltétel teljesülése mellett biztosítják a stabilitást.

16.7. Állítás. *Tegyük fel, hogy a (16.2) séma homogén verziójához tartozó $\boldsymbol{\rho}$ szorzómátrixra teljesül a Neumann-feltétel. Ha emellett az alábbi feltételek valamelyike teljesül, akkor a (16.2) séma homogén verziója stabil.*

(a) $\boldsymbol{\rho}(s)$ minden s-re Hermite-féle (valós mátrixokra ez azt jelenti, hogy szimmetrikus).

(b) Létezik olyan s-től folytonosan függő $H(s) \in \mathbb{R}^{l\times l}$ mátrixcsalád, amelyre

$$H^{-1}(s)\boldsymbol{\rho}(s)H(s)$$

Hermite-féle.

(c) *Létezik olyan s-től folytonosan függő $H(s) \in \mathbb{R}^{l\times l}$ mátrixcsalád, $\delta_0 \in \mathbb{R}$ és $C \in \mathbb{R}^+$, amelyre minden $s \in [-\pi, \pi]$ és $\delta < \delta_0$ esetén*

$$H^{-1}(s)\boldsymbol{\rho}(s)H(s) \cdot [H^{-1}(s)\boldsymbol{\rho}(s)H(s)]^* \leq (1 + C\delta^2)I$$

abban az értelemben, hogy a jobb és a bal oldal különbsége pozitív szemidefinit.

Bizonyítás. A bizonyítások előtt megjegyezzük, hogy H-nak az s-től való folytonos függése miatt H-nak akármelyik normája folytonosan függ s-től, és mivel $s \in [0, 2\pi]$, ezért a $\|H(s)\|_2$-nak van maximuma. Ugyanez teljesül H^{-1}-re is.

Az (a) pontban szereplő állítás bizonyításához megjegyezzük, hogy Hermite-féle mátrixokra is teljesül a $\sigma(\boldsymbol{\rho}(s)) = \|\boldsymbol{\rho}(s)\|_2$ egyenlőség, vagyis a 16.4. állítás segítségével kapjuk, hogy az ehhez tartozó séma stabil.

A (b) pontban szereplő állítás igazolásához először megjegyezzük, hogy a fenti feltétel olyan alakra írható, hogy a fenti, s-től folytonosan függő $H(s) \in \mathbb{R}^{l\times l}$ mátrixcsalád esetén

$$\boldsymbol{\rho}(s) = H(s)[H^{-1}(s)\boldsymbol{\rho}(s)H(s)]H^{-1}(s),$$

ahol a $H^{-1}(s)\boldsymbol{\rho}(s)H(s)$ mátrix Hermite-féle. Ekkor természetesen

$$\boldsymbol{\rho}^n(s) = H(s)[H^{-1}(s)\boldsymbol{\rho}(s)H(s)]^n H^{-1}(s), \tag{16.8}$$

vagyis

$$\begin{aligned}
\|\boldsymbol{\rho}^n(s)\|_2 &\leq \|H(s)\|_2\|H^{-1}(s)\|_2\|[H^{-1}(s)\boldsymbol{\rho}(s)H(s)]^n\|_2 = \\
&= \|H(s)\|_2\|H^{-1}(s)\|_2 s([H^{-1}(s)\boldsymbol{\rho}(s)H(s)]^n) = \\
&= \|H(s)\|_2\|H^{-1}(s)\|_2 (s([H^{-1}(s)\boldsymbol{\rho}(s)H(s)])^n \leq \\
&\leq \|H(s)\|_2\|H^{-1}(s)\|_2 s(\boldsymbol{\rho}(s))^n = \|H(s)\|_2\|H^{-1}(s)\|_2(1 + C\delta)^n,
\end{aligned}$$

amiből $\|H(s)\|_2\|H^{-1}(s)\|_2$ korlátossága, valamint a $\delta \leq \frac{T}{n}$ egyenlőtlenség alapján az állítás következik.

A (c) pontban szereplő állítás bizonyításához felhasználjuk, hogy a teljesül a (12.14) azonoság. Emiatt, a 12.19. állítás, valamint a (16.8) formula felhasználásával kapjuk, hogy

$$\begin{aligned}
\|\boldsymbol{\rho}^n(s)\|_2^2 &= \sigma(H(s)[H^{-1}(s)\boldsymbol{\rho}(s)H(s)]^n H^{-1}(s)(H^{-1})^*(s)([H^{-1}(s)\boldsymbol{\rho}(s)H(s)]^n)^* H^*(s)) \leq \\
&\leq \sigma^2(H(s))\sigma^2(H^{-1}(s))\sigma([H^{-1}(s)\boldsymbol{\rho}(s)H(s)]^n)\sigma([[H^{-1}(s)\boldsymbol{\rho}(s)H(s)]^n]^*) = \\
&= \sigma^2(H(s))\sigma^2(H^{-1}(s))\sigma(\boldsymbol{\rho}^n(s))\sigma([\boldsymbol{\rho}^n]^*(s)) = \\
&= [\sigma(H(s))\sigma(H^{-1}(s))]^2[\sigma(\boldsymbol{\rho}^n(s))]^2 = [\sigma(H(s))\sigma(H^{-1}(s))]^2(1 + C\delta)^{2n},
\end{aligned}$$

amiből ismét $\sigma(H(s))\sigma(H^{-1}(s))$ korlátossága, valamint a $\delta \leq \frac{T}{n}$ egyenlőtlenség alapján kapjuk az állítást. □

A következőkben általánosan is igazoljuk, hogy explicit alakban megadott sémák stabilitásvizsgálata egyszerűsíthető úgy, hogy a sémából a nulladrendű tagnak megfelelő komponenst elhagyjuk.

16.8. Állítás. *Tegyük fel, hogy az* $\mathbf{u}^{n+1} = Q\mathbf{u}^n$ *séma stabil. Ekkor az* $\mathbf{u}^{n+1} = Q\mathbf{u}^n + \delta B\mathbf{u}^n$ *séma is stabil.*

16.9. Megjegyzés.

1. Ugyanez igaz, ha B függ t-től vagy x-től, de az a normája, amihez tartozó konvergenciát vizsgálunk, korlátos marad.

2. Az eredeti feladatban ez azt jelenti, hogy elegendő a nulladrendű tag elhagyásával kapott séma stabilitását ellenőrizni.

3. A fenti alakban nem kötöttük ki, hogy milyen feladatról van szó; az állításban csak azt használjuk ki, hogy az ott szereplő operátorok lineárisak. Ezért akár rendszerre, akár többdimenziós feladatra is alkalmazható az eredmény. ♢

Bizonyítás. A fenti séma az $\mathbf{u}^{n+1} = (Q + \delta B)\mathbf{u}^n$ alakba is írható, vagyis teljes indukcióval azt kapjuk, hogy $\mathbf{u}^n = (Q + \delta B)^n\mathbf{u}^0$. Nyilván elegendő azt igazolni, hogy $\|Q^n\|$ korlátossága esetén $\|(Q+\delta B)^n\|$ is korlátos. A Q-val adott séma stabilitása azt is jelenti, hogy minden n-re, amelyre $n\delta < T$, teljesül, hogy $\|Q^n\| < K$; azaz akkor is, ha a kitevő $0, 1, \ldots, n-1$. Emiatt a K korlátról feltesszük, hogy legalább 1.

A $(Q + \delta B)^n$ operátort fogjuk kifejteni, ami azért nem triviális, mert Q és B nem feltétlenül felcserélhetők.

- Lesz egy Q^n tag, amelyre $\|Q^n\| \le K \le K^n$.

- Lesznek egyetlen δB-t tartalmazó

  $$\delta BQ^n, Q\delta BQ^{n-1}, \ldots, Q^{n-1}\delta BQ, Q^n\delta B$$

  tagok is, amelyek száma $\binom{n}{1}$, összegének normája felülről becsülhető a

  $$\binom{n}{1}\|\delta B\|K^2$$

  mennyiséggel.

- Lesz hasonlóan minden $j = 2, 3, \ldots, N$-re a j db B-t tartalmazó száma $\binom{n}{j}$, összegüknek normája felülről becsülhető a

  $$\binom{n}{j}\|\delta B\|^j K^{j+1}$$

  mennyiséggel.

Összeadva a fenti normákat az alábbi becslést kapjuk:

$$\|(Q+\delta B)^n\| \le \sum_{j=1}^{n} \binom{n}{j} \|\delta B\|^j K^{j+1} \le K \sum_{j=1}^{n} \frac{n^j}{j!} \delta^j \|B\|^j K^j \le$$
$$\le K \sum_{j=1}^{\infty} \frac{n^j}{j!} \delta^j \|B\|^j K^j = K e^{\delta n K \|B\|} \le K e^{TK\|B\|},$$

amivel a fentiek értelmében a stabilitást igazoltuk. □

16.10. Példa. Vizsgáljuk meg a

$$\partial_t \mathbf{v}(t,x) = B\partial_{xx}\mathbf{v}(t,x) + B_0\mathbf{v}(t,x), \quad t \in [0,T],\ x \in \mathbb{R}$$

parabolikus rendszerhez tartozó

$$\mathbf{v}^{n+1} = \mathbf{v}^n + r \cdot BD_0^2\mathbf{v}^n + \delta B_0\mathbf{v}^n$$

séma stabilitását, ahol $B, B_0 \in \mathbb{R}^{l\times l}$ adott mátrixok, valamint B (szimmetrikus) pozitív definit!

A 16.8. állítás alapján ez ekvivalens a

$$\mathbf{v}^{n+1} = \mathbf{v}^n + r \cdot BD_0^2\mathbf{v}^n$$

séma stabilitásával. Itt mindkét oldal Fourier-transzformáltját véve kapjuk, hogy

$$\mathcal{F}\mathbf{v}^{n+1}(s) = \mathcal{F}\mathbf{v}^n(s) - 4r\sin^2\frac{s}{2} \cdot B\mathcal{F}\mathbf{v}^n(s) = (I - 4r\sin^2\frac{s}{2} \cdot B)\mathcal{F}\mathbf{v}^n(s),$$

azaz

$$\boldsymbol{\rho}(s) = I - 4r\sin^2\frac{s}{2} \cdot B.$$

Mivel a feltevés szerint ez a mátrix szimmetrikus, elegendő azt ellenőrizni, hogy mikor teljesül a Neumann-feltétel. Tudjuk azt is, hogy $\boldsymbol{\rho}(s)$ minden sajátértéke $1 - 4r\sin^2\frac{s}{2}\lambda_j$ alakú, ahol λ_j a B mátrix egy sajátértéke. Ennek abszolútértéke akkor maximális, ha $\lambda_j = \lambda_{\max}$ a maximális sajátérték, vagyis a skalár eset mintájára az $r\lambda_{\max} \le \frac{1}{2}$ (szükséges és elégséges) feltétel adódik. ♢

16.11. Példa. Vizsgáljuk meg a

$$\partial_t \mathbf{v}(t,x) = A\partial_x\mathbf{v}(t,x) + A_0\mathbf{v}(t,x), \quad t \in [0,T],\ x \in \mathbb{R} \tag{16.9}$$

hiperbolikus rendszerhez tartozó

$$\mathbf{v}^{n+1} = \mathbf{v}^n + R \cdot AD_+\mathbf{v}^n + \delta A_0\mathbf{v}^n$$

séma stabilitását, ahol $A, A_0 \in \mathbb{R}^{l\times l}$ adott mátrixok, valamint A diagonalizálható valós sajátértékekkel!

A 16.8. állítás alapján ez ekvivalens a

$$\mathbf{v}^{n+1} = \mathbf{v}^n + R \cdot AD_+\mathbf{v}^n$$

séma stabilitásával. Itt mindkét oldal Fourier-transzformáltját véve kapjuk, hogy

$$\mathcal{F}\mathbf{v}^{n+1}(s) = \mathcal{F}\mathbf{v}^n(s) + R(e^{is}-1)\cdot A\mathcal{F}\mathbf{v}^n(s) = (I + R(e^{is}-1)\cdot A)\mathcal{F}\mathbf{v}^n(s),$$

azaz

$$\boldsymbol{\rho}(s) = (I + (R\cos s - R)A) + i\cdot \sin s \cdot A.$$

Ez nem feltétlenül Hermite-féle, hiszen a főátlójában bármilyen komplex szám állhat.

Jelölje $H \in \mathbb{R}^{l\times l}$ azt a mátrixot, amellyel $H^{-1}AH = D$ diagonális, valós komponensekkel. Ekkor

$$H^{-1}\boldsymbol{\rho}(s)H = (I + (R\cos s - R)D) + i\cdot \sin s \cdot D.$$

Ezért a 16.7. állítás (b) pontja szerint most is elegendő azt ellenőrizni, hogy mikor teljesül a Neumann-feltétel.

Tudjuk azt is, hogy $\boldsymbol{\rho}(s)$ minden sajátértéke

$$1 + (R\cos s - R)\lambda_j + i\cdot \sin s \cdot \lambda_j$$

alakú, ahol λ_j az A mátrix egy sajátértéke. A 12.13. példában követett levezetés alapján kapjuk, hogy a Neumann-feltétel pontosan akkor teljesül, ha

$$0 < R\lambda_j \le 1$$

teljesül minden λ_j sajátértékre. ♢

16.12. Megjegyzés. Hasonló elvet alkalmazva kapjuk, hogy a fenti (16.9) hiperbolikus rendszerre vonatkozó, a Lax–Wendroff-sémának megfelelő

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \frac{R}{2}\cdot AD_0\mathbf{u}^n + \frac{R^2}{2}A^2D_0^2\mathbf{u}^n + \delta A_0\mathbf{u}^n$$

séma pontosan akkor stabil, ha

$$|R\lambda_j| \le 1$$

teljesül A minden λ_j sajátértékére. ♢

Ezzel kapcsolatban a 21.28. feladatra utalunk. Egy további érdekes példa található a 21.29. feladatban.

17. fejezet

Többlépéses sémák

Az eddigiekben az időlépés mindig abból állt, hogy az $\mathbf{u}^n$ segítségével (esetleg valamilyen implicit módszerrel) $\mathbf{u}^{n+1}$-et előállítottuk. Ezt fogjuk általánosítani úgy, hogy $\mathbf{u}^{n+1}$ kiszámításához az $\mathbf{u}^n, \mathbf{u}^{n-1}, \dots, \mathbf{u}^{n-s}$ vektorokat is felhasználjuk. Egy egyrészt nyilván az időváltozó szerinti magasabb konzisztenciarendet eredményez, másrészt szükséges is, ha az idő szerinti magasabb rendű deriváltat kell közelíteni. A fenti eljárás természetes módon vezet rendszerek alakjában felírt sémákhoz, amelyeknek stabilitását most is a vektor értékű diszkrét idejű Fourier-transzformáció segítségével vizsgálhatjuk. Alkalmazhatjuk ugyanígy a stabilitásvizsgálat egyszerűsítésére szolgáló 16.8. állítást is, vagyis az ehhez a fejezethez tartozó elmélet fontosabb elemei az előző fejezetben találhatók. Ennek megfelelően itt konkrét példák ismertetésére szorítkozunk.

17.1. Példa. Vizsgáljuk meg a

$$\partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x)$$

egyenlethez tartozó

$$u_k^{n+1} = u_k^{n-1} + 2r(u_{k-1}^n - 2u_k^n + u_{k+1}^n) \tag{17.1}$$

séma konvergenciáját!

Tudjuk, hogy

$$\frac{1}{2\delta}(u((n+1)\delta, kh) - u((n-1)\delta, kh)) = \partial_t u(n\delta, kh) + \mathcal{O}(\delta^2),$$

továbbá

$$\frac{1}{h^2}(u(n\delta, (k-1)h) - 2u(n\delta, kh) + u(n\delta, (k+1)h) = \partial_{xx} u(n\delta, kh) + \mathcal{O}(h^2),$$

amelyeket egyenlővé téve ugyanazt kapjuk, mintha a (17.1) sémába az eredeti megoldást helyettesítettük volna, majd 2δ-val osztottunk volna. Ez a 11.9. állításnak megfelelően azt jelenti, hogy a (17.1) séma mindkét változójában másodrendben konzisztens.

A stabilitáshoz a sémát rendszer alakjába írva kapjuk az

$$\begin{pmatrix}\mathbf{u}^{n+1}\\ \mathbf{u}^{n}\end{pmatrix} = \begin{pmatrix}\mathbf{u}^{n-1}+2rD_0^2\mathbf{u}^{n}\\ \mathbf{u}^{n}\end{pmatrix} = \begin{pmatrix}2rD_0^2 & I\\ I & 0\end{pmatrix}\begin{pmatrix}\mathbf{u}^{n}\\ \mathbf{u}^{n-1}\end{pmatrix} \tag{17.2}$$

összefüggést, vagyis az egyes időpontokban vett ismeretlenek Fourier-transzformáltjaira teljesül a

$$\begin{pmatrix}\mathcal{F}\mathbf{u}^{n+1}(s)\\ \mathcal{F}\mathbf{u}^{n}(s)\end{pmatrix} = \begin{pmatrix}\mathcal{F}\mathbf{u}^{n-1}(s)+4r(\cos s-1)\mathcal{F}\mathbf{u}^{n}(s)\\ \mathcal{F}\mathbf{u}^{n}(s)\end{pmatrix} = \begin{pmatrix}4r(\cos s-1) & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}\mathcal{F}\mathbf{u}^{n}(s)\\ \mathcal{F}\mathbf{u}^{n-1}(s)\end{pmatrix}$$

összefüggés, ahol a stabilitás megállapításához a

$$\boldsymbol{\rho}(s) = \begin{pmatrix}4r(\cos s-1) & 1\\ 1 & 0\end{pmatrix} \tag{17.3}$$

mátrixot vizsgáljuk.

Ez valós és szimmetrikus, vagyis elegendő a Neumann-feltételt ellenőrizni. $\boldsymbol{\rho}(s)$ mátrix sajátértékeinek szorzata 1, összegük pedig $4r(\cos s-1)$. Emiatt valamilyen s-re fennáll, hogy a két sajátértékek nem $(1,1)$ és nem is $(-1,-1)$. Így az egyik abszolútértéke 1-nél nagyobb, vagyis a séma mindenképpen instabil. Emiatt konvergens sem lehet. ♢

17.2. Megjegyzés. A számolást gyakran „gépiesen” egyszerűsítik; azt mondják, hogy a (17.2) egyenlőségben szereplő mátrix Fourier-transzformáltját elemenként kiszámítva nyerjük a (17.2)-beli szorzómátrixot. Ez a megfogalmazás precízzé tehető, a megfelelő számolási eljárás emiatt helyes. ♢

Hasonló módosítást vizsgálunk egy egyszerű advekciós feladat esetére is.

17.3. Példa. Vizsgáljuk meg a

$$\partial_t u(t,x) = a\partial_x u(t,x)$$

egyenlethez tartozó

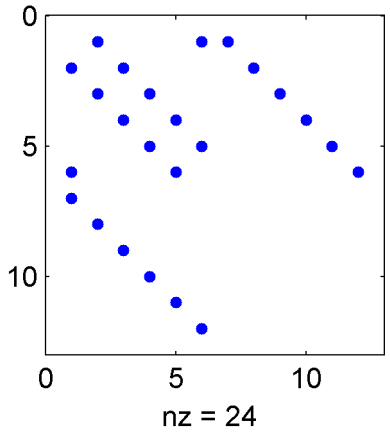
$$u_k^{n+1} = u_k^{n-1} + R(u_{k+1}^n - u_{k-1}^n) \tag{17.4}$$

séma konvergenciáját! Ezt gyakran leapfrog sémának nevezik.

Az előző példában szereplő gondolatmenet módosításával kapjuk, hogy a (17.4) séma mindkét változójában másodrendben konzisztens.

A stabilitáshoz a sémát rendszer alakjába írva kapjuk az

$$\begin{pmatrix}\mathbf{u}^{n+1}\\ \mathbf{u}^{n}\end{pmatrix} = \begin{pmatrix}\mathbf{u}^{n-1}+RD_0\mathbf{u}^{n}\\ \mathbf{u}^{n}\end{pmatrix} = \begin{pmatrix}RD_0^2 & I\\ I & 0\end{pmatrix}\begin{pmatrix}\mathbf{u}^{n}\\ \mathbf{u}^{n-1}\end{pmatrix}$$



17.1. ábra. Nemnulla elemek a (17.4) sémához tartozó lépésmátrixban, ha a feladathoz periodikus peremfeltételek tartoznak és az intervallumot hat egyforma hosszú részre osztottuk.

összefüggést. A lépésmátrix nemnulla elemeit mutatja a 17.1. ábra abban az esetben, ha a (17.4) egyenlethez tartozó feladat egy intervallumon adott periodikus peremfeltétellel. Az egyes időpontokban vett ismeretlenek Fourier-transzformáltjaira teljesül tehát az

$$\begin{pmatrix} \mathcal{F}\mathbf{u}^{n+1}(s) \\ \mathcal{F}\mathbf{u}^{n}(s) \end{pmatrix} = \begin{pmatrix} \mathcal{F}\mathbf{u}^{n-1}(s) + 2R\sin s \cdot i \cdot \mathcal{F}\mathbf{u}^{n}(s) \\ \mathcal{F}\mathbf{u}^{n}(s) \end{pmatrix} = \begin{pmatrix} 2R\sin s \cdot i & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \mathcal{F}\mathbf{u}^{n}(s) \\ \mathcal{F}\mathbf{u}^{n-1}(s) \end{pmatrix}$$

összefüggés, ahol a stabilitás megállapításához a

$$\boldsymbol{\rho}(s) = \begin{pmatrix} 2R\sin s \cdot i & 1 \\ 1 & 0 \end{pmatrix}$$

mátrixot vizsgáljuk.

Ez nem Hermite-féle, de szükségessége miatt először a Neumann-feltételt ellenőrizzük. Tudjuk, hogy $\boldsymbol{\rho}(s)$ mátrix sajátértékeinek szorzata 1, összegük pedig $2R\sin s \cdot i$.

Ha itt $R > 1$, akkor $s = \frac{\pi}{2}$ esetén a sajátértékek összege $2Ri$, amelyre $|2Ri| > 2$, vagyis valamelyik sajátérték abszolútértéke is 1-nél nagyobb.

Ha $R = 1$, akkor $s = \frac{\pi}{2}$ esetén a sajátértékek összege $2i$, vagyis csak akkor állhatna fenn stabilitás, ha mindkét sajátérték i volna. De ekkor a mátrix Jordan-féle normálalakja

$$\begin{pmatrix} i & 1 \\ 0 & i \end{pmatrix}$$

lenne, ami ismét instabilitáshoz vezet a 16.6. példa esetéhez hasonlóan.

Ha $R < 1$, akkor egyszerű számolással kapjuk, hogy a két sajátérték

$$\lambda_{1,2} = R \sin s \cdot i \pm \sqrt{1 - R^2 \sin^2 s},$$

ahol a második tag valós. Emiatt

$$|\lambda_{1,2}| = R^2 \sin^2 s + 1 - R^2 \sin^2 s = 1,$$

tehát a sajátértékek abszolútértéke 1, és azok különbözőek. Így az a mátrix, amellyel $\boldsymbol{\rho}(s)$ Jordan-féle normálalakra hozható, megfelel a 16.7. állítás (c) pontbeli követelményeinek; $R < 1$ esetében tehát a séma stabil. ♢

Ehhez kapcsolódik a 20.2.15. animáció. A többlépéses módszerekre vonatkozó részletes leírást találunk a [28] és a [29] könyvekben.

18. fejezet

Stabilitásvizsgálat másodrendű feladatokra

Motivációként a lehető legegyszerűbb másodrendű feladat megoldásának egy természetes közelítését adjuk meg, és igazoljuk, hogy az eredeti értelemben (vagyis a 11.19. tétel szerint) véve ez a séma sohasem stabil.

Vizsgáljuk tehát a $\partial_{tt}u(t,x) = \partial_{xx}u(t,x)$ PDE valamilyen megoldását közelítő

$$\frac{1}{\delta^2}(u_k^{n-1} - 2u_k^n + u_k^{n+1}) = \frac{1}{h_x^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n) \tag{18.1}$$

sémát. Ebből az új időlépésben kapott u_k^{n+1} változót kifejezve

$$u_k^{n+1} = 2u_k^n - u_k^{n-1} + \frac{\delta^2}{h_x^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n). \tag{18.2}$$

Ezt rendszer alakjába írva kapjuk az

$$\begin{pmatrix} \mathbf{u}^{n+1} \\ \mathbf{u}^n \end{pmatrix} = \begin{pmatrix} R^2 D_{0,x}^2 + I & -I \\ I & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}^n \\ \mathbf{u}^{n-1} \end{pmatrix}$$

összefüggést, vagyis az egyes időpontokban vett ismeretlenek Fourier-transzformáltjaira teljesül a

$$\mathcal{F}\begin{pmatrix} \mathbf{u}^{n+1} \\ \mathbf{u}^n \end{pmatrix}(s) = \begin{pmatrix} 2R^2(\cos s - 1) + 2 & -1 \\ 1 & 0 \end{pmatrix} \mathcal{F}\begin{pmatrix} \mathbf{u}^n \\ \mathbf{u}^{n-1} \end{pmatrix}(s)$$

összefüggés, ahol a stabilitás megállapításához a

$$\boldsymbol{\rho}(s) = \begin{pmatrix} 2R^2(\cos s - 1) + 2 & -1 \\ 1 & 0 \end{pmatrix}$$

mátrixot vizsgáljuk.

18.1. Állítás. *A fenti $\boldsymbol{\rho}$ mátrix n-edik hatványa semmilyen R érték mellett sem korlátos (itt, ahogy az előzőekben is, $n\delta \leq t$), azaz a 11.15. definíció értelmében a (18.1) séma nem lehet stabil.*

Bizonyítás. Vegyük észre, hogy $\alpha = 0$ esetén a megfelelő mátrix bal felső eleme 2, és ekkor egyszerű számolás mutatja, hogy kétszeres sajátértéke az 1. Mivel ez nem hasonló az egységmátrixhoz (mert ahhoz csak önmaga hasonló), ezért Jordan-féle normálalakja egy 2×2-es blokkból áll, a következő módon állítható elő:

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = H^{-1}\boldsymbol{\rho}(s)H$$

Emiatt

$$H^{-1}\boldsymbol{\rho}^n(s)H = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix},$$

vagyis $\|H^{-1}\|_2\|\boldsymbol{\rho}^n(s)\|_2\|H\|_2 \geq n$, amiből az következik, hogy $\|\boldsymbol{\rho}^n(s)\|_2$ valóban nem lehet korlátos. □

Ez az eredmény furcsa, mert a módszert programozva több paraméter esetén is konvergenciát kapunk. Ezért részletesebben megvizsgáljuk a stabilitás definícióját. Valójában arra vagyunk kíváncsiak, mit kell a transzformációs mátrixnak (akár az eredetinek, akár az egyes helyeken vett Fourier-transzformáltak közöttinek) teljesíteni ahhoz, hogy a megfelelő módszer a valódi megoldáshoz tartson.

Fontos észrevétel, hogy a fenti mátrix exponenciális függvénye is csak lineárisan nő. Ennek nyomán adjuk a következő definíciót.

18.2. Definíció. Azt mondjuk, hogy az $u^{n+1} = Qu^n$ séma szublineárisan instabil, ha létezik olyan K konstans, hogy minden $\mathbf{h} > 0$ és $\delta > 0$ mellett $Q^n \leq Kn$.
Hasonlóan, azt mondjuk, hogy a séma a $H(\mathbf{h}, \delta)$ feltétel mellett szublineárisan instabil, ha a fenti egyenlőtlenség $H(\mathbf{h}, \delta) > 0$ esetén teljesül. ♢

Ezen fogalom segítségével igazoljuk a Lax-ekvivalenciatétel egy változatát. Az eredeti levezetéshez képest azzal az egyszerűsítő feltevéssel élünk, hogy a lépésoperátor minden lépésben ugyanaz (ahogy a 18.2. definícióban is, ezt Q-val jelöljük), továbbá a hiba nagyságrendje is, amire emiatt szintén egységesen az $\mathcal{R}(\mathbf{h}, \delta)$ jelölést használjuk.

18.3. Tétel. *(módosított Lax-ekvivalenciatétel) Legyen a $\mathbf{v}^{n+1} = Q\mathbf{v}^n$ séma olyan, hogy abba a pontos megoldást helyettesítve a hiba $\delta^2\mathcal{R}(\mathbf{h}, \delta)$ alakú, továbbá a kezdeti közelítés hibájára*

$$\|\mathbf{e}^0\|_{\mathbf{h},p} = \|\mathbf{v}(0, \cdot) - \mathbf{v}^0\|_{\mathbf{h},p} = \delta\tilde{\mathcal{R}}(\mathbf{h}, \delta),$$

ahol $\tilde{\mathcal{R}}$ és $\mathcal{R}$ nagyságrendje megegyezik. Tegyük fel továbbá, hogy a Q lépésoperátor szublineárisan instabil. Ekkor a kapott közelítő megoldás a $\mathcal{R}(\mathbf{h}, \delta)$ hibatag rendje szerint konvergál.

Bizonyítás. Az n-edik lépésben kapott hibát most is $\mathbf{e}^n$-nel jelölve a lépésoperátor definíciója alapján teljes indukcióval kapjuk, hogy

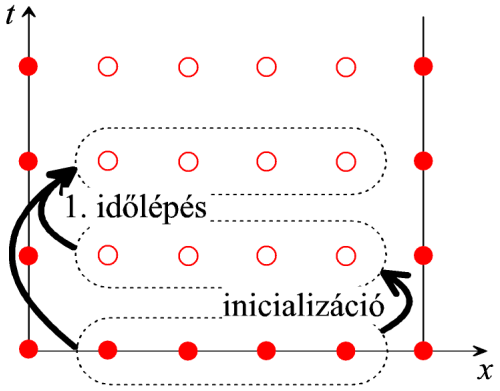
$$\begin{aligned}\mathbf{e}^n &= \mathbf{v}(n\delta,\cdot) - \mathbf{v}^n = Q\mathbf{v}((n-1)\delta,\cdot) + \delta^2\mathcal{R}(\mathbf{h},\delta) - Q\mathbf{v}^{n-1} = \\ &= Q\mathbf{e}^{n-1} + \delta^2\mathcal{R}(\mathbf{h},\delta) = \cdots = \\ &= Q^n\mathbf{e}^0 + Q^{n-1}\delta^2\mathcal{R}(\mathbf{h},\delta) + Q^{n-2}\delta^2\mathcal{R}(\mathbf{h},\delta) + \cdots + Q^0\delta^2\mathcal{R}(\mathbf{h},\delta).\end{aligned}$$

A bal oldal normáját becsülve az $\|\mathbf{e}^0\|_{\mathbf{h},p}$ nagyságrendjére vonatkozó feltevést használva ebből nyerjük az

$$\begin{aligned}\|\mathbf{e}^n\|_{\mathbf{h},p} &\le \|Q^n\|\|\mathbf{e}^0\|_{\mathbf{h},p} + \delta^2\|\mathcal{R}(\mathbf{h},\delta)\|_{\mathbf{h},p}(\|Q^{n-1}\| + \|Q^{n-2}\| + \cdots + \|Q^0\|) \le \\ &\le Kn\|\mathbf{e}^0\|_{\mathbf{h},p} + \delta^2\|\mathcal{R}(\mathbf{h},\delta)\|_{\mathbf{h},p}(n\cdot Kn) = (Kt + Kt^2)\|\mathcal{R}(\mathbf{h},\delta)\|_{\mathbf{h},p},\end{aligned}$$

vagyis a hiba olyan nagyságrendű, mint $\mathcal{R}(\mathbf{h},\delta)$, ahogy a tételben állítottuk. □

A (18.1), (illetve a (18.2)) séma vizsgálatakor nem foglalkoztunk azzal a kérdéssel, hogy ezt a gyakorlatban hogyan lehet felírni. Sőt, már a (17.1) és a (17.4) sémák elemzésekor is elhanyagoltuk ezt. Az a probléma merül ugyanis fel, hogy mind $\mathbf{u}^0$, mind az $\mathbf{u}^1$ vektorokat ismerni kellene ahhoz, hogy akár egy időlépés eredményét is ki tudjuk számolni. A kezdeti feltételben azonban csak $\mathbf{u}^0$ adott. Valahogy tehát $\mathbf{u}^1$-et is elő kellene állítani, sőt az ezt definiáló közelítésre a megfelelő rendű konvergencia érdekében a 18.3. tétel feltételeit is meg kell vizsgálnunk. Ezt az eljárást *inicializációnak* nevezzük. A következő példára vonatkozó ilyen eljárást szemléltet a 18.1. ábra.



18.1. ábra. Inicializáció és az egydimenziós hullámegyenlet rendszerként vett megoldása. Inicializáció: $\mathbf{u}^1$ előállítása $u(0,\cdot)$ és $\partial_t u(0,\cdot)$ segítségével. További lépések: $\mathbf{u}^{n+1}$ előállítása $\mathbf{u}^n$ és $\mathbf{u}^{n-1}$ segítségével.

Egy hullámegyenletre vonatkozó séma

Mindenekelőtt az egydimenziós hullámegyenletre vonatkozó korrekt kitűzésű feladatot adunk meg:

$$\begin{cases} \partial_{tt}u(t,x) = \partial_{xx}u(t,x), \quad t \in (0,T),\ x \in (0,1) \\ u(0,x) = u_0(x), \quad x \in (0,1) \\ \partial_t u(0,x) = g(x), \quad x \in (0,1) \\ u(t,0) = u_b(t),\ u(t,1) = u_j(t), \quad t \in (0,T), \end{cases} \tag{18.3}$$

ahol a u_0, g, u_b és u_j adott, 3-szor deriválható függvények. A (18.3) feladatra vonatkozó mindkét változó szerint másodrendben konvergens sémát akarunk konstruálni. Ezt úgy kell megadni, hogy egyrészt az adott peremfeltételeket is tartalmazza, másrészt a megfelelő inicializáció rendje a 18.3. tétel szerint $\delta(\mathcal{O}(\delta^2) + \mathcal{O}(h^2))$ legyen.

Diszkretizáljuk a feladatot az $\Omega_h = \{h, 2h, \dots, Nh\}$ rácson, ahol $h = \frac{1}{N+1}$; használjuk továbbá az $u_k^0 = u_0(kh)$ és a $g_k = g(kh)$ jelöléseket! A (18.2) sémát ekkor $k = 1, 2, \dots, N$ indexekre írjuk fel, ahol az $u_0^n = u_b(n\delta)$ és $u_{N+1}^n = u_j(n\delta)$ értékeket használjuk.

Ekkor a pontos megoldást az eredeti a (18.1) sémába helyettesítve teljesül, hogy annak mind a jobb, mind a bal oldala minden pontban másodrendben pontos közelítése a (18.3) feladat első sorában szereplő differenciáloperátoroknak:

$$\frac{u((n-1)\delta, kh) - 2u(n\delta, kh) + u((n+1)\delta, kh)}{\delta^2} = \partial_{tt}u(n\delta, kh) + \mathcal{O}(\delta^2),$$

továbbá

$$\frac{u(n\delta, (k-1)h) - 2u(n\delta, kh) + u(n\delta, (k+1)h)}{h^2} = \partial_{xx}u(n\delta, kh) + \mathcal{O}(h^2).$$

Ezt a számoláskor használt (18.2) séma alakjába rendezve adódik, hogy

$$\begin{aligned} u((n+1)\delta, kh) &= 2u(n\delta, kh) - u((n-1)\delta, kh) + \\ &+ \frac{\delta^2}{h^2}(u(n\delta, (k-1)h) - 2u(n\delta, kh) + u(n\delta, (k+1)h)) + \mathcal{O}(\delta^2)(\mathcal{O}(\delta^2) + \mathcal{O}(h^2)), \end{aligned}$$

vagyis a (18.2) séma pontonként másodrendben konzisztens. Mivel itt csak belső pontok szerepelnek, a 11.9. állítás eredményét egyszerűsítve kapjuk, hogy a (18.2) séma valóban másodrendben konzisztens a (18.3) feladattal.

Ezután megfelelő rendű inicializációt konstruálunk. Természetesnek tűnik az ismert első rendű derivált felhasználásával kapott

$$u_k^1 = u_k^0 + \delta g_k$$

közelítés alkalmazása. Ebbe a pontos megoldást helyettesítve

$$u(\delta, kh) = u(0, kh) + \delta u'(0, kh) + \mathcal{O}(\delta^2),$$

ami nem elégséges pontosságú közelítés.

A pontosabb eljáráshoz a (18.1) sémában szereplő egyenlőségbe a $(0, kh)$ pont körül a pontos megoldást helyettesítve és a konzisztenciarendet kiírva azt kapjuk, hogy

$$u(-\delta, kh) - 2u(0, kh) + u(\delta, kh) = \frac{u(0, (k-1)h) - 2u(0, kh) + u(0, (k+1)h)}{h^2} + \\ + \mathcal{O}(\delta^2)(\mathcal{O}(\delta^2) + \mathcal{O}(h^2)),$$

azaz

$$u(-\delta, kh) + u(\delta, kh) = 2u(0, kh) + \frac{u(0, (k-1)h) - 2u(0, kh) + u(0, (k+1)h)}{h^2} + \\ + \mathcal{O}(\delta^2)(\mathcal{O}(\delta^2) + \mathcal{O}(h^2)).$$

Ezt az

$$u(\delta, kh) - u(-\delta, kh) = \delta(u'(0, kh) + \mathcal{O}(\delta^2)) = \delta(g(kh) + \mathcal{O}(\delta^2))$$

egyenlőséggel összeadva kapjuk, hogy

$$u(\delta, kh) = u(0, kh) + \frac{1}{2} \cdot \delta g(kh) + \frac{u(0, (k-1)h) - 2u(0, kh) + u(0, (k+1)h)}{2h^2} + \\ + \mathcal{O}(\delta)(\mathcal{O}(\delta^2) + \delta\mathcal{O}(h^2)).$$

Vagyis az

$$u_k^1 = u_k^0 + \frac{1}{2} \cdot \delta g(kh) + \frac{u_0^{k-1} - 2u_0^k + u_0^{k+1}}{2h^2} + \\ + \mathcal{O}(\delta)(\mathcal{O}(\delta^2) + \delta\mathcal{O}(h^2)).$$

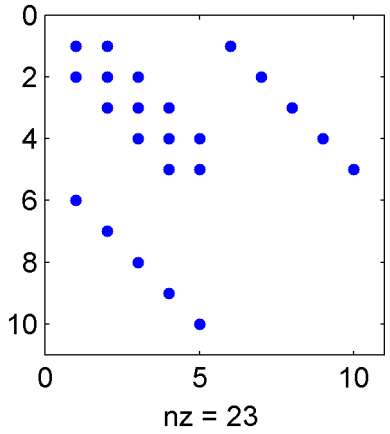
a kiszámítandó u_k^1 értékre nézve olyan inicializáció, amely a 18.3. tétel értelmében biztosítja az ezzel indított (18.2) séma mindkét változó szerint másodrendű konvergenciáját, amennyiben a séma szublineárisan instabil.

Erre vonatkozó szükséges feltételt adunk. Tudjuk, hogy a diszkretizációtól függetlenül a (t, x) ponthoz tartozó analitikus függési tartomány $[t - x, t + x]$, mert az $\mathbb{R}$ halmazon felírt egyenlet pontos megoldása

$$u(t, x) = \frac{1}{2}(u_0(x - t) + u_0(x + t)) + \int_{x-t}^{x+t} g(s)\,\mathrm{d}s.$$

Feltesszük, hogy $n\delta = t$. A séma szerint u_k^n az előző időlépésben vett u_{k-1}^{n-1}, u_k^{n-1} és u_{k+1}^{n-1} értékektől, valamint az azt megelőző lépésben kapott u_{k-1}^{n-1} értéktől függ. Ezt folytatva a 15.7. példa gondolatmenetét követve kapjuk, hogy a (t, x) ponthoz tartozó numerikus függési tartomány $[x - nh, x + nh] = [x - \frac{t}{R}, x + \frac{t}{R}]$ intervallum. Vagyis a konvergenciához szükséges feltétel $\frac{t}{R} \geq t$, azaz $R \leq 1$.

18.4. Megjegyzés.



18.2. ábra. Nemnulla elemek a (18.2) sémához tartozó lépésmátrixban, ha a (18.3) feladatban nulla peremfeltételek adottak és az intervallumot hat egyforma hosszú részre osztottuk.

1. A kapott szükséges feltétel elégséges is a konvergenciához, mert kissé hosszadalmasabb számolással látszik, hogy a (18.2) séma $R \leq 1$ esetén szublineárisan instabil.

2. Habár itt a levezetés során felhasználtuk, hogy a megoldás $t = -\delta$-ban is létezik, az inicializációs sémába helyettesítve a kívánt konzisztenciarendet akkor is megkapjuk, ha ezt nem használjuk.

3. Fontos, hogy az inicializációnak nem kell semmilyen stabilitási feltételt teljesítenie. Elegendő a lehető legegyszerűbb eljárást használni, amely megfelelő rendben konzisztens. ♢

Az egydimenziós hullámegyenlet numerikus megoldását szemléltetik a 20.2.17. és a 20.2.18. animációk.

A hullámegyenlet numerikus megoldásának kvalitatív tulajdonságait vizsgálja az [1] könyv.

egyenletek esetén]$\boldsymbol{L_2}$-norma megőrzése folytonos idejű egyenletek esetén

18.1. L_2-norma megőrzése folytonos idejű egyenletek esetén

18.5. Lemma. *Legyen μ az Ω tartományon értelmezett mérték. Tegyük fel, hogy valamilyen $\alpha \in \boldsymbol{\alpha}$-ra a $\partial_t u + F(u, \partial_x u, \ldots, \partial_x^{(\alpha)} u) = 0$ egyenlethez tartozó valamilyen Ω*

tartományon értelmezett kezdetiérték-feladat megoldása korlátos valamilyen $(0,T)$ *intervallumon, valamint* $\partial_t u$ *is az. Ekkor pontosan abban az esetben lesz* $t \to \|u(t,\cdot)\|_{L_2(\Omega)}$ *konstans, illetve monoton csökkenő, ha az* u *megoldásra*

$$\int_\Omega u\, F(u, \partial_x u, \dots, \partial_x^{(\alpha)} u)\, \mathrm{d}\mu = 0$$

teljesül, illetve ha

$$\int_\Omega u\, F(u, \partial_x u, \dots, \partial_x^{(\alpha)} u)\, \mathrm{d}\mu \le 0.$$

Bizonyítás. Felhasználva a korlátossági feltételeket, nyerjük az

$$\frac{1}{2}\partial_t \int_\Omega u^2\, \mathrm{d}\mu = \int_\Omega u\, \partial_t u\, \mathrm{d}\mu,$$

egyenlőtlenséget, vagyis az eredeti egyenlet mindkét oldalát az u megoldással szorozva kapjuk, hogy

$$\frac{1}{2}\partial_t \|u(t,\cdot)\|^2_{L_2(\Omega),\mu} = \frac{1}{2}\partial_t \int_\Omega u^2(t,\cdot)\, \mathrm{d}\mu = \int_\Omega u\, \partial_t u\, \mathrm{d}\mu = \int_\Omega u\, F(u, \partial_x u, \partial_{xx} u)\, \mathrm{d}\mu.$$

A jobb oldal itt pontosan akkor nulla, illetve legfeljebb nulla, ha a bal oldal is ilyen, ez viszont azt jelenti, hogy az $\|u(t,\cdot)\|_{L_2(\Omega),\mu}$ mennyiség állandó, illetve monoton csökken.□

18.6. Megjegyzés. Amennyiben μ a számláló-mérték, akkor az állítást a szemidiszkretizációra is lehet használni. Ekkor az Ω tartományon való integrálás az Ω_h halmazbeli rácspontokon való összegzést jelenti. ♢

A következő példákban az L_2 norma változását vizsgáljuk három fontos feladat esetében

18.7. Példa. Diffúziós feladat, azaz

$$\partial_t u(t,\mathbf{x}) = \Delta u(t,\mathbf{x}) \quad t \in (0,T),\ \mathbf{x} \in \Omega$$

homogén Dirichlet- vagy homogén Neumann-peremfeltétellel.
Ekkor a Gauss-tétel alkalmazásával

$$\int_\Omega u\, \Delta u = \int_\Omega u\, \nu \cdot \nabla u - \int_\Omega \nabla u\, \nabla u = -\int_\Omega \nabla u\, \nabla u \le 0,$$

azaz nem növekszik, sőt nem konstans megoldás esetén csökken az $\|u(t,\cdot)\|_{L_2(\Omega)}$ norma.♢

18.8. Példa. KdV egyenlet homogén peremfeltételekkel vagy periodikus peremfeltételekkel egydimenziós esetben.
Az egyenlet általános alakja ekkor

$$\begin{cases} \partial_t u(t,x) = \partial_x(\alpha u^2(t,x) + k\partial_{xx}u(t,x)) \quad t \in (0,T),\ x \in (a,b) \\ u(t,a) = u(t,b),\ u'(t,a) = u'(t,b),\ u''(t,a) = u''(t,b), \end{cases}$$

ahol α és k pozitív konstansok. Ekkor tagonként vizsgálva a lemmában szereplő mennyiséget (mindenhol $\Omega = (a,b)$)

$$\int_\Omega u\,\partial_x(\alpha u^2) = -\int_\Omega \partial_x u\,\alpha u^2 = -\frac{1}{2}\int_\Omega \alpha u\,2u\,\partial_x u = -\frac{1}{2}\int_\Omega u\,\partial_x(\alpha u^2), \tag{18.4}$$

vagyis az elején szereplő mennyiség önmagának $-\frac{1}{2}$-szerese, azaz nulla. Másodszor

$$\begin{aligned} &\int_\Omega u\,\partial_{xxx}ku = -k\int_\Omega \partial_x u\,\partial_{xx}u + ku(b)\,\partial_{xx}u(b) - ku(a)\,\partial_{xx}u(a) = \\ &= -k\int_\Omega \partial_x u\,\partial_{xx}u = k\int_\Omega \partial_{xx}u\,\partial_x u - k\partial_x u(b)\,\partial_x u(b) + k\partial_x u(a)\,\partial_x u(a) = \\ &= k\int_\Omega \partial_{xx}u\,\partial_x u, \end{aligned}$$

ahol a szereplő tagok egymás ellentettjei, vagyis az összes mennyiség nulla. ♢

18.9. Példa. A Burgers-egyenlet egydimenziós esetben.

$$\begin{cases} \partial_t u(t,x) + u(t,x)\,\partial_x u(t,x) = \partial_t u(t,x) + \partial_x u^2(t,x) = 0 \quad t \in (0,T),\ x \in (a,b) \\ u(t,a) = u(t,b),\ u'(t,a) = u'(t,b) \end{cases}$$

alakú, azaz a fent szereplő (18.4) tagot kell vizsgálni, és mivel az nulla, ezért a megoldás $L_2(a,b)$ normája itt is állandó. ♢

18.2. Megmaradó mennyiségek a (szemi-) diszkretizált egyenletekben

Mivel az itteni átalakítások fő eleme a parciális integrálás elve, meg kell vizsgálnunk, hogy egyrészt ez hogyan és milyen feltételekkel teljesül az operátorok véges differencia közelítésére, másrészt pedig, hogy milyen véges differencia közelítéseket alkalmazhatunk. A továbbiakban $\mathbf{u} \cdot v$ az $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ és $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ vektorok $\mathbb{R}^n$-beli skalárszorzatát jelöli. Használjuk továbbá az $\mathbf{uv}$ jelölést is a fenti két vektor komponensenkénti szorzatára, ahogy az $\mathbf{u}^2$-vel jelölt négyzetre emelést is komponensenként értjük.

A továbbiakban a D_0 jelölést (amelyhez nincs feltétlenül egy h_x felosztásparaméter rendelve) az alábbi értelemben használjuk:

$$D_0 : \mathbb{R}^n \to \mathbb{R}^n, \text{ ill. } \mathbb{R}^n \to \mathbb{R}^{n-2},$$
$$D_0(\mathbf{u})[i] = u_{i+1} - u_{i-1}, \ i = 1, 2, \ldots, n, \text{ ill. } i = 2, 3, \ldots, n-1,$$

attól függően, hogy értelmezzük-e periodikus vagy éppen Dirichlet-peremfeltétellel az $i = 1, i = n$ esetre vonatkozó különbségeket.

18.10. Lemma. *Tegyük fel, hogy az* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ *vektorokra periodikus peremfeltételek teljesülnek. Ekkor*

$$D_0\mathbf{u} \cdot \mathbf{v} = -\mathbf{u} \cdot D_0\mathbf{v}.$$

Bizonyítás. Itt a periodikus peremfeltételek miatt lehet a $D_0\mathbf{u}$ mennyiséget a vektorok első és utolsó komponensében is értelmezni. Ekkor teljesül, hogy $u_{n+1}v_n = u_1v_0$, valamint $u_0v_1 = u_nv_{n+1}$, azaz kapjuk, hogy

$$\begin{aligned} D_0\mathbf{u} \cdot \mathbf{v} &= (u_2 - u_0)v_1 + (u_{n+1} - u_{n-1})v_n + \sum_{i=2}^{n-1}(u_{i+1} - u_{i-1})v_i \\ &= u_2v_1 + u_1v_0 - u_nv_{n+1} - u_{n-1}v_n + \sum_{i=2}^{n-1} u_{i+1}v_i - \sum_{i=2}^{n-1} u_{i-1}v_i \\ &= \sum_{i=0}^{n-1} u_{i+1} \cdot v_i - \sum_{i=2}^{n+1} u_{i-1} \cdot v_i = \sum_{j=1}^{n} u_j \cdot v_{j-1} - \sum_{j=1}^{n} u_j \cdot v_{j+1} \\ &= \sum_{j=1}^{n}(u_j, (v_{j-1} - v_{j+1})) = -\mathbf{u} \cdot D_0\mathbf{v}, \end{aligned}$$

ami bizonyítja a lemma állítását. □

18.11. Megjegyzés. A parciális integrálás további alkalmazásainak megfelelő formulák találhatók a 21.30. és a 21.31. feladatokban. ♢

Először a szemidiszkretizált problémát vizsgáljuk a periodikus peremfeltétellel ellátott Burgers-egyenletre vonatkozó kezdetiérték-feladat esetében. Megjegyezzük, hogy habár

$$\frac{1}{2}\partial_x u^2 = u\,\partial_x u$$

teljesül, a diszkretizált verzióra ez nem feltétlenül igaz, azaz például

$$\frac{1}{2}D_0\mathbf{u}^2 \neq \mathbf{u}D_0\mathbf{u}.$$

Ezért a Burgers-egyenletet nem a logikusnak tűnő egyszerű módon diszkretizáljuk, hanem amint a következő lemma mutatja, egy összetettebb (szemi-) diszkretizáció célszerűbb.

18.12. Állítás. *Az l_2 norma konstans marad minden időpontban az alábbi (szemidiszkretizációval kapott) séma megoldása során:*

$$\partial_t \mathbf{u}(t,\cdot) + \frac{\theta}{2} D_0 \mathbf{u}^2(t,\cdot) + (1-\theta)\mathbf{u}(t,\cdot) D_0 \mathbf{u}(t,\cdot) = 0, \tag{18.5}$$

ahol $\mathbf{u} : (0,T) \to \mathbb{R}^n$ periodikus peremfeltétellel, valamint $\theta = \frac{2}{3}$.

Bizonyítás. A (18.5)-beli szemidiszkretizációra alkalmazzuk a 18.5. lemma állítását, így elegendő azt belátni, hogy a

$$\mathbf{u} \cdot \left[\frac{\theta}{2} D_{0,h} \mathbf{u}^2 + (1-\theta)\mathbf{u} D_{0,h} \mathbf{u} \right]$$

skalárszorzat a szemidiszkretizált feladat $\mathbf{u}$ megoldása esetén nulla. A továbbiakban a $(t,\cdot)$ argumentumot elhagyjuk. A 18.10. lemma eredményét használva (a feltételek szerint periodikus peremfeltétellel rendelkező vektorokra) kapjuk, hogy

$$\begin{aligned} \mathbf{u} \cdot \left[\frac{\theta}{2} D_{0,h} \mathbf{u}^2 + (1-\theta)\mathbf{u} D_{0,h} \mathbf{u} \right] &= -D_{0,h}\mathbf{u} \cdot \frac{\theta}{2}\mathbf{u}^2 + \mathbf{u}^2 \cdot (1-\theta) D_{0,h}\mathbf{u} = \\ = -\mathbf{u} D_{0,h}\mathbf{u} \cdot \frac{\theta}{2}\mathbf{u} - D_{0,h}\mathbf{u}^2 \cdot (1-\theta)\mathbf{u} &= -\mathbf{u} \cdot (1-\theta) D_{0,h}\mathbf{u}^2 - \mathbf{u} \cdot \frac{\theta}{2} \mathbf{u} D_{0,h}\mathbf{u}, \end{aligned}$$

ahol az első és az utolsó kifejezés valóban egymás ellentettje, ha $\theta = \frac{2}{3}$, azaz ekkor mindkettő nulla, amiből a 18.5. lemma szerint az l_2 norma megmaradása valóban következik. □

Ezután a Burgers-egyenletre vonatkozó kezdetiérték-feladat teljes diszkretizációját adjuk meg:

$$0 = \frac{1}{\delta}(u_k^{n+1} - u_k^n) + \frac{1}{3} D_0 [u_k^{n+\frac{1}{2}}]^2 + \frac{1}{3} u_k^{n+\frac{1}{2}} D_0 u_k^{n+\frac{1}{2}}, \quad k = 1, 2, \ldots, N \tag{18.6}$$

ahol $u_k^{n+\frac{1}{2}} = \frac{u_k^{n+1} + u_k^n}{2}$.

18.13. Állítás. *A periodikus peremfeltételekkel, azaz $u_0^n = u_N^n$ és $u_{N+1}^n = u_1^n$ egyenlőségekkel adott (18.6) séma olyan, hogy minden időlépésben (azaz tetszőleges n-re) $\|\mathbf{u}^n\|_2$ konstans.*

Bizonyítás. Szorozzuk meg (18.6) mindkét oldalát $\mathbf{u}^{n+\frac{1}{2}}$-del! Ekkor a 18.12. állítás eredményét az $\mathbf{u}^{n+\frac{1}{2}}$ vektorra alkalmazva azt kapjuk, hogy

$$\begin{aligned} 0 &= \mathbf{u}^{n+\frac{1}{2}} \cdot \left[\frac{1}{\delta}(\mathbf{u}^{n+1} - \mathbf{u}^n) + \frac{1}{2h} \left(\frac{1}{3} D_{0,h} [\mathbf{u}^{n+\frac{1}{2}}]^2 + \frac{1}{3} \mathbf{u}^{n+\frac{1}{2}} D_{0,h} \mathbf{u}^{n+\frac{1}{2}} \right) \right] \\ &= \mathbf{u}^{n+\frac{1}{2}} \cdot \frac{1}{\delta} [\mathbf{u}^{n+1} - \mathbf{u}^n] = \frac{1}{2\delta} [\mathbf{u}^{n+1} + \mathbf{u}^n] \cdot [\mathbf{u}^{n+1} - \mathbf{u}^n] = \frac{1}{2\delta} \left(\|\mathbf{u}^{n+1}\|_2^2 - \|\mathbf{u}^n\|_2^2 \right), \end{aligned}$$

amiből valóban az következik, hogy az $\|\cdot\|_2$ norma az időlépés során nem változik. □

A fenti eljárás, azaz (18.6) időlépésének végrehajtása azonban nem magától értetődő, ugyanis közben egy nemlineáris egyenletet megoldása adja $\mathbf{u}^{n+1}$ értékét. Ezért részletesebben is leírjuk az erre szolgáló algoritmust.

A $(0,1)$ intervallum $0 = x_0, x_0 + h = x_1, \ldots, x_N = 1$ felosztását tekintjük, ahol az ismeretlenek az $x_1, x_2, \ldots, x_N$-beli értékek.

A lehető legegyszerűbb esetet tárgyaljuk, a Newton-iterációval egyetlen lépést teszünk. A fenti okoskodás alapján a (18.6)-beli időlépés végrehajtásához a

$$0 = [f(\mathbf{v})]_k = \frac{1}{\delta} v_k + \frac{1}{3}\left(\left(\frac{v_{k+1}+u_{k+1}^n}{2}\right)^2 - \left(\frac{v_{k-1}+u_{k-1}^n}{2}\right)^2\right)$$
$$+ \theta \cdot \frac{v_k + u_k^n}{2} \cdot \left(\frac{v_{k+1}+u_{k+1}^n}{2} - \frac{v_{k-1}+u_{k-1}^n}{2}\right) - \frac{1}{\delta} u_k^n, \quad k = 1, 2, \ldots, N$$

egyenletrendszert kell megoldanunk a $\mathbf{v}$ ismeretlenre nézve, illetve megoldását közelítenünk. Ezt a Newton-iteráció egyetlen lépésével tesszük, vagyis az

$$\mathbf{u}^{n+1,1} = \mathbf{u}^n - [\nabla f(\mathbf{u}^n)]^{-1} f(\mathbf{u}^n) \tag{18.7}$$

képlettel adott mennyiséget számítjuk ki. Kézenfekvő módszer az, ha az iterációt az előző időlépésben felvett értékből indítjuk. Itt $\nabla f(\mathbf{u}^n) \in \mathbb{R}^{N\times N}$, azaz

$$\nabla f(\mathbf{u}^n)[j,k] = \partial_{v_k} [f(\mathbf{v})]_j |_{\mathbf{v}=\mathbf{u}^n}$$

ahol speciálisan

$$\nabla f(\mathbf{u}^n)[k,k] = \frac{1}{\delta} + \frac{1}{3}\left(\frac{v_{k+1}+u_{k+1}^n}{2} - \frac{v_{k-1}+u_{k-1}^n}{2}\right)|_{\mathbf{v}=\mathbf{u}^n} = \frac{1}{\delta} + \frac{1}{3}\left(u_{k+1}^n - u_{k-1}^n\right)$$
$$\nabla f(\mathbf{u}^n)[k,k-1] = -\frac{1}{3}(v_{k-1}+u_{k-1}^n) - \frac{1}{6}(v_k+u_k^n)|_{\mathbf{v}=\mathbf{u}^n} = \frac{2}{3}u_{k-1}^n - \frac{1}{3}u_k^n$$
$$\nabla f(\mathbf{u}^n)[k,k+1] = -\frac{1}{3}(v_{k+1}+u_{k-1}^n) + \frac{1}{6}(v_k+u_k^n)|_{\mathbf{v}=\mathbf{u}^n} = \frac{2}{3}u_{k+1}^n + \frac{1}{3}u_k^n.$$

A (18.7) formulában szereplő mennyiségek tehát a következők:

$$f(\mathbf{u}^n) = \begin{pmatrix} \frac{1}{3}([u_2^n]^2 - [u_N^n]^2) + \frac{2}{3}u_1^n(u_2^n - u_N^n) \\ \frac{1}{3}([u_3^n]^2 - [u_1^n]^2) + \frac{2}{3}u_k^n(u_3^n - u_1^n) \\ \vdots \\ \frac{1}{3}([u_1^n]^2 - [u_{N-1}^n]^2) + \frac{2}{3}u_k^n(u_1^n - u_{N-1}^n) \end{pmatrix},$$

valamint

$$\nabla f(\mathbf{u}^n) = \begin{pmatrix} \frac{1}{\delta} + \frac{u_2^n - u_N^n}{3} & \frac{2}{3}u_2^n + \frac{1}{3}u_1^n & 0 & \ldots & 0 & \frac{2}{3}u_N^n - \frac{1}{3}u_1^n \\ \frac{2}{3}u_1^n - \frac{1}{3}u_2^n & \frac{1}{\delta} + \frac{u_3^n - u_1^n}{3} & \frac{2}{3}u_3^n + \frac{1}{3}u_2^n & \ldots & 0 & 0 \\ \vdots & \vdots & & & \vdots & 0 \\ \frac{2}{3}u_1^n + \frac{1}{3}u_N^n & 0 & \ldots & 0 & \frac{2}{3}u_{N-1}^n - \frac{1}{3}u_N^n & \frac{1}{\delta} + \frac{u_1^n - u_{N-1}^n}{3} \end{pmatrix}.$$

18.3. Egy séma a Korteweg–de Vries-egyenletre

A fenti eljárást a KdV egyenletre is alkalmazni akarjuk.

Látjuk, hogy a Burgers-egyenlet megoldására vonatkozó (18.6) séma kedvező tulajdonsággal rendelkezik. De ez implicit, méghozzá az ismeretlen komponensben másodfokú. Ezért a nemlineáris egyenlet kezelésére egy konkrét eljárást is mutatunk, amelyet kiterjesztünk a KdV egyenlet esetére is, azaz az $u : (0,T)\times(0,1)$ függvényre vonatkozó alábbi feladatra

$$\begin{cases} \partial_t u + \partial_x(u\partial_x u + \partial_{xx}u) = 0, & t \in (0,T) \\ u(0,x) = g(x), & x \in (0,1) \\ u(t,0) = u(t,1),\ \partial_x u(t,0) = \partial_x u(t,1), & t \in (0,T), \end{cases}$$

ahol $g : (0,1) \to \mathbb{R}$ adott. A (18.6) sémának megfelelően az alábbi közelítést szeretnénk használni:

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \delta\left(\frac{1}{3}D_0[\mathbf{u}^{n+\frac{1}{2}}]^2 + \frac{1}{3}\mathbf{u}^{n+\frac{1}{2}}D_0\mathbf{u}^{n+\frac{1}{2}} + D_0D_0^2\mathbf{u}^{n+\frac{1}{2}}\right) =$$
$$= \mathbf{u}^n - \delta\left(\frac{1}{12}D_0(\mathbf{u}^{n+1}+\mathbf{u}^n)^2 + \frac{1}{12}(\mathbf{u}^{n+1}+\mathbf{u}^n)D_0(\mathbf{u}^{n+1}+\mathbf{u}^n) + \frac{1}{2}D_0D_0^2(\mathbf{u}^{n+1}+\mathbf{u}^n)\right),$$

ahol a jobb oldalon megjelenik $[\mathbf{u}^{n+1}]^2$. Az $[\mathbf{u}^{n+1}]^2$ értékére vonatkozólag a nemlineáris rendszer megoldását közelítjük, méghozzá úgy, hogy ennek egy $[u_k^{n+1,j}]^2$ iterációs lépésében (komponensenként) az alábbi approximációt használjuk:

$$[u_k^{n+1,j}+S]^2 \approx [u_k^{n+1,j-1}+S]^2 + 2(u_k^{n+1,j-1}+S)(u_k^{n+1,j}-u_k^{n+1,j-1}) =$$
$$= 2(u_k^{n+1,j-1}+S)(u_k^{n+1,j}+S) - [u_k^{n+1,j-1}+S]^2.$$

A séma ezzel a következőképpen változik:

$$\begin{aligned} u_k^{n+1,j} - \delta\Bigg(&\frac{\theta}{8}D_{0,h}(2(u_k^{n+1,j-1}+u_k^n)(u_k^{n+1,j}+u_k^n) - (u_k^{n+1,j-1}+u_k^n)^2) \\ &+ \frac{1-\theta}{4}(u_k^{n+1,j}+u_k^n)D_{0,h}(u_k^{n+1,j-1}+u_k^n) + \frac{1}{2}D_{0,h}D_{0,h}^2(u_k^{n+1,j-1}+u_k^n)\Bigg), \end{aligned} \tag{18.8}$$

amelyben $u_k^{n+1,0} := u_k^n$, majd $u_k^{n+1,1}$ értékét egy *lineáris* egyenletrendszer megoldásával kapjuk ebből, és ugyanígy $u_k^{n+1,2}$ értékét a (18.8) formulából $u_k^{n+1,1}$ felhasználásával. Ezután legyen $u_k^{n+1} := u_k^{n+1,2}$, amivel egy két iterációt tartalmazó időlépést adtunk meg.

18.14. Megjegyzés. Használhattuk volna (18.8) utolsó tagjában a $u_k^{n+1,j-1}$ helyett az $u_k^{n+1,j}$ értéket is, a séma ugyanúgy lineáris maradt volna, azonban a rendszer megoldásához szükséges egyenletrendszerben kapott mátrixszal való számolás időigényesebb lenne. ♢

19. fejezet

Időfüggő PDE-ek megoldása végeselem-módszerrel

Ebben a fejezetben lineáris parabolikus feladatok megoldásának végeselemes közelítésével foglalkozunk. A végeselem-módszer részleteinek megértéséhez az I. részben leírt ismeretek elsajátítása szükséges. Itt az elmélet vázát és az időfüggő PDE-ek megoldására vonatkozó egy konkrét példát ismertetünk. Az egyszerűség kedvéért az elméleti vizsgálatot homogén problémákra korlátozzuk.

19.1. A vizsgált feladat, feltevések, jelölések

A végeselem-módszer bevezetéséhez a vizsgált feladatot az

$$\begin{cases} \partial_t u(t) = -Au(t) \\ u(0) = u_0 \end{cases} \tag{19.1}$$

alakba írjuk, ahol a keresett ismeretlen függvény $u : (0,T) \to L_2(\Omega)$ típusú. Az $A : L_2(\Omega) \to L_2(\Omega)$ operátorról feltesszük, hogy az egy erősen folytonos félcsoportot generál, ami annak felel meg, hogy a (19.1)-beli *absztrakt Cauchy-probléma* megoldása létezik és egyértelmű. A t időpontbeli megoldást a

$$T(t)(u_0) = e^{-At}(u_0)$$

formula adja meg, ahol $\{T(t)\}_{t\geq 0}$ az A által generált erősen folytonos félcsoport.
Formálisan azt is mondhatjuk, hogy a $\{T(t)\}_{t\geq 0}$ félcsoport a $-A$ differenciáloperátor exponenciális függvénye.
Az A operátorra vonatkozó feltevés alapján annak sűrűn definiáltnak kell lennie. A pontos értelmezési tartományát a következő szakaszban jellemezzük.

Ahogy az első részben is, a $\langle \cdot, \cdot \rangle_0$ szimbólum az $L_2(\Omega)$-beli skalárszorzatot jelöli.

19.1. Feltevés. Az egész fejezetben feltesszük, hogy A sajátértékei pozitívak, emellett A^{-1} kompakt és önadjungált. ♢

Ekkor a kompakt operátorokra vonatkozó Schauder-elméletet felhasználva fennáll a következő:

19.2. Következmény. *Az A operátor $\{b_j\}_{j\in S}$ sajátvektorai teljes ortogonális rendszert alkotnak, amelyhez tartozó sajátértékek $\{\lambda_j\}_{j\in S}$, ahol S egy véges halmaz vagy $S=\mathbb{N}$. Ezekkel tetszőleges $v\in H$ esetén*

$$Av=\sum_{j\in S}\lambda_j b_j\langle b_j,v\rangle_0. \tag{19.2}$$

19.3. Példa. A leggyakrabban vizsgált esetben

- $H=L_2(\Omega)$
- $A=-\Delta$, ahol $\mathcal{D}(-\Delta)=H_0^1(\Omega)\cap H^2(\Omega)$. ♢

19.2. Gyenge alak, a numerikus megoldás módszere

A (19.1) feladat gyenge alakját kétféle ok miatt is célszerű felírni. Egyrészt a vizsgált esetekben A mindig egy differenciáloperátor, amelyet eleve csak így tudunk értelmezni olyan függvények esetén, amelyek nem deriválhatók klasszikus értelemben annyiszor, amennyi az A rendje. Másrészt ez teszi lehetővé, hogy természetes módon értelmezzük a megoldás végeselemes közelítését.
A gyenge alakhoz többféle módon is eljuthatunk.

- A (19.1) feladatban szereplő egyenletben mindkét oldalt disztribúciónak tekintve mondhatjuk, hogy az egyenlet ekvivalens az

  $$\langle\partial_t u(t),v\rangle_0=\langle Au(t),v\rangle_0=a(u(t),v)\quad\forall v\in\mathcal{D}(\Omega)$$

  alakkal, ahol az $a:H\times H\to\mathbb{R}$ bilineáris formát a disztribúciós deriválás képlete, illetve valamilyen integrálátalakító formula segítségével kapjuk. Ha ez folytonos valamilyen $H^*\subset L_2(\Omega)$ altéren értelmezett (valamilyen skalárszorzásból származó) $\|\cdot\|_*$ norma szerint, akkor a $H^*\supset\mathcal{D}(\Omega)$ teret tegyük esszerint teljessé! Az így kapott Hilbert-teret H jelöli. Ekkor a $H:=\overline{H^*}$ tér elemeire teljesül, hogy

  $$\langle\partial_t u(t),v\rangle_0=a(u(t),v)\quad\forall v\in\overline{H^*}=H.$$

  Ennek megfelelően (19.1) gyenge alakja a következő: Olyan $u(t):(0,T)\to H$ függvényt keresünk, amelyre minden $v\in H$ esetén teljesül, hogy

  $$\langle\partial_t u(t),v\rangle_0=a(u(t),v). \tag{19.3}$$

- Nyilvánvaló, hogy a (19.1) feladatban szereplő egyenlet két oldala pontosan akkor egyezik meg, ha minden $v \in H$ esetén a v-vel vett skalárszorzatuk azonos. Így azon $u \in H$ megoldásokra és $v \in H$ elemekre, amelyekre valamilyen integrálátalakító formulát lehet használni, kapjuk, hogy

$$\langle \partial_t u(t), v\rangle_0 = \langle -Au(t), v\rangle_0 = a(u(t), v),$$

ahol $a : H \times H \to \mathbb{R}$ bilineáris függvény. Azonban csupán ebből a szokásos levezetésből nem világos, hogy mi pontosan a értelmezési tartománya, továbbá az sem, hogy azon milyen normát érdemes értelmezni.

A 19.1. feltevés alapján kapjuk, hogy $a : \overline{H^*} \times \overline{H^*} \to \mathbb{R}$ pozitív definit.

19.4. Példa. A 19.3. példa folytatásaként legyen $H^* = C_0^2(\Omega)$; ekkor $u(t), v \in C_0^2(\Omega)$ esetén

$$\langle \partial_t u(t), v\rangle_0 = \langle -\Delta u(t), v\rangle_0 = \langle \nabla u(t), \nabla v\rangle_0.$$

Itt

$$\|u\|_*^2 = \langle \nabla u, \nabla u\rangle_0,$$

vagyis $C_0^2(\Omega)$-nek az erre vonatkozó lezártja $H = H_0^1(\Omega)$. Azaz a feladat gyenge alakjában olyan $u \in H_0^1(\Omega)$ függvényt keresünk, amelyre minden $u \in H_0^1(\Omega)$ függvény esetén

$$\langle \partial_t u(t), v\rangle_0 = \langle \nabla u, \nabla v\rangle_0 \tag{19.4}$$

teljesül. ♢

19.3. Szemidiszkretizáció

A közelítő megoldás kiszámításához a feladatot először térben diszkretizáljuk, a közelítő megoldást egy rögzített $V_h \subset H$ véges dimenziós altérben keressük. A (19.3) formula alapján kapjuk a következő feladatot:
Olyan $u_h : (0, t) \to V_h$ függvényt keresünk, hogy

$$\langle \partial_t u_h(t), v_h\rangle_0 = a(u_h(t), v_h) \quad \forall v \in V_h. \tag{19.5}$$

Mivel V_h véges dimenziós, ezért a keresett függvény

$$u_h(t) = \sum_{j=1}^{S} C_j(t) b_{h,j}$$

alakba írható, ahol $\{b_{h,j}\}_{j=1}^S$ a V_h egy bázisa. Fennáll továbbá, hogy a (19.5) egyenlőség pontosan akkor teljesül, ha az az összes $b_{j,h}$ báziselemre igaz.

Ezeket a (19.5) formulába beírva kapjuk, hogy

$$\langle\sum_{j=1}^{S}\dot{C}_j(t)b_j,b_k\rangle_0 = a(\sum_{j=1}^{S}C_j(t)b_j,b_k) \quad \forall b_k\in V_h,\ k=1,2,\ldots,S. \tag{19.6}$$

Az ebből származó közönséges differenciálegyenlet-rendszer megadásához szükségünk lesz a $\{b_{h,j}\}_{j=1}^{S}$ bázishoz tartozó B-vel jelölt *merevségi mátrixra* és E-vel jelölt *tömegmátrixra*, amelyeket a

$$B[j,k] = a(b_k,b_j)$$
$$E[j,k] = \langle b_k,b_j\rangle_0$$

egyenlőségekkel definiálunk.
Ezekkel a (19.6) gyenge alak a következő módon írható fel:

$$E[\dot{C}_1(t),\dot{C}_2(t),\ldots,\dot{C}_S(t)]^T = B[C_1(t),C_2(t),\ldots,C_S(t)]^T, \tag{19.7}$$

amely valóban az $[C_1(t),C_2(t),\ldots,C_S(t)]$ (vektor)függvényre vonatkozó közönséges differenciálegyenlet-rendszerre vezet.

19.3.1. Közelítő megoldás kiszámítása egy példán

A szemidiszkretizáció után a (19.7) közönséges differenciálegyenlet-rendszer megoldásának közelítését kell már csak elvégeznünk. Az általános formalizmus helyett egy konkrét módszer esetén mutatjuk ezt be.

Az implicit Euler-módszert választjuk. Ekkor a (19.6)-beli ismeretlen függvények

$$C_j(n\delta)\approx C_j^n,\quad j=1,2,\ldots,S,\ n=1,2,\ldots,\frac{T}{\delta}$$

közelítését alkalmazva azt kapjuk, hogy

$$\langle\sum_{j=1}^{S}\frac{C_j^{n+1}-C_j^n}{\delta}b_j,b_k\rangle_0 = a(\sum_{j=1}^{S}C_j^{n+1}b_j,b_k)\ \forall b_k\in\overline{V_h},\ k=1,2,\ldots,S. \tag{19.8}$$

vagyis az $\mathbf{C}^{n+1}=(C_1^{n+1},C_2^{n+1},\ldots,C_S^{n+1})^T$ ismeretlenekből álló vektorra a (19.7) alaknak megfelelően a következő teljesül:

$$E\frac{1}{\delta}[\mathbf{C}^{n+1}-\mathbf{C}^n] = B\mathbf{C}^{n+1}.$$

Innen kapjuk a megoldásra vonatkozó időlépést:

$$\mathbf{C}^{n+1} = (E-\delta B)^{-1}E\mathbf{C}^n.$$

Végeselem-módszerekre vonatkozik a 20.2.11. és a 20.2.12. animáció.
A kapcsolódó elmélet részletes leírása a [30] könyvben található.

20. fejezet

Számítógépes alkalmazások

20.1. Programok

Az időfüggő feladatok numerikus megoldásához is megadunk két mintaprogramot a korábban bemutatott eljárások gyakorlati bemutatására. Az első az egydimenziós hővezetési egyenlet numerikus megoldását mutatja be, a második pedig az advekciós egyenletre konstruálja meg a Lax–Wendroff-sémát. A programokkal a következő fejezet több animációja is rekonstruálható a paraméterek megfelelő beállításával.

20.1. Példa. Tekintsük az egydimenziós hővezetési feladatot a $[0, \pi]$ intervallumon homogén Dirichlet-peremfeltétel mellett (legyen $\sigma_D = 1$). Az alábbi program a numerikus megoldást mutatja. A hőmérsékletértékeket az egyes időrétegeken a v vektor elemei közelítik. A programhoz az $u0$ vektorban adhatjuk meg a kezdeti feltételt, az n paraméter adja meg a belső osztópontok számát, T az az időpont, ameddig szeretnénk megkapni a numerikus megoldást, r a rácsparaméter, a θ paraméterrel pedig beállíthatjuk, hogy az időintegrációt milyen módszerrel hajtjuk végre: $\theta = 0$ explicit Euler-módszer, $\theta = 1$ implicit Euler-módszer, $\theta = 1/2$ Crank–Nicolson-módszer.

```
clear all; close all; % minden korábbi adat törlése

% Input adatok megadása

n=50; h=pi/(n+1); x=h*[1:n]; % osztópontok száma a [0,1]-en
T=1; % t [0,T]-ben fut
theta=0; % teta paraméter
r=0.4; % rácsparaméter
u0=pi-2*abs(x-pi/2); % kezdeti feltétel

delta=r*h^2;
```

```
kmax=round(T/delta);

% Az iterációs mátrixok konstrukciója

N=sparse(2:n,1:n-1,ones(n-1,1),n,n);
I=speye(n);
Q=-2*speye(n)+N+N';

A1=I-theta*((delta/h^2)*Q);
A2=I+(1-theta)*((delta/h^2)*Q);

% Az iterációs vektor inicializálása

v=u0';

% Az iterációs lépések megvalósítása

for k=1:kmax

   v=A1\(A2*v);

   % Ábrázolás (peremfeltétellel együtt)

   plot(0:h:pi,[0,v',0],'bo')
   axis([0,pi,0,pi])
   xlabel('x','FontSize',14)
   ylabel('u','FontSize',14)
   title(['t= ',num2str(k*delta,'%2.4f')],'Color','r','FontSize',14)
   pause(delta*100)

end;
```

20.2. Példa. Tekintsük az advekciós egyenletet $a = 1$ választással a $[0, 1]$ intervallumon periodikus peremfeltétel mellett. Az alábbi program a Lax–Wendroff-módszerrel nyert numerikus megoldást adja abban az esetben, ha a kezdeti feltételt $(\sin(2\pi x))$ az $u0$ vektorban tároljuk, a [0,1] intervallumot 20 ekvidisztáns intervallumra osztottuk, és az R rácsparaméter értékét 0.75-nek választottuk. A program az animáció során mutatja a pontos megoldást is.

```
clear all; close all; % minden korábbi adat törlése
```

```
% Input adatok megadása

R=0.75; % CFL-együttható
n=20; h=1/n; delta=R*h; % osztóintervallumok száma, rácstávolság
T=5; % t [0,T]-ben fut
kmax=round(T/delta)% % lépések száma
u0=sin(pi*2*h*[1:n]); % kezdeti feltétel

% Lépésmátrix konstrukciója

B=sparse(1:n-1,2:n,ones(1,n-1),n,n);
A=B'*1/2*R*(1+R)+B*(-1/2*R)*(1-R)+(1-R^2)*speye(n);
A(1,n)=1/2*R*(1+R); A(n,1)=-1/2*R*(1-R); % periodikus peremfelt. miatt

% A kezdeti vektor inicializációja

v=u0';

% Az iterációs lépések (ábrázolással együtt)

for k=1:kmax

v=A*v;

vpontos=(sin(([1:100]/100-k*delta)*2*pi))';
plot([0:1/100:1],[vpontos(100),vpontos'],'r-',[0:h:1],[v(n),v'],'bo-');
xlabel('x','FontSize',14)
ylabel('u','FontSize',14,'Rotation',0)
title(['t= ',num2str(k*delta,'%2.3f')],'Color','r','FontSize',14)
pause(0.2);

end
```

20.2. Animációk

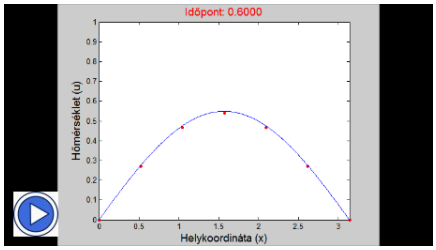
20.2.1. Az egydimenziós hővezetési egyenlet numerikus megoldása

Ebben a részben a

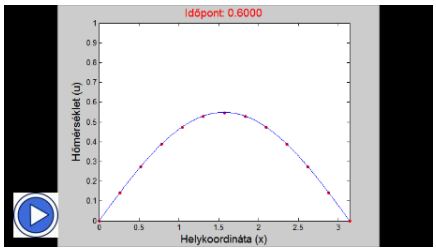
$$\begin{aligned} \partial_t u(t,x) &= \partial_{xx} u(t,x), \quad t \in (0,T),\ x \in (0,\pi) \\ u(t,0) &= u(t,\pi) = 0 \quad t \in (0,T) \\ u(0,x) &= u_0(x) \quad x \in [0,\pi] \end{aligned}$$

(vö. (10.10)) egydimenziós hővezetési egyenlet megoldásának közelítésére adunk meg szimulációkat. Mindegyik esetben a véges differenciák módszerét használjuk. A szimulációk közti különbséget a kezdeti feltétel, a rácstávolságok ill. az időintegrációban alkalmazott numerikus módszer megválasztása adja. A numerikus megoldást piros pontokkal jelöltük, a pontos megoldást (amennyiben ismert) kék folytonos vonal jelöli.

20.2.1 Animáció. A (10.16) séma (explicit Euler-módszer) eredménye a $h = \pi/6$, $\delta = 0.1$ ($r = 0.3647$), $u_0(x) = \sin x$, $T = 1$ választással. A 20.2.2. animáción ugyanezen feladat numerikus megoldása látható ugyanekkora rácsparaméterrel egy finomabb rácson.

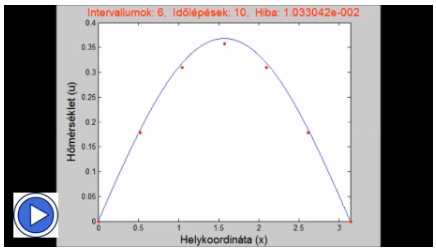


20.2.2 Animáció. A (10.16) séma (explicit Euler-módszer) eredménye a $h = \pi/12$, $\delta = 0.025$ ($r = 0.3647$), $u_0(x) = \sin x$, $T = 1$ választással. A 20.2.1. animáción ugyanezen feladat numerikus megoldása látható ugyanekkora rácsparaméterrel egy durvább rácson.

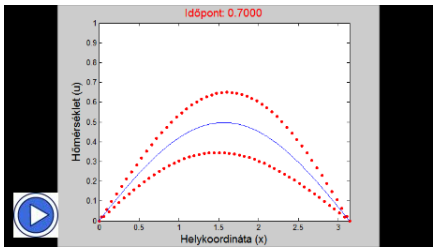


20.2.3 Animáció. A (10.16) séma (explicit Euler-módszer) konvergenciájának szemléltetése. Az animáció azt mutatja, hogy hogyan változik a maximum-normabeli hiba a

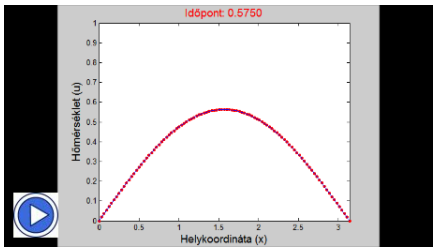
$T = 1$ időrétegen abban az esetben, ha a rácsparamétert konstansnak tartva felezzük a rácstávolságot és negyedeljük az időlépést. Vegyük észre, hogy minden finomítási lépésben kb. negyedelődik a hiba, ami másodrendű konvergenciát mutat.



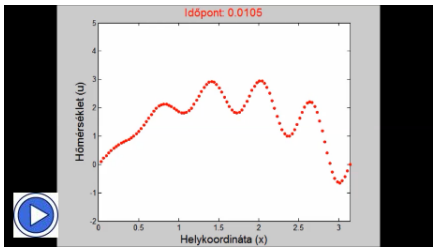
20.2.4 Animáció. A (10.16) séma (explicit Euler-módszer) eredménye a $h = \pi/100$, $\delta = 0.0005$ ($r = 0.5066$), $u_0(x) = \sin x$, $T = 0.75$ választással. Mivel $r > 0.5$, a numerikus megoldás nem stabil. Ennek eredménye, hogy a lépésmátrix 1-nél nagyobb abszolút értékű sajátértékéhez tartozó sajátvektor (ami a kezdeti vektor előállításában nem is szerepel) a gépi kerekítések miatt felerősödik az iterációs vektorban. Ez a jelenség figyelhető meg az animációban a $t = 0.6$ időponttól (az animációban ez a rész lassítva szerepel).



20.2.5 Animáció. A (12.7) séma (Crank–Nicolson-módszer) eredménye a $h = \pi/100$, $\delta = 0.005$ ($r = 5.066$), $u_0(x) = \sin x$, $T = 2$ választással. Az animáció azt mutatja, hogy a feltétel nélkül stabil Crank–Nicolson-módszer $r = 5$-ös rácsparaméterrel is (10-szer akkorával, ami az explicit Euler-módszer esetén megengedett) teljesen kielégítő megoldást szolgáltat. Ilyenkor nem a stabilitás, hanem az elérni kívánt pontosság határozza meg r és az időlépés értékét.



20.2.6 Animáció. A (12.7) séma (Crank–Nicolson-módszer) eredménye a $h = \pi/100$, $\delta = 0.0001$ ($r = 0.1013$), $T = 1$ választással. Az animáció a hővezetés jelenségét szemlélteti egy általános kezdeti függvény esetén.



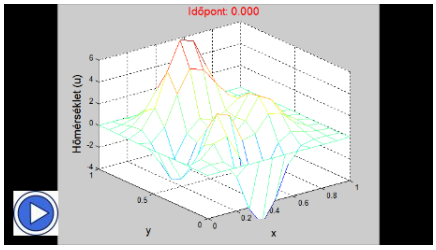
20.2.2. A kétdimenziós hővezetési egyenlet numerikus megoldása

Most áttérünk a kétdimenziós hővezetési egyenlet megoldásainak szimulációira. A

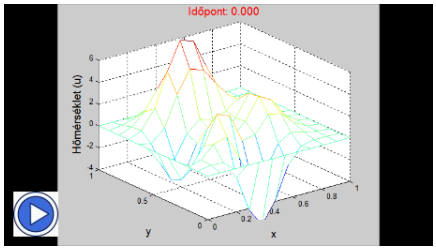
$$\partial_t u(t,x,y) = \partial_{xx} u(t,x,y) + \partial_{yy} u(t,x,y) \tag{20.1}$$

(vö. (13.2)) egyenletet oldjuk meg az egységnégyzeten és a $(0,T]$ időintervallumon homogén Dirichlet-peremfeltétellel. Kezdeti függvényként a MATLAB peaks nevű beépített kétváltozós függvényét használjuk. Vizsgáljuk először a véges differenciás megoldásokat!

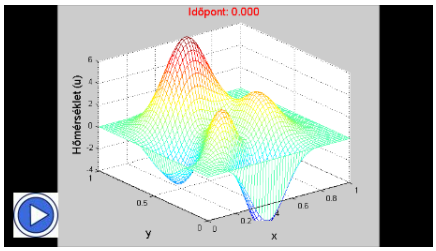
20.2.7 Animáció. A (13.3) séma (explicit Euler) eredménye a $h_x = h_y = 1/11$, $\delta = 0.001983$ ($r_x = r_y = 0.24$), $T = 0.2$ választással.



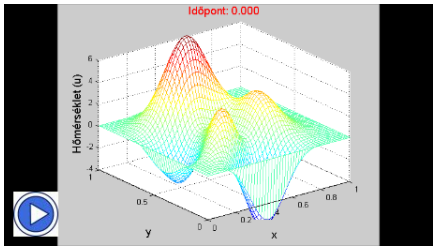
20.2.8 Animáció. A (13.3) séma (explicit Euler) eredménye a $h_x = h_y = 1/11$, $\delta = 0.002149$ ($r_x = r_y = 0.26$), $T = 0.8$ választással. Egy kicsit megnövelve az időlépést már sérül a stabilitási feltétel. Emiatt lép fel a $t = 0.5$ időponttól látható jelenség.



20.2.9 Animáció. A (13.3) séma (explicit Euler) eredménye a $h_x = h_y = 1/51$, $\delta = 1.1534e-004$ ($r_x = r_y = 0.3$), $T = 0.008$ választással. Az animáció ismét a nem stabil esetet szemlélteti, most egy finomabb rács esetén.

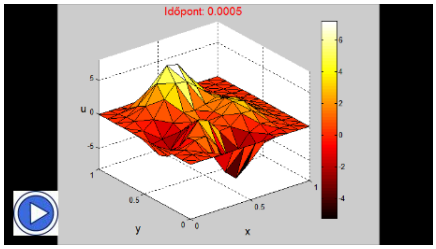


20.2.10 Animáció. A (13.12) séma (Crank–Nicolson) eredménye a $h_x = h_y = 1/51$, $\delta = 7.6893e-004$ ($r_x = r_y = 2$), $T = 0.1$ választással. Az animáció azt szemlélteti, hogy a Crank–Nicolson-séma nagyobb rácsparaméterek esetén is stabil megoldást ad.

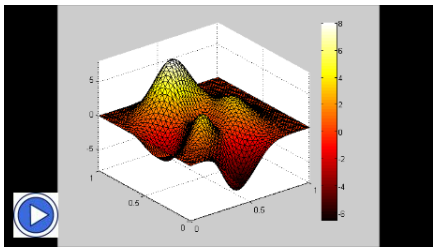


Most áttérünk a feladat végeselemes szimulációinak bemutatására. Mindegyik esetben egyenletes háromszögrácsot és szakaszonként lineáris bázisfüggvényeket használunk. Mutatunk példát Neumann-típusú peremfeltételre is.

20.2.11 Animáció. A numerikus megoldás szimulációja a $T = 0.0455$ időpontig abban az esetben, ha 10-10 belső osztópontot veszünk fel x és y irányban is. A szemidiszkrét feladat megoldására a Cranck–Nicolson-módszert használtuk $\Delta t = 0.0005$ választással.



20.2.12 Animáció. Ebben a szimulációban a peremfeltétel az $y = 0$ oldalon homogén Neumann, a többin homogén Dirichlet. Az animáció a numerikus megoldást mutatja a $T = 0.045$ időpontig abban az esetben, ha 40-40 belső osztópontot veszünk fel x és y irányban is. A szemidiszkrét feladat megoldására a Crank–Nicolson-módszert használtuk $\Delta t = 0.0005$ választással.



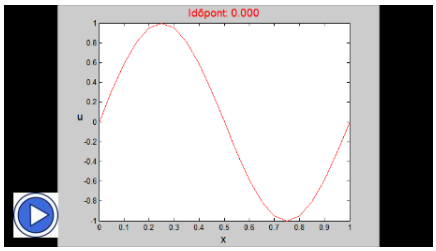
20.2.3. Az advekciós egyenlet numerikus megoldása

Az advekciós egyenlet numerikus megoldásainak bemutatásához a

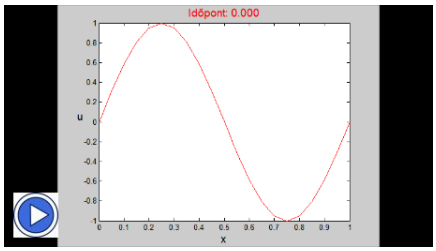
$$\begin{cases} \partial_t u(t,x) + a\partial_x u(t,x) = 0, \quad t \in (0,T),\ x \in (0,1) \\ u(0,x) = \sin(2\pi x), \quad x \in (0,1) \end{cases} \tag{20.2}$$

feladatot választottuk periodikus peremfeltétellel ($u(t,0) = u(t,1)$). Feltesszük, hogy az advekció sebessége pozitív. Az alábbi animációk három sémát hasonlítanak össze: upwind, Lax-Wendroff- és leapfrog. A sémák diszkrét Fourier-analíziséből tudjuk, hogy mindhárom sémára a CFL-feltétel ($R \leq 1$) egyúttal elégséges is a stabilitáshoz. $R = 1$ esetén a rácspontokban a pontos megoldást kapjuk, így ebben az esetben nem változik az amplitúdó. $R < 1$ esetén csak a leapfrog séma tudja ezt. A Lax–Wendroff-séma kevésbé, az upwind jobban csökkenti az amplitúdót. A numerikus megoldás fázisa is általában eltér a pontos megoldásétól. Általában lemarad a numerikus megoldás a pontos megoldáshoz képest. A Lax–Wendroff-sémánál és a leapfrognál kb. ugyanakkora ez a lemaradás. Az upwind séma esetén a numerikus megoldás $R > 1/2$ esetén siet, $R < 1/2$ esetén lemarad. $R = 1/2$ esetén lényegében nincs fázistolás.

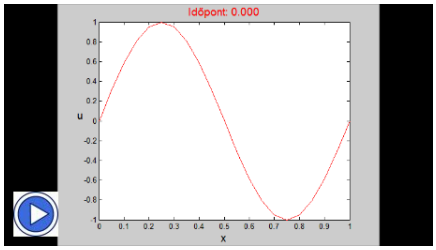
20.2.13 Animáció. Az upwind sémával nyert numerikus megoldás $R = 0.75$, $h = 1/20$ választással.



20.2.14 Animáció. A Lax–Wendroff-sémával nyert numerikus megoldás $R = 0.75$, $h = 1/20$ választással.

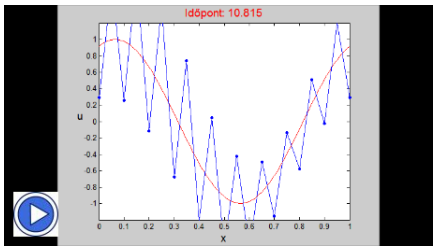


20.2.15 Animáció. A leapfrog sémával nyert numerikus megoldás $R = 0.75$, $h = 1/20$ választással.



A következő animáció azt az esetet mutatja, amikor a Lax-Wendroff sémánál nem teljesül a CFL-feltétel.

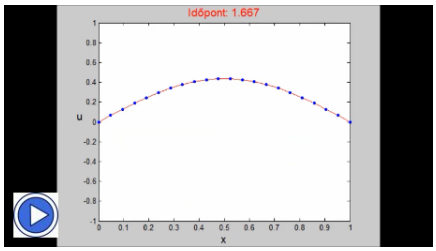
20.2.16 Animáció. A Lax–Wendroff-sémával nyert numerikus megoldás $R = 1.05$, $h = 1/20$ választással. Az instabilitás jelensége körülbelül $t = 9$-től látható.



20.2.4. Az egydimenziós hullámegyenlet numerikus megoldása

A $\partial_{tt}u(t,x) = \partial_{xx}u(t,x)$ egydimenziós hullámegyenletet tekintjük homogén Dirichlet-féle peremfeltétellel. A numerikus megoldáshoz a véges differenciák módszerét használjuk ((18.1) séma).

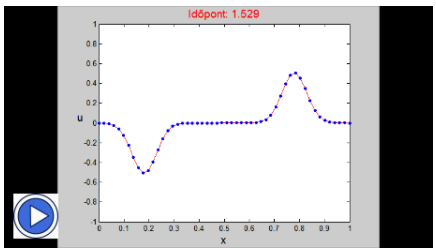
20.2.17 Animáció. Ebben az animációban a kezdeti feltétel $u(0,x) = \sin(\pi x)$, $\partial_t u(0,x) = 0$. A numerikus megoldáshoz 20 belső pontot vettünk fel a [0,1] intervallumon. A sémát az $R = 1$ választással alkalmazzuk a [0,6] időintervallumon.



20.2.18 Animáció. Ebben az animációban a kezdeti feltétel

$$u(0,x) = \exp(-(x-0.3)^2/0.005), \quad \partial_t u(0,x) = 0.$$

A numerikus megoldáshoz 50 belső pontot vettünk fel a [0,1] intervallumon. A sémát az $R = 1$ választással alkalmazzuk a [0,6] időintervallumon. Az animáció jól szemlélteti, hogy az u megoldásfüggvény egy jobbra és egy balra haladó hullám összege.



21. fejezet

Feladatok

21.1. Feladat. Adjuk meg az alábbi közelítések rendjét! Ha egy f függvény k-adrendű deriváltját közelítjük, akkor mindenhol feltesszük, hogy $f \in C^{k+1}(\bar{\Omega})$-beli.

$$\text{(a) } f'(x) \approx \frac{1}{2h}(f(x+h) - f(x-h))$$
$$\text{(b) } f'(x) \approx \frac{1}{2h}(-3f(x) + 4f(x+h) - f(x+2h))$$
$$\text{(c) } \partial_{xy} f(x,y) \approx \frac{1}{4h^2}(f(x+h,y+h) - f(x+h,y-h) -$$
$$- f(x-h,y+h) + f(x-h,y-h)$$

21.2. Feladat. Közelítsük a $\partial_x[q(x)\partial_x]$ differenciáloperátort azzal a véges differenciával, amely egy olyan rácson értelmezett, ahol az egész indexek egy egyenletes felosztás rácspontjait, a tört indexűek értelemszerűen az ezek közötti felezőpontokat jelölik:

$$\frac{1}{h_x^2}\left(q^{n-\frac{1}{2}}u^{n-1} - (q^{n+\frac{1}{2}} + q^{n-\frac{1}{2}})u^n + q^{n+\frac{1}{2}}u^{n+1}\right).$$

Számítsuk ki a közelítés rendjét, ha a $q \in C^2(\mathbb{R})$ feltevéssel élünk!

21.3. Feladat. Igazoljuk, hogy ha (pl. egy véges differencia közelítéssel kapott megoldás-) sorozat konvergens $\|\cdot\|_{\mathbf{h},\infty}$ normában, akkor tetszőleges $p \in \mathbb{R}^+$ esetén konvergens a $\|\cdot\|_{\mathbf{h},p}$ norma szerint is! Mutassunk példát arra, hogy a fordított irányú következtetés nem igaz semmilyen véges p esetén sem!

21.4. Feladat. Tekintsük a

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x), \quad t \in \mathbb{R}^+,\ x \in (0,\pi) \\ \partial_x u(t,0) = \partial_x u(t,\pi) = 0, \quad t \in \mathbb{R}^+ \\ u(0,x) = \sin x \quad x \in (0,\pi). \end{cases}$$

feladat közelítésére vonatkozó

$$\begin{cases} u_k^{n+1} = u_k^n + \frac{\sigma_D \delta}{h^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n), \quad k = 1, 2, \ldots, N \\ 3u_0^n = 4u_1^n - u_2^n, \ -u_{N-1}^n + 4u_N^n = 3u_{N+1}^n, \\ u_k^0 = \sin \frac{k\pi}{N+1}, \quad k = 0, 1, \ldots, N, N+1 \end{cases}$$

sémát, amelyet a $(0, \pi)$ intervallum egy egyenletes felosztásán adunk meg, ahol $x_0 = 0$ és $x_{N+1} = \pi$ a peremrácspontok. Számítsuk ki ennek konzisztenciarendjét!

21.5. Feladat. Tekintsük a

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x) + f(t,x), \quad t \in \mathbb{R}^+, \ x \in (0,\pi) \\ u(t,0) = u(t,\pi) = 0, \quad t \in \mathbb{R}^+ \\ u(0,x) = \sin x \quad x \in (0,\pi). \end{cases}$$

feladat közelítésére vonatkozó

$$\begin{cases} u_k^{n+1} = u_k^n + \frac{\sigma_D \delta}{h^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n) + \frac{\delta}{2}(f(n\delta, x_k) + f((n+1)\delta, x_k)), \quad k = 1, 2, \ldots, N \\ u_0^n = u_{N+1}^n = 0, \\ u_k^0 = \sin \frac{k\pi}{N+1}, \quad k = 0, 1, \ldots, N, N+1 \end{cases}$$

sémát, amelyet a $(0, \pi)$ intervallum egy egyenletes felosztásán adunk meg, ahol $x_0 = 0$ és $x_{N+1} = \pi$ a peremrácspontok, továbbá $f \in C(0, \pi)$ adott. Számítsuk ki ennek konzisztenciarendjét!

21.6. Feladat. Igazoljuk, hogy a

$$\begin{cases} \partial_t u(t,x) + a\partial_x u(t,x) = \sigma \partial_{xx} u(t,x), \quad t \in (0,T), x \in \mathbb{R} \\ u(0,x) = f(x), \quad x \in \mathbb{R} \end{cases}$$

feladat

$$\begin{cases} u_k^{n+1} + a\frac{\delta}{2h}(u_{k+1}^{n+1} - u_{k-1}^{n+1}) = u_k^n + \sigma \frac{\delta}{h^2}(u_{k+1}^{n+1} - 2u_k^{n+1} + u_{k-1}^{n+1}) \\ u_k^0 = f(kh) \end{cases}$$

implicit sémával való közelítése pontonként konzisztens, és adjuk is meg a (pontonkénti) konzisztencia rendjét!

21.7. Feladat. Igazoljuk, hogy a

$$\partial_t u(t,x) = \partial_{xx} u(t,x), \quad t \in (0,T), x \in \mathbb{R}$$

egyenlethez tartozó

$$u_k^{n+1} - u_k^{n-1} = \frac{2\delta}{h^2}(u_{k+1}^{n+1} - 2u_k^{n+1} + u_{k-1}^{n+1})$$

séma pontonként konzisztens a fenti egyenlettel! Számítsuk ki a konzisztenciarendet!

21.8. Feladat. Igazoljuk, hogy adott $f \in C^2(\mathbb{R}^+ \times \mathbb{R})$ esetén a

$$\partial_t u(t,x) = \partial_{xx} u(t,x) + f(t,x), \quad t \in (0,T), x \in \mathbb{R}$$

egyenlethez tartozó

$$u_k^{n+1} - u_k^{n-1} = \frac{2\delta}{h^2}(u_{k+1}^{n+1} - 2u_k^{n+1} + u_{k-1}^{n+1}) + \delta(f((n+1)\delta, x_k) + f((n-1)\delta, x_k))$$

séma pontonként konzisztens a fenti egyenlettel! Számítsuk ki a konzisztenciarendet!

21.9. Feladat. Igazoljuk, hogy a

$$\begin{cases} \partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x), & t \in (0,T), x \in \mathbb{R} \\ u(0,x) = f(x), & x \in \mathbb{R} \end{cases}$$

feladathoz tartozó

$$u_k^{n+1} = u_k^n + \sigma_D \frac{\delta}{h^2}(u_{k+1}^{n+1} - 2u_k^{n+1} + u_{k-1}^{n+1})$$

implicit séma a $\|\cdot\|_{h,2}$ norma szerint konzisztens a fenti egyenlettel!

21.10. Feladat. Milyen feltétellel teljesül, hogy a

$$\partial_t u(t,x) = \partial_{xx} u(t,x), \quad t \in (0,T), x \in \mathbb{R}$$

egyenlethez tartozó

$$u_k^{n+1} - u_k^{n-1} = \frac{2\delta}{h^2}(u_{k+1}^n - (u_k^{n+1} + u_k^{n-1}) + u_{k-1}^n)$$

séma pontonként konzisztens a fenti egyenlettel?

21.11. Feladat. Igazoljuk, hogy a

$$\begin{cases} \partial_t u(t,x) = a \partial_x u(t,x), & t \in (0,T), x \in \mathbb{R} \\ u(0,x) = f(x), & x \in \mathbb{R} \end{cases}$$

feladathoz tartozó

$$u_k^{n+1} = u_k^n + 2R(u_{k+1}^{n+1} - u_{k-1}^{n+1})$$

implicit séma pontonként konzisztens a fenti egyenlettel! Határozzuk meg a konzisztenciarendet!

21.12. Feladat. Igazoljuk, hogy a kétdimenziós

$$\partial_t u = a\partial_x u + b\partial_y u$$

hullámegyenletre vonatkozó természetesnek tűnő

$$u_{j,k}^{n+1} = u_{j,k}^n - R_x D_{0,x} u_{j,k}^n + \frac{R_x^2}{2} D_{0,x}^2 u_{j,k}^n - R_y D_{0,y} u_{j,k}^n + \frac{R_y^2}{2} D_{0,y}^2 u_{j,k}^n$$

séma nem konzisztens másodrendben az időváltozó szerint! Hogy módosítsuk, hogy valóban másodrendben konzisztens legyen? (Itt érdemes általánosítani a (14.5) formulát, és az azt követő gondolatmenetet.)

21.13. Feladat. Igazoljuk, hogy a Peaceman–Rachford-séma forrástaggal vett

$$\begin{cases} \left(I - \frac{r_x}{2} D_{0,x}^2\right) u^{n+\frac{1}{2}} = \left(I + \frac{r_y}{2} D_{0,y}^2\right) u^n + f^{n+\frac{1}{2}} \\ \left(I - \frac{r_y}{2} D_{0,y}^2\right) u^{n+1} = \left(I + \frac{r_x}{2} D_{0,x}^2\right) u^{n+\frac{1}{2}} + f^{n+\frac{1}{2}} \end{cases}$$

verziója minden változó szerint másodrendben konzisztens a következő egyenlettel:

$$\partial_t u(t,x,y) = \partial_{xx} u(t,x,y) + \partial_{yy} u(t,x,y) + f(t,x,y).$$

21.14. Feladat. Tekintsük a

$$\begin{cases} \partial_t u(t,x) = \partial_{xx} u(t,x) + tx \quad t \in \mathbb{R}^+,\ x \in (0,\pi) \\ u(t,0) = 1, \quad u(t,\pi) = 2 \quad t \in \mathbb{R}^+ \\ u(0,x) = \sin x \quad x \in (0,\pi). \end{cases}$$

feladat közelítésére vonatkozó

$$\begin{cases} u_k^{n+1} = u_k^n + \frac{\delta}{h^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n) + \delta \cdot n\delta x_k, \quad k = 1,2,\ldots,N \\ u_0^n = 1, \quad u_{N+1}^n = 2, \\ u_k^0 = \sin\frac{k\pi}{N+1} \quad k = 0,1,\ldots,N,N+1 \end{cases}$$

sémát, amelyet a $(0,\pi)$ intervallum egy egyenletes felosztásán definiálunk, ahol $x_0 = 0$ és $x_{N+1} = \pi$ a peremrácspontok!

Írjuk ezt fel (inhomogén) lineáris rendszer alakjában, azaz adjuk meg az

$$\mathbf{u}^{n+1} = A\mathbf{u}^n + \mathbf{b}^n$$

alakú lépésoperátorban szereplő A mátrixot és $\mathbf{b}$ vektort!

21.15. Feladat. Tekintsük a

$$\begin{cases} \partial_t u(t,x) = \partial_{xx} u(t,x) - tx \quad t \in \mathbb{R}^+,\ x \in (0,\pi) \\ \partial_x u(t,0) = 1, \quad \partial_x u(t,\pi) = -1 \quad t \in \mathbb{R}^+ \\ u(0,x) = \sin x \quad x \in (0,\pi). \end{cases}$$

feladat közelítésére vonatkozó

$$\begin{cases} u_k^{n+1} = u_k^n + \frac{\delta}{h^2}(u_{k-1}^n - 2u_k^n + u_{k+1}^n) - \delta \cdot n\delta x_k, \quad k = 1,2,\dots,N \\ u_1^n - u_0^n = \frac{1}{N+1}, \quad u_{N+1}^n - u_{N+1}^n = -\frac{1}{N+1}, \\ u_k^0 = \sin\frac{k\pi}{N+1} \quad k = 0,1,\dots,N,N+1 \end{cases}$$

sémát, amelyet a $(0,\pi)$ intervallum egy egyenletes felosztásán definiálunk, ahol $x_0 = 0$ és $x_{N+1} = \pi$ a peremrácspontok!

Írjuk ezt fel (inhomogén) lineáris rendszer alakjában, azaz adjuk meg az

$$\mathbf{u}^{n+1} = A\mathbf{u}^n + \mathbf{b}^n$$

alakú lépésoperátorban szereplő A mátrixot és $\mathbf{b}$ vektort!

21.16. Feladat. Írjuk fel a $Q = (0,1) \times (0,1)$ tartományon adott

$$\begin{cases} \partial_t u(t,x,y) = \partial_{xx} u(t,x,y) + \partial_{yy} u(t,x,y) \quad (x,y) \in Q \\ u(x,0) = 0,\ \partial_y u(x,1) = 0, \quad x \in (0,1) \\ u(0,y) = 0,\ \partial_y u(1,y) = 0, \quad x \in (0,1) \end{cases} \tag{21.1}$$

feladat numerikus megoldására vonatkozó azon sémához tartozó lépésmátrixot, ahol egy egyenletes rácson diszkretizáljuk a feladatot, továbbá a második derivált közelítésére az $r_x D_{0,x}^2 + r_y D_{0,y}^2$ sémát használjuk, a homogén Neumann-peremfeltételt pedig a legegyszerűbb jobb-, illetve baloldali elsőrendű véges differenciával közelítjük!

21.17. Feladat. Módosítsuk a 21.16. feladatot úgy, hogy a perem alján, illetve bal oldalán az $u(x,0) = x$, illetve az $u(0,y) = y$ Dirichlet-peremfeltételek legyenek adottak. Ekkor a megfelelő rendszer $\mathbf{u}^{n+1} = A\mathbf{u}^n + \mathbf{b}^n$ alakú lesz. Számítsuk ki A és $\mathbf{b}^n$ értékét!

21.18. Feladat. Igazoljuk, hogy a

$$\partial_t u(t,x) + a\partial_x u(t,x) = 0, \quad t \in (0,T), x \in \mathbb{R}$$

egyenlethez tartozó

$$u_k^{n+1} = (1-R)\,u_k^n + Ru_{k-1}^n$$

séma pontosan akkor stabil, ha $a > 0$ és $0 < R \leq 1$ teljesül!

21.19. Feladat. Igazoljuk, hogy a

$$\partial_t u(t,x) + a\partial_x u(t,x) = 0, \quad t \in (0,T), x \in \mathbb{R}$$

egyenlethez tartozó

$$u_k^{n+1} = u_k^n - \frac{R}{2}(u_{k+1}^n - u_{k-1}^n)$$

séma semmilyen a és R érték mellett sem lehet stabil!

21.20. Feladat. Igazoljuk, hogy az egydimenziós

$$\partial_t u + a\partial_x u = 0$$

egyenletre vonatkozó

$$u_k^{n+1} = \frac{1}{2}(u_{k+1}^n + u_{k-1}^n) - \frac{R}{2}(u_{k+1}^n - u_{k-1}^n)$$

séma (Lax–Friedrichs-séma) pontosan akkor stabil, ha $|R| \leq 1$ teljesül!

21.21. Feladat. Igazoljuk, hogy a

$$\partial_t u(t,x) = \sigma_D \partial_{xx} u(t,x), \quad t \in (0,T), x \in \mathbb{R}$$

egyenlethez tartozó

$$u_k^{n+1} - \frac{r}{2}(u_{k+1}^{n+1} - 2u_k^{n+1} + u_{k-1}^{n+1}) = u_k^n + \frac{r}{2}(u_{k+1}^n - 2u_k^n + u_{k-1}^n)$$

Crank–Nicolson-séma feltétel nélkül stabil!

21.22. Feladat. Igazoljuk, hogy a

$$\partial_t u(t,x) = a\partial_x u(t,x) + \sigma_D \partial_{xx} u(t,x), \quad t \in (0,T), x \in \mathbb{R}$$

egyenlethez tartozó

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \frac{R}{2} D_0 \mathbf{u}^{n+1} + r D_0^2 \mathbf{u}^{n+1}$$

séma feltétel nélkül stabil!

21.23. Feladat. Tekintsük az $u : (0,T) \times \mathbb{R}^3 \to \mathbb{R}$ típusú függvényre vonatkozó

$$\partial_t u = \partial_{xx} u + \partial_{yy} u + \partial_{zz} u$$

egyenlet (valamilyen konkrét kezdeti feltételhez tartozó) megoldásának numerikus közelítésére felírt

$$\begin{cases} \left(I - \frac{r_x}{3} D_{0,x}^2 - \frac{r_y}{3} D_{0,y}^2\right) \mathbf{u}^{n+\frac{1}{3}} = \left(I + \frac{r_z}{3} D_{0,z}^2\right) \mathbf{u}^n \\ \left(I - \frac{r_y}{3} D_{0,y}^2 - \frac{r_z}{3} D_{0,z}^2\right) \mathbf{u}^{n+\frac{2}{3}} = \left(I + \frac{r_x}{3} D_{0,x}^2\right) \mathbf{u}^{n+\frac{1}{3}} \\ \left(I - \frac{r_z}{3} D_{0,z}^2 - \frac{r_x}{3} D_{0,x}^2\right) \mathbf{u}^{n+1} = \left(I + \frac{r_y}{3} D_{0,y}^2\right) \mathbf{u}^{n+\frac{2}{3}} \end{cases}$$

sémát!
Igazoljuk, hogy ez feltétel nélkül stabil!

21.24. Feladat. Tudjuk, hogy a homogén Neumann-peremfeltételekkel rendelkező $\partial_t u = \sigma \partial_{xx} u$ feladat megoldásának közelítésére vonatkozó egy egyszerű explicit Euler séma lépésmátrixa

$$\begin{pmatrix} 1-r & r & 0 & 0 & \dots & 0 \\ r & 1-2r & r & 0 & \dots & 0 \\ \vdots & & & & & \vdots \\ 0 & \dots & 0 & r & 1-2r & r \\ 0 & \dots & 0 & 0 & r & 1-r \end{pmatrix}.$$

Mutassuk meg, hogy a stabilitás egy elégséges feltétele most is $0 \le r \le \frac{1}{2}$ teljesülése! Ez szükséges is?

21.25. Feladat. Írjuk fel a fenti 21.24. feladat módosításaként az implicit Euler sémához tartozó lépésmátrixot, majd igazoljuk, hogy az (a peremfeltétel nélküli esethez hasonlóan) most is feltétel nélkül stabil sémát ad!

21.26. Feladat. Adjuk meg a 21.16. feladat azon módosításához tartozó lépésmátrixot, ahol mindenütt homogén Neumann-peremfeltételt tekintünk! Mutassuk meg, hogy a stabilitás (itt is) pontosan akkor teljesül, ha $0 \le r_x + r_y \le \frac{1}{2}$ fennáll!

21.27. Feladat. Írjuk fel a a homogén Dirichlet- és a homogén Neumann-peremfeltételekkel rendelkező $\partial_t u = \sigma \partial_{xx} u$ feladat megoldására felírt Crank–Nicolson-sémához tartozó lépésmátrixot! Igazoljuk, hogy mindkét séma feltétel nélkül stabil! (Segítség: Írjuk fel a lépésmátrixokat $(I+Q)^{-1}(I-Q)$ alakba!)

21.28. Feladat. Vizsgáljuk a $\partial_t \mathbf{v}(t,x) = A \partial_x \mathbf{v}(t,x)$ advekciós egyenlethez tartozó

$$\mathbf{v}^{n+1} = \mathbf{v}^n + \frac{R}{2} A D_0 \mathbf{v}^n + \frac{R^2}{2} A^2 D_0^2 \mathbf{v}^n, \quad R = \frac{\delta}{h}$$

Lax–Wendroff-sémát, ahol A valós, diagonalizálható mátrix valós $\lambda_1, \lambda_2, \dots, \lambda_l$ sajátértékekkel! A skalár esetre vonatkozó eredmény felhasználásával igazoljuk, hogy ez pontosan akkor stabil, ha $R|\lambda_j| \le 1$ minden $j = 1, 2, \dots, l$ esetén!

21.29. Feladat. Vizsgáljuk a 21.28. feladatban szereplő egyenlethez tartozó

$$\mathbf{v}^{n+1} - \frac{R}{2} A D_0 \mathbf{v}^{n+1} = \mathbf{v}^n, \quad R = \frac{\delta}{h}$$

sémát, ahol A most is valós, diagonalizálható mátrix valós $\lambda_1, \lambda_2, \dots, \lambda_l$ sajátértékekkel! Igazoljuk, hogy a skalár esethez hasonlóan ez feltétel nélkül stabil!

21.30. Feladat. Milyen $\mathbb{R}^n \to \mathbb{R}^n$ operátor legyen $D_\square$, hogy a periodikus vektorok halmazán minden $\mathbf{u}, \mathbf{v}$ párra teljesüljön a

$$(D_+\mathbf{u}, \mathbf{v}) = (\mathbf{u}, D_\square\mathbf{v})$$

egyenlőség?

21.31. Feladat. Legyenek az $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$ vektorok periodikus peremfeltétellel ellátva. Konstruáljuk meg az ilyen peremfeltétellel ellátott u és v függvényekre vonatkozó

$$(\partial_{xx}u, v) = -(\partial_x u, \partial_x v)$$

azonosságnak megfelelő formulát $\mathbf{u}$ és $\mathbf{v}$ esetére!

Irodalomjegyzék

[1] Ascher, U., *Numerical Methods for Evolutionary Differential Equations*, SIAM, 2008.

[2] Atkinson, K., Han. W., *Theoretical numerical analysis: a functional analysis framework*, Springer, 2009.

[3] Axelsson, O., *Iterative Solution Methods,* Cambridge University Press, 1994.

[4] Arnold, D.N., Lecture notes on partial differential equations, University of Minnesota, 2011.

[5] Arnold, D.N., Brezzi, F., Cockburn, B., Marini, L.D., Unified analysis of discontinuous Galerkin methods for elliptic problems, *SIAM J. Numer. Anal.* 39(5):1749-1779, 2002.

[6] Benzi, M., Golub, G. H., Liesen, J., Numerical solution of saddle point problems, *Acta Numer.* 14 (2005), 1–137.

[7] Braess, D., *Finite Elements*, Cambridge University Press, 1997.

[8] Ciarlet, P. G., *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam, 1978.

[9] Czách L., Simon L., *Parciális differenciálegyenletek*, egyetemi jegyzet, Tankönyvkiadó, Budapest, 1970.

[10] Elman, H. C., Silvester, D. J., Wathen, A. J., *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics,* Oxford University Press, New York, 2005.

[11] Faragó I., Horváth R., *Numerikus módszerek*, 2011.

[12] Faragó I., Karátson J., *Numerical Solution of Nonlinear Elliptic Problems via Preconditioning Operators. Theory and Applications.* Advances in Computation, Vol. 11, NOVA Science Publishers, New York, 2002.

[13] Hackbusch, W., *Multigrid methods and applications*, Springer Series in Computational Mathematics 4, Springer, Berlin, 1985.

[14] Hackbusch, W., *Elliptic differential equations. Theory and numerical treatment*, Springer Series in Computational Mathematics 18, Springer, Berlin, 1992.

[15] Kadlec, J., On the regularity of the solution of the Poisson problem on a domain with boundary locally similar to the boundary of a convex open set, *Czechosl. Math. J.*, 14 (89), (1964), pp. 386-393.

[16] Karátson J., *Numerikus funkcionálanalízis*, elektronikus jegyzet, Typotex, 2013.

[17] Korotov, S., Kropac, A., Krizek, M., Strong regularity of a family of face-to-face partitions generated by the longest-edge bisection algorithm, *Zh. Vychisl. Mat. Mat. Fiz.*, 48 (2008), No. 9, 1728.

[18] Křižek, M., Neittaanmäki, P., *Mathematical and Numerical Modelling in Electrical Engineering: Theory and Applications*, Kluwer Academic Publishers, 1996.

[19] Lyness, J. N., Cools, R., A survey of numerical cubature over triangles, In: Proceedings of Symposia in Applied Mathematics, American Mathematical Society, 1994; pp. 127–150.

[20] Mao, S., Zhongci, S., Explicit error estimates for mixed and nonconforming finite elements, *J. Comp. Math.*, 27 (2009), pp. 425-440.

[21] Morton, K. W., Mayers, D. F., *Numerical Solution of Partial Differential Equations*, Cambridge, 2005.

[22] Nečas, J., *Equations aux dérivées partielles,* Presse de l'Université de Montréal, Canada, 1965.

[23] Rosser, J.B., Nine-point difference solutions for Poisson's equation, *Comp. Math. Appl.* 1 (1975), pp. 351-360.

[24] Repin, S., Sauter, S., Smolianski, A., A posteriori error estimation for the Poisson equation with mixed Dirichlet/Neumann boundary conditions, *J. Comput. Appl. Math.* 164-165 (2004), 601–612.

[25] Simon L., Baderko, E., *Másodrendű parciális differenciálegyenletek*, Tankönyvkiadó, Budapest, 1983.

[26] Stoyan G., Towards discrete velte decompositions and narrow bounds for inf-sup constants, *Comput Math Appl.* Volume 38, Issues 7–8, 1999, pp. 243–261.

[27] Stoyan G., Takó G., *Numerikus módszerek*, I–III, TypoTEX, 1997.

[28] Strikwerda, J. C., *Finite Difference Schemes and Partial Differential Equations*, SIAM, 2004.

[29] Thomas, J. W., *Numerical Partial Differential Equations: Finite Difference Methods*, Springer, 1998.

[30] Thomee, V., *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, 2006.

[31] Vlagyimirov, V. Sz., *Parciális differenciálegyenletek* (Feladatgyűjtemény), Műszaki Könyvkiadó, 1980.