Eötvös Loránd Tudományegyetem
Informatikai Kar
Numerikus Analízis Tanszék

# Diszkretizációk stabilitási tulajdonságai

Habilitációs tézisek

**Lóczi Lajos**

Budapest, 2018. április 30.

# Tartalomjegyzék

# 1. Bevezetés

Ez a dolgozat áttekinti a szerző PhD-fokozatának 2007-es megszerzése után – több esetben társszerzőkkel közösen – elért tudományos eredményeit, valamint felsorolja az oktatáshoz és ismeretterjesztéshez kapcsolódó publikációit.

A tudományos eredmények mindegyike bizonyos folytonos és – a belőlük származó – diszkrét modellek kapcsolatát vizsgálja. A folytonos modellek többnyire parciális differenciálegyenletek vagy közönséges differenciálegyenletek megoldásai, a diszkrét modellek pedig ezen egyenletek numerikus megoldásakor adódó (lineáris vagy nemlineáris) differenciaegyenletek, illetve rekurziók. E vizsgálatokban a központi kérdés általában az, hogy az eredeti folytonos rendszer megoldásainak különféle tulajdonságai hogyan öröklődnek át a diszkrét rendszerekre: ez a kérdés gyakran úgy hangzik, hogy a numerikus megoldás milyen kvalitatív-, megőrzési-, vagy stabilitási tulajdonságokkal rendelkezik.

A dolgozatban a differenciálegyenletek és dinamikai rendszerek, valamint a numerikus- és komplex analízis fogalom- és eszköztárát használjuk. Emellett fontosnak tartjuk kiemelni, hogy az elért eredményekhez döntő módon járult hozzá a *Mathematica* programcsomag [1]. A *Mathematica*-t – 1988-as debütálása óta – Stephen Wolfram és csapata folyamatosan fejleszti. E sorok írója a szoftvert 1991 óta használja aktívan: a Wolfram-nyelv és a *Mathematica* programozási környezete messzemenően alkalmasnak bizonyult arra, hogy benne interaktív módon kísérleteket és szimulációkat végezzünk sejtések kialakításához, vagy szimbolikus eszközökkel formális tételbizonyításokat hozzunk létre.

E dolgozat felépítése a következő.

Az 1.1. szakaszban néhány jelölést és definíciót idézünk fel. Az 1.2–1.3. szakaszban ismertetjük a későbbi vizsgálataink tárgyát képező klasszikus egy-, illetve többlépéses numerikus módszereket és legfontosabb tulajdonságaikat.

A 2. fejezetben tipikus problémákon keresztül mutatjuk be azt a kontextust, amelybe a [2]–[12] cikkekre épülő 3. fejezet illeszkedik. A 2. fejezet részletesebb történeti áttekintést is ad az SSP-módszerekről (*strong-stability preserving methods*). E nagyobb lélegzetvételű fejezet hosszát az indokolja, hogy az SSP-módszerekről magyar nyelven itt olvashatunk először.

A 3. fejezetben az új tudományos eredményeket vázoljuk fel (az alábbiakban a szakasz száma után zárójelben azokat a publikációkat soroljuk fel, amelyekre az illető rész épít). Ezek mindegyikében a *stabilitás* fogalmának valamely aspektusa jelenik meg.

• A 3.1. szakasz (lásd [12, 11, 10, 6]) egylépéses diszkretizációk *numerikus strukturális* stabilitását vizsgálja egy-, illetve két kodimenziós bifurkációs pontok környezetében.

• A 3.2. szakaszban (lásd [8]) bizonyos egylépéses diszkretizációk *belső* stabilitását tanulmányozzuk.

• A 3.3.1. szakasz az [5] cikket ismerteti, ahol az irodalomban először konstruáltunk *változó* lépésközű SSP-tulajdonságú többlépéses módszereket.

• A 3.3.2. szakasz (lásd [4]) egylépéses SSP-módszerek *folytonos kiterjesztésével* foglalkozik.

• A 3.3.3. szakaszban (lásd [3]) differenciálegyenletek egy bizonyos osztályában az általános SSP-elmélet eredményeit *javítjuk meg*.

• A 3.4.1. szakasz (lásd [9]), illetve a 3.4.2. szakasz (lásd [2]) rendre egylépéses, illetve többlépéses módszerek *optimális* lépésköz-együtthatóiról szól. A lépésköz-együtthatók az SSP-együtthatók rokonai.

• A 3.5.1. szakaszban (lásd [7]) egylépéses módszerek stabilitásának szemszögéből elemezzük az exponenciális függvény *Taylor-sorának részletösszegeit*.

A 4. fejezet végül a szerző oktatáshoz és ismeretterjesztéshez kapcsolódó publikációit sorolja fel.

## 1.1. Néhány jelölés és elnevezés

### 1.1.1. Bizonyítási módszerek

A későbbiekben többször hivatkozunk majd *számítógépes bizonyításokra*, melyek helyessége a *Mathematica* belső (számunkra tehát nem hozzáférhető) algoritmusainak helyességén múlik. Ezzel szemben *hagyományos bizonyításon* a „papíron" leírt és a szokásos módon olvasható bizonyításokat értjük. Érdekes módon az ebben a dolgozatban szereplő csaknem valamennyi hagyományos bizonyítás számítógépes kísérletek és módszerek által nyújtott támogatás eredményeként született.

### 1.1.2. Algebrai számok és abszolút monotonitás

• A közelítő numerikus értékek esetén tizedesvessző helyett tizedespontot használunk.

• A természetes számok $\mathbb{N}$ halmaza a nullát is tartalmazza.

• Komplex számok valós- és képzetes részét a Re és az Im szimbólumokkal jelöljük. A $z$ komplex szám konjugáltja $\bar{z}$.

• A későbbiekben felbukkanó algebrai számokat definiáló $\sum_{j=0}^{n} a_j x^j$ polinomokat (ahol $3 \leq n \in \mathbb{N}$, $a_n \neq 0$ és $a_j \in \mathbb{Z}$) a rövidség kedvéért legtöbbször az együtthatók

$$\{a_n, a_{n-1}, \ldots, a_0\} \tag{1}$$

sorozatával jelöljük.

• A $\psi$ valós függvényt az $x \in \mathbb{R}$ pontban *abszolút monotonnak* mondjuk, ha tetszőleges $k \in \mathbb{N}$ esetén a $k$-adik deriváltja létezik és nemnegatív: $\psi^{(k)}(x) \geq 0$. A $\psi$ függvény *abszolút monotonitási sugara*

$$R(\psi) := \sup\Big(\{r \in [0,+\infty) : \psi \text{ abszolút monoton } \forall x \in [-r,0] \text{ pontban}\} \cup \{0\}\Big) \in [0,+\infty], \tag{2}$$

azaz a számegyenes bal felén fekvő és origóban végződő legnagyobb olyan intervallum hossza, amelyen $\psi$ abszolút monoton. A definíció garantálja, hogy a szuprémumban szereplő halmaz sosem üres.

### 1.1.3. Vektorterek és funkcionálok

• Jelölje a továbbiakban $\mathbb{1}$ a csupa 1-eseket tartalmazó vektort. Az $\mathbb{1}$ vektor dimenziója a szövegkörnyezetből világos lesz.

• Ha $v, w \in \mathbb{R}^m$, illetve ha $A$ és $B$ azonos méretű valós mátrixok, akkor ebben a dolgozatban a $v \leq w$, illetve az $A \leq B$ relációt *komponensenként* értjük – természetesen hasonló megállapodás vonatkozik a $\geq$ relációra.

• A $v \in \mathbb{R}^m$ vektor legkisebb, illetve legnagyobb komponensét $\min v$, illetve $\max v$ jelöli.

• Legyen $\mathbb{V}$ valós vektortér. Azt mondjuk, hogy a $\|\cdot\| : \mathbb{V} \to \mathbb{R}$ funkcionál *konvex*, ha

$$\|\theta v + (1-\theta)w\| \leq \theta\|v\| + (1-\theta)\|w\| \qquad (\forall v, w \in \mathbb{V},\ \forall \theta \in [0,1]).$$

A $\|\cdot\| : \mathbb{V} \to \mathbb{R}$ funkcionál *félnorma*, ha

$$\|v + w\| \leq \|v\| + \|w\|, \quad \|\lambda v\| = |\lambda|\|v\| \qquad (\forall v, w \in \mathbb{V},\ \forall \lambda \in \mathbb{R});$$

a félnormák egyúttal nyilván konvex funkcionálok is. Például az $x = (x_j) \in \mathbb{R}^m$ vektor *teljes variációját* az

$$\|x\|_{\mathrm{TV}} := \sum_{j=1}^{m-1} |x_{j+1} - x_j| \tag{3}$$

félnormával definiáljuk. A (3) egyenlőség egy folytonos függvény teljes variációjának, más szóval teljes megváltozásának egyik lehetséges diszkrét változatát definiálja. Az $x = (x_j) \in \mathbb{R}^m$ vektor $\infty$-normája (amely nyilván félnorma is) az $\|x\|_\infty := \max_{1 \leq j \leq m} |x_j|$ szám.

### 1.1.4. Differenciálegyenletek

• A KDE, illetve a PDE rövidítés közönséges differenciálegyenletet, illetve parciális differenciálegyenletet jelent.

• Közönséges differenciálegyenlet *kezdetiérték-problémáján* az

$$u'(t) = f(t, u(t)), \quad u(t_0) = u_0 \tag{4}$$

feladatot értjük, ahol $u$ az ismeretlen függvény, $f$ valamely függvényosztályból vett adott függvény, a $(t_0, u_0)$ pár pedig a kezdeti értéket határozza meg. Mindvégig feltesszük, hogy a (4) problémának egyértelműen létezik az $u$ megoldása valamely $t \in [t_0, t_0 + T]$ intervallumon. Vektorértékű $f$ és $u$ esetén differenciálegyenlet-rendszerre vonatkozó kezdetiérték-problémát nyerünk.

• Állandó együtthatós lineáris KDE-rendszeren az

$$u'(t) = Lu(t), \quad u(t_0) = u_0 \tag{5}$$

problémát értjük, ahol $L$ adott valós elemű négyzetes mátrix, amely nem függ a $t$ változótól.

### 1.1.5. Folytonos rendszerek néhány kvalitatív tulajdonsága

• A (4) problémát *pozitívnak* mondjuk, ha tetszőleges $u_0 \geq 0$ kezdőérték és $\forall t \geq t_0$ esetén a megoldásvektorra fennáll az $u(t) \geq 0$ egyenlőtlenség. Ebben a szövegkörnyezetben a *pozitivitást* tehát gyenge értelemben használjuk, azon általában *nemnegativitást* értünk – pontosabban a nemnegativitás megőrzését.

• Jelölje $u^*$, illetve $u^{**}$ a (4) egyenlet $u_0^*$, illetve $u_0^{**}$ kezdeti értékekből induló megoldásait. A (4) problémában szereplő differenciálegyenletet a $\|\cdot\|$ normában *disszipatívnak* mondjuk, ha $\|u^*(t) - u^{**}(t)\| \leq \|u_0^* - u_0^{**}\|$ fennáll $\forall t \geq t_0$ pontban az $u_0^*$, illetve $u_0^{**}$ kezdeti értékek tetszőleges megválasztása mellett.

• A (4) probléma teljesíti a *maximum-elvet*, ha $\forall t \geq t_0$ pontban az $u(t)$ megoldásvektorra igaz, hogy $(\min u_0)\mathbb{1} \leq u(t) \leq (\max u_0)\mathbb{1}$ (vagyis $u(t)$ minden $u_k(t)$ komponensére $\min u_0 \leq u_k(t) \leq \max u_0$ teljesül). Ezt a tulajdonságot az *értékkészlet korlátosságának* (*range boundedness*) is szokás nevezni.

• Jelöljön $\|\cdot\|$ egy konvex funkcionált. Azt mondjuk, hogy a (4) probléma *monoton* a $\|\cdot\|$ funkcionálra nézve, ha $\forall t^* \geq t \geq t_0$ mellett fennáll az $\|u(t^*)\| \leq \|u(t)\|$ egyenlőtlenség.

• A fenti fogalmak között számos összefüggés fogalmazható meg (lásd [13]). Például, ha a (4) probléma teljesíti a maximum-elvet, akkor $\|u(t)\|_\infty \leq \|u_0\|_\infty$ is igaz ($\forall t \geq t_0$). Speciálisan, ha a (4) rendszer lineáris (vagyis (5) alakú) és teljesíti a maximum-elvet, akkor $u(t) = e^{L(t-t_0)}u_0$ miatt (a $\infty$-norma által indukált operátornormára) $\|e^{L(t-t_0)}\|_\infty \leq 1$ is fennáll, vagyis a rendszer *stabil* a $\infty$-normában. Bizonyos feltételek mellett a pozitivitás maga után vonja a maximum-elvet. A pozitivitás ellenőrzésére pedig egyszerű kritériumok léteznek: az (5) rendszer például pontosan akkor pozitív, ha az $L$ mátrix főátlón kívüli elemei mind nemnegatívak.

**1.1. megjegyzés.** *A [14] könyvben KDE-rendszerek és késleltetett differenciálegyenletek (delay differential equations) monotonitási tulajdonságairól, illetve a monoton dinamikai rendszerek elméletéről kapunk részletesebb áttekintést.*

### 1.1.6. Diszkretizációk néhány kvantitatív és kvalitatív tulajdonsága

• A Runge–Kutta-módszereket (lásd 1.2. szakasz) RK-módszereknek, a lineáris többlépéses módszereket (lásd 1.3. szakasz) LT-módszereknek rövidítjük.

• Az explicit Euler-módszerre – amely a numerikus módszerek prototípusa – az EE-módszer rövidítéssel hivatkozunk (lásd (10)).

• A numerikus módszerek kvantitatív elméletében többek között definiálják egy RK- vagy LT-módszer *konzisztenciáját*, *stabilitását* és *konvergenciáját*. Ezeket a fogalmakat itt nem definiáljuk, csak az alábbi heurisztikus módon érzékeltetjük. Tekintsük a (4) probléma numerikus megoldását egy állandó $h$-lépésközű módszerrel. A módszer (bizonyos rendben) *konzisztens*, ha a folytonos rendszer megoldásának egy pontjából (például a $(t_0, u_0)$ kezdeti feltételből) kiindulva a numerikus módszer által egy lépésben elkövetett hiba (vagyis a *lokális hiba*) $h \to 0^+$ esetén (megfelelő sebességgel) 0-hoz tart. A módszer *stabil*, ha a $[t_0, t_0 + T]$ intervallumon – sok lépésen keresztül – elkövetett lokális hibák nem halmozódnak túlságosan, vagyis ha a kis perturbációk hosszú távú hatása mérsékelt. A módszer *konvergens*, ha a $[t_0, t_0 + T]$ intervallumon elkövetett *globális hiba* $h \to 0^+$ esetén (megfelelő sebességgel) 0-hoz tart. A módszer konzisztenciáját többnyire egyszerű, míg konvergenciáját általában nehéz ellenőrizni. Ezért jelentősek a numerikus módszerek azon alaptételei, amelyek például a konzisztencia és stabilitás együttes fennállása esetén garantálják a módszer konvergenciáját.

A kvalitatív tulajdonságok közül az 1.1.5. szakaszban megemlített fogalmak alábbi diszkrét megfelelőit emeljük ki.

• Egy RK-módszer által generált $u_0$ kezdőértékű $u_n$ vektorsorozat teljesíti a *diszkrét pozitivitási* tulajdonságot, ha

$$u_0 \geq 0 \implies u_n \geq 0 \quad (n \in \mathbb{N}),$$

míg $k$-lépéses LT-módszer alkalmazása esetén az $u_0$, $u_1$, …, $u_{k-1}$ kezdőértékekkel generált sorozat diszkrét pozitivitási tulajdonsággal rendelkezik, ha

$$u_0 \geq 0, u_1 \geq 0, \ldots, u_{k-1} \geq 0 \implies u_n \geq 0 \quad (n \geq k).$$

• Diszkretizáljuk a (4) problémát egy RK-módszerrel. Jelölje $u_n^*$, illetve $u_n^{**}$ az $u_0^*$, illetve $u_0^{**}$ kezdőértékekből induló folytonos megoldások diszkrét approximációit. Ha a (4) probléma disszipatív, akkor természetes elvárás, hogy a diszkretizáltak teljesítsék az

$$\|u_{n+1}^* - u_{n+1}^{**}\| \leq \|u_n^* - u_n^{**}\| \quad (n \in \mathbb{N})$$

*kontraktivitási* tulajdonságot.

• Egy RK-módszer által generált $u_n$ sorozat teljesíti a *diszkrét maximum-elvet*, ha

$$(\min u_0)\mathbb{1} \leq u_n \leq (\max u_0)\mathbb{1} \quad (n \in \mathbb{N}).$$

• Egy RK-módszer által generált $u_n$ sorozat rendelkezik az *értékkészlet korlátosságának* tulajdonságával, ha adott $v$ és $w$ vektorokra

$$v \leq u_0 \leq w \implies v \leq u_n \leq w \quad (n \in \mathbb{N}).$$

• Egy RK-módszer által generált $u_n$ sorozat a $\|\cdot\|$ konvex funkcionálra nézve teljesíti a *diszkrét* (*külső*) *monotonitási* tulajdonságot (*external monotonicity*), ha

$$\|u_{n+1}\| \leq \|u_n\| \quad (n \in \mathbb{N}). \tag{6}$$

• Az előbbiekben felsorolt folytonos-diszkrét fogalompárok közötti egyik kapcsolatot mutatja be a következő állítás. Az egyszerűség kedvéért fel fogjuk tenni, hogy a (4) egyenlet egyértelműen létező $u$ megoldása értelmezve van a $[t_0, +\infty)$ intervallumon.

**1.2. definíció.** *Azt mondjuk, hogy az* $f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ *függvény teljesíti az* explicit Euler-feltételt *(EE-feltételt, forward Euler condition, FE condition) a* $\|\cdot\|$ *konvex funkcionálra nézve, ha*

$$\exists \varrho \in (0, +\infty) \quad \forall t \in \mathbb{R} \quad \forall y \in \mathbb{R}^m : \quad \|y + \varrho f(t, y)\| \leq \|y\|. \tag{7}$$

A $\|\cdot\|$ funkcionál konvexitása miatt a (7) feltételből a

$$\forall \tau \in [0, \varrho] \;\; \forall t \in \mathbb{R} \;\; \forall y \in \mathbb{R}^m : \quad \|y + \tau f(t, y)\| \leq \|y\|$$

tulajdonság is egyszerűen következik.

**1.3. állítás (lásd [15]).** *Jelöljön $\|\cdot\|$ egy félnormát és tegyük fel, hogy a* (4) *problémában szereplő $f$ függvény teljesíti a* (7) *EE-feltételt. Ekkor*

$$\forall\, t, t^* \in [t_0, +\infty),\; t^* \geq t \quad \Longrightarrow \quad \|u(t^*)\| \leq \|u(t)\|.$$

**1.4. megjegyzés.** *Az* EE-feltétel *elnevezés oka az, hogy a fenti $y + \varrho f(t, y)$ kifejezés nem más, mint a* (10)*-beli $h = \varrho$ lépésközű EE-módszer egy lépésének alkalmazása a* (4) *egyenletre. Az állítás tehát diszkrét monotonitási tulajdonságból következtet folytonos monotonitási tulajdonságra.*

**1.5. megjegyzés.** *Az itt felsorolt néhány kvalitatív tulajdonságon kívül az alkalmazások tükrében számos egyéb fogalom vizsgálata bizonyult hasznosnak. Ismert például, hogy a Hamilton-rendszerek megőrzik a fázistérfogatot, természetes kívánalom tehát, hogy ezen folytonos rendszerek diszkretizációi is rendelkezzenek ezzel a tulajdonsággal. Ebből a célból fejlesztették ki a szimplektikus RK-módszereket (*symplectic Runge–Kutta methods*, lásd [16]). Ezzel és az ehhez hasonló megőrzési tulajdonságokkal ebben a dolgozatban nem foglalkozunk.*

## 1.2. A Runge–Kutta-módszercsalád

A (4) kezdetiérték-feladatot gyakran egylépéses, de többlépcsős numerikus módszerrel oldjuk meg. Ezen módszerek fontos osztályát alkotja a Runge–Kutta-módszercsalád.

Adott RK-módszer esetén $p \in \mathbb{N}^+$ jelölje a módszer (konvergencia)rendjét, $s \in \mathbb{N}^+$ a lépcsők számát, és (állandó lépésközű módszer esetén) $h > 0$ a lépésközt. Egy RK-módszert az $(A, b)$ Butcher-táblájával szokás megadni, ahol $A = (a_{i,j})_{i,j=1}^{s}$ egy $s \times s$ méretű valós mátrix, míg $b = (b_j)_{j=1}^{s}$ egy $s$ valós komponensből álló (oszlop)vektor. Legyen $t_j := t_0 + jh$, és szokás szerint $1 \leq i \leq s$ esetén $c_i := \sum_{j=1}^{s} a_{i,j}$. Ekkor az RK-módszer egy $u_n$ $(n \in \mathbb{N})$ sorozatot határoz meg az alábbi módon:

$$u_{n+1} = u_n + h \sum_{j=1}^{s} b_j f(t_n + c_j h, y_j), \tag{8}$$

ahol $1 \leq i \leq s$ esetén

$$y_i = u_n + h \sum_{j=1}^{s} a_{i,j} f(t_n + c_j h, y_j). \tag{9}$$

Az $y_i$ mennyiség tehát $n$-től is függ, és egy (algebrai) egyenletrendszer megoldása után határozható meg. Az $u_n$ rekurzió $u_0$ kezdőértékét (4) tartalmazza. Az RK-módszert úgy tervezik, hogy $t_0 + nh \in [t_0, t_0 + T]$ esetén $u_n \approx u(t_0 + nh)$ legyen, ahol $u$ jelöli a (4) kezdetiérték-probléma pontos megoldását. Ha (4) vektorértékű, akkor természetesen az $u_n$ sorozat is vektorokból áll. Az $y_i$ értékeket egyes szerzők *belső lépcsőknek*, mások *belső approximációnak* (*internal approximation*) nevezik, míg az $u_n$ értékeket *külső approximációnak* (*external approximation*) hívják.

Az RK-módszer *explicit* (ERK), ha $A$ szigorú alsóháromszög-mátrix, egyébként a módszer *implicit* (IRK). Egy implicit RK-módszer diagonálisan implicit (DIRK), ha az $A$ mátrix főátlója fölötti háromszögben csak nullák állnak. Egy DIRK-módszer egyszeresen diagonálisan implicit (SDIRK), ha $A$ főátlójában minden elem ugyanaz a nemnulla konstans.

Általában minél nagyobb a módszer $p$ rendje, annál jobban approximálja a módszer a differenciálegyenlet megoldását, viszont annál nehezebb az $(A, b)$ Butcher-táblát megkonstruálni. Adott $p$ és $s$ esetén az $(A, b)$ Butcher-tábla meghatározásához sokismeretlenes polinomiális egyenletrendszereket kell megoldani. Ezeket az egyenletrendszereket kombinatorikus módszerekkel és fagráfokkal lehet kényelmesen leírni, lásd például a [17] könyvet.

Jelölje ERK$(p, s)$ a $p$-edrendű, $s$-lépcsős ERK-módszereket. Ismert, hogy ERK-módszerek esetén $s \geq p$, és $p \in \{1, 2, 3, 4\}$ esetén $s = p$ is elérhető, ám ha $p \geq 5$, akkor $s > p$. Az alábbiakban $p \leq 4$ esetén felsorolunk néhány ERK-módszert, amelyekre a későbbiekben hivatkozni fogunk, és amelyek az előző mondat értelmében adott rend esetén minimális lépcsőszámú módszerek.

• Az ERK$(1, 1)$ halmaz egyetlen eleme az

$$u_{n+1} = u_n + hf(t_n, u_n) \tag{10}$$

explicit Euler-módszer.

• Az ERK$(2, 2)$ halmazhoz tartozó módszerek az

$$A = \begin{pmatrix} 0 & 0 \\ \alpha & 0 \end{pmatrix}, \quad b^\top = \left(1 - \tfrac{1}{2\alpha}, \tfrac{1}{2\alpha}\right) \tag{11}$$

egyparaméteres családdal írhatóak le, ahol $\alpha \in \mathbb{R} \setminus \{0\}$.

• Az ERK$(3, 3)$ halmazt három diszjunkt család alkotja: egy kétparaméteres család, és két, egyparaméteres család. A kétparaméteres család az

$$A = \begin{pmatrix} 0 & 0 & 0 \\ \alpha & 0 & 0 \\ \beta - \frac{(\alpha-\beta)\beta}{\alpha(3\alpha-2)} & \frac{(\alpha-\beta)\beta}{\alpha(3\alpha-2)} & 0 \end{pmatrix}, \quad b^\top = \left(\frac{6\alpha\beta - 3\alpha - 3\beta + 2}{6\alpha\beta}, \frac{2 - 3\beta}{6\alpha(\alpha - \beta)}, \frac{3\alpha - 2}{6\beta(\alpha - \beta)}\right) \tag{12}$$

alakú Butcher-táblákkal reprezentálható, ahol $\alpha, \beta \in \mathbb{R} \setminus \{0\}$ és $\frac{2}{3} \neq \alpha \neq \beta$.

• Az ERK$(4, 4)$ halmaz elemeit itt nem írjuk le részletesen, csak megjegyezzük, hogy a klasszikus negyedrendű RK-módszer[1], ahol

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad b^\top = (1/6, 1/3, 1/3, 1/6)\,, \tag{13}$$

is ide tartozik.

Az RK-módszerek más szempontok szerint is csoportosíthatók: beszélhetünk például reducibilis vagy irreducibilis módszerekről. (Valójában több különböző (ir)reducibilitási fogalom létezik, lásd [18, 19]; a pontos definíciókra itt nem lesz szükségünk.) A gyakorlatban csak az *irreducibilis módszerek* érdekesek; reducibilis módszerek alkalmazása esetén ugyanis végezhetünk például olyan részszámításokat, melyeket nem is használunk fel (következésképp a módszer ekvivalens módon átírható egy kisebb méretű RK-módszerré).

A numerikus módszerek – 1.1.6. szakaszban említett – stabilitása aszimptotikus fogalom, amely a módszer viselkedéséről a $h \to 0^+$ határesetben mond ki valamit; a gyakorlatban azonban fontos tudni, hogy pontosan mekkora is ez az „elegendően kicsiny" $h > 0$ érték. Ennek kvantitatív jellemzésére vezették be a RK-módszerek abszolút stabilitásának fogalmát. Ezzel kapcsolatban az alábbiakban emlékeztetünk egy RK-módszer *stabilitási függvényére* és *abszolút stabilitási tartományára*.

[1] Ezt a módszert gyakran jelölik az RK44 szimbólummal, és az 1900-as évek eleje óta ismert.

Legyen $\lambda \in \mathbb{C}$ tetszőleges konstans. Ha az $(A, b)$ Butcher-táblázattal megadott RK-módszert alkalmazzuk a (lineáris és skaláris)

$$u'(t) = \lambda u(t) \tag{14}$$

tesztegyenletre, akkor jól ismert, hogy a $z := h\lambda$ választással (8)–(9)-ből egy

$$u_{n+1} = \psi(z) u_n \tag{15}$$

alakú rekurzió adódik, ahol $\psi \equiv \psi_{A,b}$ a módszerhez tartozó (lineáris) stabilitási függvény.

**1.6. megjegyzés.** *Az egyszerű szerkezetű tesztegyenletben szereplő $\lambda$ konstansra állandó együtthatós lineáris differenciálegyenlet-rendszer esetén gondolhatunk úgy, mint az együtthatómátrix egy tipikus sajátértékére; nemlineáris rendszer esetén pedig a linearizálás után kapott Jacobi-mátrix valamelyik sajátértékére. Általában azt feltételezzük, hogy ha megértjük a módszer viselkedését stabilitás szempontjából a tesztegyenleten, akkor abból következtetni tudunk a stabilitásra bonyolultabb szituációkban is.*

ERK-módszerek esetén a $\psi$ stabilitási függvény polinom, míg IRK-módszerek esetén racionális törtfüggvény. Általánosan igaz, hogy

$$\psi(z) = \frac{\det(I - zA + z\mathbb{1}b^\top)}{\det(I - zA)}, \tag{16}$$

ahol $\mathbb{1}b^\top$ diadikus szorzat és $I \in \mathbb{R}^{s\times s}$ az egységmátrix.

Az RK-módszer $\mathcal{S}$ abszolút stabilitási tartománya a komplex számok azon részhalmaza, melyre a stabilitási függvény definiálva van és legfeljebb 1 abszolút értékű, azaz

$$\mathcal{S} := \{z \in \mathbb{C} : \exists\, (I - zA)^{-1}, |\psi(z)| \le 1\}. \tag{17}$$

**1.7. megjegyzés.** *Az abszolút stabilitási tartomány fenti definícióját az alábbi egyszerű észrevétel motiválja. Ha valamely rögzített $\lambda \in \mathbb{C}$ esetén a (14) egyenlet megoldását egy $h > 0$ lépésközű RK-módszerrel közelítjük és e lépésköz úgy van megválasztva, hogy*

$$z = h\lambda \in \mathcal{S}$$

*fennáll, akkor a (15) rekurzió által generált $u_n$ sorozat tetszőleges $u_0$ kezdőérték esetén korlátos marad.*

## 1.3. Lineáris többlépéses módszerek

Az egylépéses RK-módszerek alternatívájaként a (4) kezdetiérték-probléma megoldására gyakran használunk lineáris többlépéses módszereket. Az LT-módszerek[2] részletes leírása megtalálható például a [20, 21] monográfiákban.

Legyen a *lépések száma* $2 \le k \in \mathbb{N}$ rögzített. Ekkor az $\alpha_j$ $(j = 1, \dots, k)$ és $\beta_j$ $(j = 0, \dots, k)$ valós együtthatók által meghatározott (állandó) $h > 0$ lépésközű $k$-lépéses LT-módszeren $n \in \mathbb{N}$ esetén az

$$u_{n+k} = \sum_{j=1}^{k} \alpha_j u_{n+k-j} + h \sum_{j=0}^{k} \beta_j f_{n+k-j} \tag{18}$$

alakú $k$-adrendű differenciaegyenletet értjük, ahol $f_m$ $(m \in \mathbb{N})$ az $f(t_0 + mh, u_m)$ kifejezés rövidítése és $f$ a (4) egyenletben adott függvény. Az LT-módszer *explicit*, ha $\beta_0 = 0$; különben a módszer *implicit*. Azért, hogy a differenciaegyenlet rendje pontosan $k$ legyen, feltesszük, hogy $\alpha_k^2 + \beta_k^2 > 0$.

[2] Az első LT-módszereket az 1880-as évek közepe táján használták. Modern elméletüket G. Dahlquist (1925–2005) dolgozta ki az 1950-es években.

**1.8. megjegyzés.** *Ahhoz, hogy egy $k$-lépéses módszert elindíthassunk, szükségünk van a (megfelelően pontos) $u_0, u_1, \dots, u_{k-1}$ kezdőértékekre. Ezt a $k$ darab értéket a gyakorlatban sokszor egy alkalmas RK-módszerrel állítjuk elő.*

Az RK-módszerekhez hasonlóan itt is igaz, hogy az $u_n$ numerikus megoldás $t_0 + nh \in [t_0, t_0 + T]$ esetén az $u(t_0 + nh)$ pontos megoldás egy közelítése. Az RK-módszerekhez hasonlóan az LT-módszerek esetén is beszélhetünk a módszer $p \in \mathbb{N}^+$ rendjéről. Az RK-módszerekkel ellentétben azonban, az LT-módszerek rendfeltételei sokkal egyszerűbb szerkezetűek.

Az RK-módszerekhez hasonlóan egy LT-módszernek is definiálható az *abszolút stabilitási tartománya.* Ehhez tekintsük a módszer $\alpha_j$ és $\beta_j$ együtthatóiból képzett

$$\varrho(\zeta) := \zeta^k - \sum_{j=1}^{k} \alpha_j \zeta^{k-j} \quad \text{és} \quad \sigma(\zeta) := \sum_{j=0}^{k} \beta_j \zeta^{k-j}$$

polinomokat. Az LT-módszer karakterisztikus polinomja a $\Phi(\zeta, \mu) := \varrho(\zeta) - \mu\sigma(\zeta)$ $(\zeta, \mu \in \mathbb{C})$ kétváltozós polinom. Azt mondjuk, hogy egy egyváltozós (legalább elsőfokú, komplex együtthatós) $P$ polinom *teljesíti a gyökfeltételt*, ha $P$ minden $\zeta_j$ gyökére igaz, hogy $|\zeta_j| \le 1$, és ha egy $\zeta_m$ gyök többszörös, akkor $|\zeta_m| < 1$ is fennáll. Ezek után az LT-módszer $\mathcal{S}$ abszolút stabilitási tartománya a komplex számok alábbi részhalmaza:

$$\mathcal{S} := \big\{\mu \in \mathbb{C} : 1 - \mu\beta_0 \neq 0 \text{ és a } \Phi(\cdot, \mu) \text{ egyváltozós polinom teljesíti a gyökfeltételt}\big\}. \tag{19}$$

**1.9. megjegyzés.** *Az 1.2. szakaszban említett 1.7. megjegyzés analogonja LT-módszerek esetén is igaz: rögzített $\lambda \in \mathbb{C}$ esetén a (14) egyenlet megoldását egy $h > 0$ lépésközű LT-módszerrel közelítve, a (18) rekurzió által generált $u_n$ sorozat tetszőleges $u_0, u_1, \dots, u_{k-1}$ kezdőérték esetén korlátos marad, ha $h\lambda \in \mathcal{S}$.*

**1.10. megjegyzés.** *Az $\mathcal{S}$ tartomány (19)-beli definíciója némileg szigorúbb az LT-módszerek abszolút stabilitási tartományának szokásos definíciójánál. A korábbi irodalmi hivatkozásokban (lásd például [20, 21]) az $1 - \mu\beta_0 \neq 0$ feltétel általában nincs megkövetelve. Könnyen látható, hogy ez a feltétel éppen azt jelenti, hogy a $\Phi(\zeta, \mu)$ polinom $\zeta$ változó szerinti főegyütthatója nem tűnik el. Célszerű azonban feltenni, hogy e főegyüttható nem nulla (lásd például [2, Remark 1.3], illetve [22, 23]). Az ellenkező esetben ugyanis például a (14) tesztegyenletre alkalmazott (18) rekurzió rendje $k$-nál szigorúan kisebb (vagyis bizonyos kivételes $h\lambda$ pontokban az LT-módszer nem $k$-lépéses módszer), s így az $u_0, u_1, \dots, u_{k-1}$ kezdőértékek sem választhatók meg általában tetszőlegesen.*

A későbbiekben előforduló LT-módszerek közül most csak egyetlen implicit módszercsaládot emelünk ki: a $k$-lépéses BDF-módszer képlete

$$\sum_{j=1}^{k} \frac{1}{j} \nabla^j u_{n+1} = h f_{n+1}. \tag{20}$$

Itt a $\nabla$ differenciaoperátort a $\nabla u_{n+1} := u_{n+1} - u_n$ és $j > 1$ esetén a $\nabla^j u_{n+1} := \nabla^{j-1} u_{n+1} - \nabla^{j-1} u_n$ képletek definiálják. A 2-lépéses $h > 0$ lépésközű BDF-módszer például

$$u_{n+2} - \frac{4}{3} u_{n+1} + \frac{1}{3} u_n = \frac{2}{3} h f(t_{n+2}, u_{n+2})$$

alakban is felírható, ahol $t_{n+2} := t_0 + (n+2)h$.

**1.11. megjegyzés.** *Az angol* backward differentiation formula *(BDF) kifejezést magyarul a retrográd differenciák módszerének is nevezik. A gyakorlatban csak a $k \le 6$ lépéses BDF-módszereket használják, ugyanis $k \ge 7$ esetén e család rekurziói nem 0-stabilak (a 0-stabilitás a módszer konvergenciájának egy szükséges feltétele).*

**1.12. megjegyzés.** *Az $u_{n+1} = u_n + h f(t_n + h, u_{n+1})$ implicit Euler-módszer (amely egyúttal az egyik legegyszerűbb implicit RK-módszer is) a (20) formula $k = 1$ speciális eseteként adódik.*

## 2. Néhány tipikus kérdés

Ebben a fejezetben néhány olyan problémát és gondolatot ismertetünk az elmúlt pár évtizedből, amelyek a 3. fejezet eredményeihez teremtik meg a kontextust.

### 2.1. Megmaradási elvek: hiperbolikus parciális differenciálegyenletek

Számos fontos fizikai jelenség írható le megmaradási elvek segítségével, melyeket matematikailag gyakran hiperbolikus parciális differenciálegyenletekkel modelleznek. Az ilyen egyenletek numerikus megoldási módszereit foglalja össze a [24] mű, míg a [13] könyv általánosabb keretek között tárgyalja parciális differenciálegyenletek diszkretizációit.

Valamely (megfelelő osztályból vett) $F : \mathbb{R} \to \mathbb{R}$ függvény esetén tekintsük például a

$$\partial_t U(x,t) + \partial_x (F(U(x,t))) = 0 \tag{21}$$

nemlineáris, elsőrendű, egydimenziós skaláris hiperbolikus egyenletet, ahol $U : \mathbb{R} \times [0, +\infty) \to \mathbb{R}$ jelöli az ismeretlen függvényt valamilyen $U(x,0) = U_0(x)$ kezdeti feltétel mellett; az $U$ megoldást általában nem lehet analitikus formulával előállítani. A (21) egyenlet speciális eseteként például a jól ismert Burgers-egyenlet alábbi változata adódik:

$$\partial_t U(x,t) + U(x,t)\, \partial_x U(x,t) = 0 \quad (x \in \mathbb{R},\, t \geq 0). \tag{22}$$

Ez a kvázilineáris PDE többek között összenyomható gázok matematikai modelljeként szolgál. Ismert (lásd [24]), hogy akár sima $U_0$ kezdeti feltétel esetén is szakadás (más szóval lökéshullám, *shock wave*) alakulhat ki a megoldásban véges idő alatt. A lökéshullámok léte egyrészt komplikálttá teszi az ilyen típusú egyenletek matematikai elméletét (a megoldás egzisztenciájának és unicitásának kérdése problematikussá válik), másrészt megnehezíti a lökéshullámot megfelelő módon reprodukáló numerikus módszerek kifejlesztését.

**2.1. megjegyzés.** *A felmerülő elméleti nehézségek közül az alábbiakat emeljük ki. A szakadási pontban a derivált klasszikus módon nem értelmezhető. Ennek orvoslására bevezették a* gyenge megoldás *fogalmát. Sajnos azonban gyenge megoldásból több is lehet (még azonos kezdeti feltétel esetén is). Ezek közül a fizikailag releváns gyenge megoldást az egyenlethez csatolt további feltételekkel (például* entrópiafeltétellel*) választják ki.*

Az $U$ függvényt numerikusan gyakran az *egyenesek módszerével* (*method of lines*, MOL) approximálják, amelynek során szétválasztják a térbeli és az időbeli diszkretizációt. A MOL-módszer első fázisában térben diszkretizálunk, a $\partial_x$ operátort például egy végesdifferencia-operátorral helyettesítve. A $\Delta x > 0$ térbeli diszkretizációs lépésközzel dolgozva a folytonos $x$ változó helyett egy diszkrét $\{x_j\}$ rács adódik, ám a $t \geq 0$ változó továbbra is folytonos marad. Az ily módon nyert szemidiszkrét rendszer nem más, mint közönséges differenciálegyenletek kezdetiérték-problémáinak egy (4) alakú rendszere. Fontos azonban kiemelni, hogy most a (4) egyenletrendszer jobb oldalán álló $f$ függvény *szinguláris módon függhet* a térbeli diszkretizációs paramétertől: tipikusan például az egyenletrendszer mérete végtelenhez tart, amint $\Delta x \to 0^+$. A MOL-módszer második fázisában – valamely konkrét $\Delta x$ által generált $\{x_j\}$ rács és tetszőlegesen rögzített $x = x_j$ esetén – a (4) problémát időben, azaz $t$ szerint is diszkretizáljuk egy $\Delta t > 0$ idődiszkretizációs lépésközzel, alkalmazva például a korábban említett RK vagy LT módszerek valamelyikét ($\Delta t$ szerepét az ottani jelölésekkel $h$ játssza).

**2.2. megjegyzés.** *A MOL (egyenesek módszere) kifejezés valójában tehát (i) nem konkrét módszert jelent, hanem egy megközelítési módot, ahogyan diszkretizációk készíthetők és vizsgálhatók, illetve (ii)* egyenesek *helyett* $\{(x_j, t) \in \mathbb{R}^2 : t \geq 0\}$ *alakú félegyenesek szerepelnek benne.*

*A parciális differenciálegyenletek a térbeli és időbeli diszkretizációk szétválasztása nélkül, direkt módon is diszkretizálhatók (például a Lax–Wendroff-sémával, amely a MOL-megközelítés keretein kívülre esik). Az*

*ilyen típusú* teljes diszkretizációk *stabilitási- vagy konvergenciatulajdonságait azonban elméleti szempontból nehezebb vizsgálni.*

A lökéshullámok jelenléte miatt egyáltalán nem nyilvánvaló, hogy konkrétan milyen térbeli és időbeli diszkretizációkat érdemes alkalmazni. Ha a diszkretizációk rendje alacsony, akkor egyrészt a numerikus módszer kevésbé hatékony, másrészt a séma felbontása a lökéshullám környékén rossz lehet, és a szakadás „elkenődik". Ha a diszkretizációk rendje magas, akkor viszont gyakran tapasztalni a lökéshullám környékén a numerikus megoldásban oszcillációkat, amelyek a folytonos modell megoldásában nincsenek jelen.

**2.3. megjegyzés.** *Hasonló jelenséget látunk a Fourier-sorok klasszikus elméletében: szakadásos függvény Fourier-sorának részletösszegei a szakadási pont környékén szükségszerűen „túlhullámzást" mutatnak, melyet Gibbs–Wilbraham-jelenségnek neveznek (a részletösszegek pontonként konvergálnak ugyan, de egyenletes konvergencia nincsen).*

A numerikus megoldások ezen oszcillációi azért sem kívánatosak, mert az eredeti PDE megoldásának egyes komponensei gyakran a sűrűséget vagy nyomást jelentik: nem szerencsés tehát, ha nemnegatív fizikai mennyiségek numerikus approximációjakor negatív értékek is fellépnek.

Az elmúlt néhány évtizedben kifejlesztett, és a fenti szempontokat figyelembe vevő térbeli diszkretizációk közé tartoznak például

- a TVD-módszerek (*total-variation-diminishing methods*): A. Harten[3] (1983);
- a TVB-módszerek (*total-variation-bounded methods*);
- az ENO-módszerek (*essentially-non-oscillatory methods*): A. Harten, B. Engquist, S. Osher, S. Chakravarthy (1987);
- a harmadrendű WENO-módszerek (*weighted-essentially-non-oscillatory methods*): X.-D. Liu, S. Osher, T. Chan (1994);
- az ötöd- és magasabb rendű WENO-módszerek: G. Jiang, C.-W. Shu (1996).

Ezeknek a módszereknek a haszna az alábbiakban rejlik. Bizonyos általános feltételek mellett igaz, hogy egydimenziós skaláris megmaradási törvények (tetszőleges gyenge) $U$ megoldása TVD-tulajdonságú, azaz a megoldás teljes variációja időben nem nőhet:

$$TV(U(\cdot,t^*)) \le TV(U(\cdot,t)) \qquad (\forall\, t^* \ge t \ge 0).$$

A TVD-módszerekkel a fenti tulajdonság a szemidiszkrét rendszerre is átvihető: az $\{x_j\}$ diszkrét rácson nyert $\{u_j(t)\}$ közelítéseket jelölje $\widetilde{u}(t)$. Itt $u_j(t)$ például a PDE megoldásának $U(x_j,t)$ értékét, vagy a megoldás $\frac{1}{\Delta x}\int_{x_j-\Delta x/2}^{x_j+\Delta x/2} U(x,t)dx$ cellaátlagát közelíti. Ha most az $\widetilde{u}(t)$ közelítés diszkrét teljes variációját a

$$TV(\widetilde{u}(t)) \equiv \|\widetilde{u}(t)\|_{\mathrm{TV}} = \sum_j |u_{j+1}(t) - u_j(t)|$$

formulával definiáljuk (lásd (3)), akkor $\forall\, t^* \ge t \ge 0$ esetén fennáll, hogy

$$TV(\widetilde{u}(t^*)) \le TV(\widetilde{u}(t)), \text{ vagyis } \|\widetilde{u}(t^*)\|_{\mathrm{TV}} \le \|\widetilde{u}(t)\|_{\mathrm{TV}}.$$

**2.4. megjegyzés.** *A 2.1. megjegyzés folytatásaként megemlítjük, hogy P. D. Lax és B. Wendroff (1960) egyik eredménye szerint, ha egy* konzervatív *tulajdonságú szemidiszkretizáció konvergál, akkor az a PDE egyik gyenge megoldásához fog tartani; A. Harten, J. M. Hyman és P. D. Lax (1976) eredménye pedig kimondja, hogy a TVD-tulajdonság bizonyos feltételek mellett elégséges ahhoz, hogy a szemidiszkretizáció az entrópiafeltételt kielégítő, fizikailag releváns gyenge megoldáshoz konvergáljon.*

[3] Amiram Harten doktori témavezetője 1971 és 1974 között Lax Péter (P. D. Lax, 1926–) volt New Yorkban

*Néha a TVD-tulajdonság túl erős megkötés: a TVB-módszerek olyan numerikus közelítéseket adnak, melyek teljes változása korlátos (ha az $U_0$ kezdeti feltételben szereplő függvény is ilyen tulajdonságú). A TVD-módszerek hátránya, hogy elméletileg korlátozott a velük elérhető térbeli approximáció rendje. Ennek kiküszöbölésére alkották meg a (térben) magasabb rendű WENO-módszereket. A WENO-módszerek hátránya viszont, hogy nincs elméleti garancia arra, hogy ne növelhetnék meg a diszkrét teljes variációt (noha a WENO-módszereket úgy tervezték, hogy az oszcillációk a „lehető legkevésbé" nőjenek).*

Továbbmenve, a MOL-módszer második fázisában az $\widetilde{u}(t)$ közelítésben szereplő $u_j(t)$ függvényeket egy $\Delta t$ lépésközű EE-módszerrel időben is diszkretizálva az eredeti PDE megoldásának teljes diszkretizáltjához jutunk, melynek $t = t_n$ időréteghez tartozó értékét jelölje $\widetilde{u}_n$. A TVD-módszer EE-módszerrel való ötvözésének lényege az, hogy amennyiben a $\Delta t$ lépésköz értékét a TVD-módszer alapján alkalmasan választjuk meg, akkor *a teljes diszkretizációra is* igaz marad, hogy a diszkrét teljes variáció nem nőhet:

$$\|\widetilde{u}_{n+1}\|_{\mathrm{TV}} \leq \|\widetilde{u}_n\|_{\mathrm{TV}} \qquad (n \in \mathbb{N}). \tag{23}$$

A fent vázolt TVD+EE teljes diszkretizációs eljárás (még magasrendű térbeli approximáció esetén is) időben csak $p = 1$ elsőrendű közelítést ad az EE-módszer miatt. Hogyan lehet olyan TVD-tulajdonságú teljes diszkretizációkat kifejleszteni, amelyek térben is és időben is magas rendben approximálnak? Erre a kérdésre térünk vissza a 2.3. szakaszban.

## 2.2. Közönséges differenciálegyenletek diszkretizációinak nemnegativitása, kontraktivitása, monotonitása

Az előző, 2.1. szakaszban bizonyos típusú parciális differenciálegyenletek vizsgálatakor felmerülő kérdéseket mutattunk be. A jelen 2.2. szakaszban felsorolt problémákat és eredményeket közönséges differenciálegyenletek kvalitatív tulajdonságainak tanulmányozása motiválta – a numerikus módszerek klasszikus hibaanalízise ugyanis nem ad kielégítő választ ezekre a kérdésekre.

• C. Bolley, M. Crouzeix (1978) és M. N. Spijker (1983) *lineáris* KDE-rendszerek pozitivitási, illetve kontraktivitási tulajdonságait vizsgálja. Megállapítják, hogy ha ilyen tulajdonságú folytonos rendszereket *legalább másodrendű* (azaz $p \geq 2$) RK- vagy LT-módszerrel (időben) diszkretizálunk, akkor a pozitivitási, illetve kontraktivitási tulajdonság nem öröklődhet át a diszkrét rendszerre

$$\text{a } h > 0 \text{ lépésköz nagyságára tett megszorítás} \tag{24}$$

nélkül.

• A (24) feltétellel a továbbiakban még sokszor fogunk találkozni, világítsuk meg ezért konkrétabban, ám az egyszerűség kedvéért pozitivitás vagy kontraktivitás helyett monotonitással megfogalmazva. A (7) definícióhoz hasonlóan azt mondjuk, hogy az $L$ valós négyzetes mátrix teljesíti az *EE-feltételt*, ha

$$\exists\, h_{\mathrm{EE}} \in (0, +\infty) \text{ hogy } \forall h \in (0, h_{\mathrm{EE}}] \text{ és } \forall v \in \mathbb{V} \text{ esetén } \|v + hLv\| \leq \|v\|, \tag{25}$$

ahol $\|\cdot\|$ egy adott konvex funkcionál. Az *EE-feltétel* kifejezés helyett *lineáris* esetben az angol szakirodalomban gyakran a *körfeltétel* (*circle condition*) kifejezéssel találkozunk, mert a (25) feltétel (vagy annak variánsai) kapcsolatba hozható(k) azzal, hogy az $L$ mátrix sajátértékei a komplex síkon egy (alkalmas $\delta > 0$ sugarú) $\{z \in \mathbb{C} : |z + \delta| \leq \delta\}$ körben helyezkednek el. Jelölje végül az RK-módszer (16) stabilitási függvényének (2) abszolút monotonitási sugarát $R(\psi_{A,b})$, vagy röviden csak $R(\psi)$; e mennyiség neve az irodalomban gyakran *threshold factor*. Ez után az előkészítés után kimondható az alábbi állítás.

**2.5. állítás (lásd [25]).** *Tegyük fel, hogy az* (5) *lineáris rendszerben szereplő $L$ mátrix teljesíti a* (25) *EE-feltételt, és jelölje $u_n$ az* (5) *rendszer egy $h$-lépésközű konzisztens RK-módszerrel való diszkretizációjának eredményét, ahol a $h$ lépésköz a*

$$0 < h \leq R(\psi) h_{EE} \tag{26}$$

*egyenlőtlenség alapján van megválasztva (ilyen típusú feltételre utal* (24)*). Ekkor fennáll az*

$$\|u_{n+1}\| \leq \|u_n\| \qquad (n \in \mathbb{N}) \tag{27}$$

*monotonitási tulajdonság (lásd* (6)*).*

A tétel lényege tehát az, hogy a (26) feltétel kapcsolatot teremt egy $h$-lépésközű (magasabb rendű) RK-módszer monotonitási tulajdonsága és a $h_{\mathrm{EE}}$-lépésközű EE-módszer (mint legegyszerűbb, elsőrendű RK-módszer) monotonitási tulajdonsága között. Kicsit másképp megfogalmazva, a (26) egyenlőtlenség jobb oldalának két faktora közül az első, $R(\psi)$, csak az aktuális RK-diszkretizációtól függ, míg a második, (25) szerint, csak a differenciálegyenlet (vagyis az $L$ mátrix) tulajdonságától. Megjegyezzük, hogy a (26) egyenlőtlenség optimális, vagyis az $R(\psi)$ faktor a lehető legnagyobb olyan tényező, amely mellett az állítás adott RK-módszer esetén általánosan igaz – gyakorlati szempontból pedig világos, hogy nagyobb $h$-lépésközű monoton RK-módszer hatékonyabb diszkretizációt jelent.

• J. A. van de Griend és J. F. B. M. Kraaijevanger (1986, [26]) lineáris KDE-rendszerek megoldásának approximációjára optimális (azaz maximális) lépésközű ERK-, illetve IRK-módszereket konstruál. Fontos azonban kiemelni, hogy míg az előző bekezdésben említett optimalitási tulajdonság egy rögzített RK-módszerre vonatkozott, addig az itteni optimalitási eredmény RK-módszerek (valamilyen rendű vagy lépcsőszámú) osztályáról szól.

• J. Sand (1986) és H. W. J. Lenferink (1989, 1991) kontraktivitás szempontjából optimális (állandó) lépésközű explicit, illetve implicit LT-módszerek osztályát vizsgálja lineáris vagy nemlineáris közönséges differenciálegyenletek megoldására.

• J. F. B. M. Kraaijevanger (1991, [27]) elegáns módon kiterjeszti a lineáris KDE-rendszerek megoldására szolgáló kontraktív RK-módszerek elméletét disszipatív *nemlineáris* KDE-rendszerekre. A (26) feltétellel analóg egyenlőtlenséget fogalmaz meg, melyben az $R(\psi)$ mennyiség (vagyis az RK-módszer stabilitási függvénye abszolút monotonitási sugarának) szerepét egy $R(A,b)$-vel jelölt mennyiség veszi át. Az $R(A,b) \in [0,+\infty]$ konstans neve *az RK-módszer abszolút monotonitási sugara*, vagy röviden *Kraaijevanger-együtthatója.*

• Horváth Zoltán (1998, [28]) olyan RK-módszerek lépésközére ad becslést, amelyek megőrzik nemlineáris KDE-rendszerek pozitivitási tulajdonságait. A lépésközre tett megszorításban megjelenik az RK-módszer abszolút monotonitási sugara, $R(A,b)$.

**2.6. megjegyzés.** *Az 1.1.5. szakasz 1.1. megjegyzésében említett monoton KDE-rendszerek és monoton késleltetett differenciálegyenletek RK-módszerrel történő monotonitásőrző diszkretizációiról olvashatunk például a [29, 30] cikkben, ahol azonban a lépésközre tett megszorítások nincsenek expliciten kifejezve az RK-módszer Butcher-táblájával.*

## 2.3. SSP-módszerek

### 2.3.1. Az SSP-módszerek bevezetése

A 2.1. és 2.2. szakaszban felvázolt problémákat sok éven keresztül egymástól függetlenül vizsgálták, és a két terület módszerei és eredményei nem voltak egymásra hatással. Az alábbiakból kiderül, hogy ezek a kérdések közös elmélettel kezelhetők. Az így adódó diszkretizációkat nevezzük ma *SSP-módszereknek.* Az angol szakirodalomban az SSP (*strong-stability preserving*, vagy *strong stability preserving*, néha *strong stability-preserving*) tulajdonsággal rendelkező módszereket magyarul hívhatjuk *az erős stabilitást megőrző* (vagy *erősebb értelemben stabilitásőrző*) módszereknek, de az egyszerűség kedvéért mi is megtartjuk az angol rövidítést. Az SSP kifejezést egyébként S. Gottlieb, C.-W. Shu és E. Tadmor (2001) használta először. Az SSP-tulajdonságú idődiszkretizációs módszerekre gyakran érdemes úgy gondolni, mint amelyek *nemlineáris stabilitási tulajdonsággal*, vagy másképpen mondva, *általános monotonitási tulajdonsággal* rendelkeznek. RK-módszerek és LT-módszerek szisztematikus vizsgálatával az SSP-tulajdonság

szemszögéből a [25] monográfiában olvashatunk. Az alábbiakban röviden felsoroljuk az SSP-módszerek fejlődésének néhány lényeges mozzanatát.

• Az RK-módszerek (8)–(9) klasszikus alakja helyett C.-W. Shu és S. Osher (1988) egy adott RK-módszert *EE-lépések konvex kombinációjaként* reprezentál (melyet ma Shu–Osher-alaknak nevezünk), s így sikerül megalkotni az első, térben és időben is magasabb rendű TVD-tulajdonságú diszkretizációkat.

• S. Gottlieb és C.-W. Shu (1998) azonosítja az optimális SSP RK-módszereket az ERK(2,2) és ERK(3,3) osztályban. Ebben a kontextusban egy SSP-módszert optimálisnak mondunk, ha *SSP-együtthatója* az adott osztályban a legnagyobb. Az SSP-együtthatót a 2.3.3. szakaszban definiáljuk majd, és értéke a $(0, +\infty]$ intervallumba eshet. A nagyobb SSP-együtthatójú numerikus módszerek hatékonyabbak, ugyanis gyorsabban tud haladni a nemlineáris stabilitási tulajdonságokat megtartó időbeli diszkretizáció. (Néhány szerző más szóhasználatot követve megengedi, hogy az SSP-együttható értéke 0 is lehessen: mindenesetre a 0 SSP-együtthatójú módszerek a gyakorlat szempontjából érdektelenek, célszerű ezért azt mondani, hogy egy módszer pontosan akkor SSP-tulajdonságú, ha SSP-együtthatója a $(0, +\infty]$ intervallum eleme.)

• S. Gottlieb, C.-W. Shu és E. Tadmor (2001) SSP-tulajdonságú IRK-módszereket, valamint explicit, illetve implicit SSP LT-módszereket konstruál.

• R. J. Spiteri és S. J. Ruuth (2002) bebizonyítja, hogy az SSP-módszerek maximális rendje korlátozott: pozitív SSP-együtthatójú irreducibilis ERK-módszer rendje legfeljebb $p \leq 4$, illetve pozitív SSP-együtthatójú IRK-módszer rendje legfeljebb $p \leq 6$ lehet. A további kutatások során kiderül, hogy az SSP-elmélet módosításával (*downwinding* alkalmazásával) ezek a korlátok átléphetők.

• A Shu–Osher-elmélet (pontosabban az RK-módszerek *optimális* Shu–Osher-reprezentációja) és az RK-módszerek abszolút monotonitási sugarára vonatkozó elmélet közötti szoros kapcsolatot 2004–2005-ben csaknem egyszerre fedezi fel L. Ferracina és M. N. Spijker, illetve I. Higueras. A két terület egyesítésével megszülető új elmélet hatékonyabb és teljesebb mindkét elődjénél.

• E felfedezés egyik következményeként egyszerűbben tudjuk megfogalmazni az optimális SSP-együtthatójú RK-módszerek keresésére szolgáló algoritmusokat.

• A 2004–2005-ös felfedezés másik következménye, hogy az SSP-elmélet keretein belül egységesen kezelhetővé válnak például az 1.1.5., 1.1.6., illetve a 2.2. szakaszban felsorolt fogalmak, ha az eddigiekben szereplő $\|\cdot\|$ konvex funkcionált alkalmas módon választjuk meg (lásd [31]): egységesen tárgyalható a diszkrét pozitivitás, a diszkrét maximum-elv, a diszkrét értékkészlet korlátossága, a diszkrét kontraktivitás, vagy a (23)-ban, illetve (27)-ben szereplő diszkrét monotonitási tulajdonság.

• Az elmúlt 10–12 évben az SSP-elméletet kiterjesztették például additív RK-módszerekre, illetve (az RK- és LT-módszereket is magukban foglaló) *általános lineáris módszerek* (*general linear methods*) osztályára (lásd [32, 33]).

### 2.3.2. Példa SSP RK-módszerre

Az SSP-módszerek formális definíciójának megfogalmazása előtt bemutatunk egy konkrét háromlépcsős, másodrendű, SSP-tulajdonságú explicit Runge–Kutta-módszert.

Tekintsük például az

$$u'(t) = f(t, u(t)) := \sin(10t)\, u(t)\, (1 - u(t)) \tag{28a}$$

$$u(0) = u_0 \tag{28b}$$

kezdetiérték-feladatot, valamely adott $u_0 \in \mathcal{I}_1 := [0, 1]$ kezdőértékkel. Belátható, hogy a (28) KDE egyértelmű megoldása az

$$u(t) = u_0 \left( u_0 + (1 - u_0) \exp\left( \frac{\cos(10t) - 1}{10} \right) \right)^{-1} \tag{29}$$

függvény, melyre igaz, hogy

$$\forall u_0 \in \mathcal{I}_1,\ \forall t \geq 0 \quad \Longrightarrow \quad u(t) \in \mathcal{I}_1,$$

vagyis az $\mathcal{I}_1$ intervallumból indított folytonos megoldás mindvégig $\mathcal{I}_1$-ben marad.

Diszkretizáljuk most a (28) problémát a (10)-beli elsőrendű, $h$-lépésközű

$$u_{n+1} = u_n + hf(t_n, u_n) \tag{30}$$

EE-módszerrel, azaz most $t_n = nh$. Legyen $h_{\mathrm{EE}} := 1$, és rögzítsünk egy $u_0 \in \mathcal{I}_1$ kezdőértéket, valamint egy $h \in (0, h_{\mathrm{EE}}]$ lépésközt. Ekkor teljes indukcióval a nyilvánvaló

$$0 \leq u_n \left(1 + h\left(1 - u_n\right) \sin(10nh)\right) = u_n + hf(nh, u_n) =$$

$$1 + \left(1 - u_n\right)\left(hu_n \sin(10nh) - 1\right) \leq 1$$

egyenlőtlenségekből azt kapjuk, hogy

$$\forall u_0 \in \mathcal{I}_1,\ \forall n \in \mathbb{N} \quad \Longrightarrow \quad u_n \in \mathcal{I}_1, \tag{31}$$

vagyis az $\mathcal{I}_1$ intervallumból indított diszkrét megoldás mindvégig $\mathcal{I}_1$-ben marad. Az is egyszerűen megmutatható, hogy ha a $h$ lépésközt az $(1, 2)$ nyílt intervallumból alkalmasan választjuk, akkor a (31) invarianciatulajdonság már megsérülhet: $h = 101/100$ és $u_0 = 9861/10000 \in \mathcal{I}_1$ esetén például $u_8 > 50001/50000$, azaz $u_8 \notin \mathcal{I}_1$.

Az elsőrendű EE-módszer helyett alkalmazzuk most a (28)-as egyenletre az

$$y_1 = u_n \tag{32a}$$

$$y_2 = y_1 + \frac{h}{2} f(t_n, y_1) \tag{32b}$$

$$y_3 = y_2 + \frac{h}{2} f(t_n + h/2, y_2) \tag{32c}$$

$$u_{n+1} = \frac{1}{3} u_n + \frac{2}{3}\left(y_3 + \frac{h}{2} f(t_n + h, y_3)\right) \tag{32d}$$

képlettel adott $s = 3$ lépcsős, $p = 2$ másodrendű SSP Runge–Kutta-módszert, ahol természetesen ismét $t_n = nh$. Legyen $\mathcal{C} := 2$. Ekkor belátható, hogy tetszőleges $h \in (0, \mathcal{C} \cdot h_{\mathrm{EE}}] = (0, 2]$ esetén a (32) RK-módszer által generált $u_n$ sorozatra fennáll a (31) invarianciatulajdonság.

**2.7. megjegyzés.** *A* (32)*-ben megadott RK-módszer nem a* (8)–(9) *Butcher-alakban, hanem egy (vele ekvivalens) Shu–Osher-alakban van felírva, mert így az invarianciatulajdonság bizonyítása könnyebb.*

A fentiekben szereplő mindkét diszkretizáció ((30) és (32)) megőrzi tehát a folytonos megoldás nemnegativitását (sőt, az értékkészletének korlátosságát), ha a numerikus módszer lépésközét alkalmasan megszorítjuk (vö. (24)). Azt is figyeljük meg, hogy a (32) módszer lépésközére vonatkozó $0 < h \leq \mathcal{C}\, h_{\mathrm{EE}}$ korlát pont olyan szerkezetű, mint a (26)-ban szereplő: a lineáris esetben fellépő $R(\psi)$ konstans szerepét a nemlineáris esetben a $\mathcal{C}$-vel jelölt SSP-együttható veszi át. Másrészt a két numerikus módszer közül (32) hatékonyabb, mint (30), hiszen

• (32) rendje magasabb, mint az EE-módszer rendje, így kis $h > 0$ lépésköz mellett (32) jobban approximálja a (29)-beli megoldást;

• (32) nagyobb diszkretizációs lépésköz mellett ($0 < h \leq 2$) is rendelkezik az SSP-tulajdonsággal, vagyis időben gyorsabban haladhat az erős stabilitási tulajdonsággal bíró diszkretizáció.

### 2.3.3. Az SSP RK-módszerek definíciója

Tegyük fel, hogy a (4) problémában szereplő $f$ vektormező kielégíti az 1.1.6. szakasz 1.2. definíciójában szereplő (7) EE-feltételt egy $h_{\mathrm{EE}}$ konstanssal és adott $\|\cdot\|$ konvex funkcionállal, vagyis fennáll, hogy

$$\exists h_{\mathrm{EE}} \in (0,+\infty) \;\; \forall t \in \mathbb{R} \;\; \forall y \in \mathbb{R}^m : \quad \|y + h_{\mathrm{EE}}\, f(t,y)\| \leq \|y\|. \tag{33}$$

Rögzítsünk egy $(A,b)$ Butcher-táblázattal megadott RK-módszert. Valamely $h>0$ lépésköz mellett jelölje az RK-módszer által generált sorozatot $u_n$ (amely sorozat tehát az $(f,t_0,u_0,A,b,h)$ adatoktól függ). Azt mondjuk, hogy ez az $(A,b)$ RK-módszer *SSP-tulajdonságú*, ha létezik olyan $c>0$ konstans, hogy

$$\forall\, h \in (0, c\cdot h_{\mathrm{EE}}] \text{ és } \forall\, n \in \mathbb{N} \text{ esetén } \|u_{n+1}\| \leq \|u_n\|.$$

A legnagyobb ilyen tulajdonságú $c>0$ számot az RK-módszer *SSP-együtthatójának* nevezzük, és $\mathcal{C}$-vel jelöljük.

**2.8. megjegyzés.** *Az SSP-együttható csak az $(A,b)$ RK-módszertől függ, és nem függ $f$-től, sem az $f$ értelmezési tartományában szereplő vektortér $m$ dimenziójától, sem a konvex funkcionáltól, sem az $u_0$ kezdőértéktől. Az alkalmazások szempontjából fontos, hogy az SSP-együttható* nem *függ például $m$-től: gondoljunk a 2.1. szakaszban említett szituációra, amikor $f$ a $\Delta x>0$ térbeli diszkretizációs paramétertől függ úgy, hogy $\Delta x \to 0^+$ esetén $m \to +\infty$.*

Tegyük fel most, hogy az RK-módszer SSP-tulajdonságú és teljesít bizonyos irreducibilitási feltételeket. Ekkor az RK-módszer Shu–Osher-alakjából az SSP-együttható kiszámítható. Az $(\alpha,\beta)$-val jelölt Shu–Osher-alakot itt nem definiáljuk; az $\alpha$, $\beta$ szimbólumok mátrixok, melyek elemeit jelölje rendre $\alpha_{i,j} \in \mathbb{R}$ és $\beta_{i,j} \in \mathbb{R}$. Egy adott $(A,b)$ Butcher-táblázatú RK-módszerhez általában végtelen sok $(\alpha,\beta)$ Shu–Osher-alak tartozik – ez felel meg annak a korábban említett ötletnek, hogy egy RK-módszer EE-lépések *konvex* kombinációjaként is reprezentálható, s emiatt tesszük fel, hogy $\|\cdot\|$ *konvex* funkcionál. Az RK-módszer SSP-együtthatóját egy

$$\mathcal{C} = \sup_{\alpha,\beta} \min_{i,j} \frac{\alpha_{i,j}}{\beta_{i,j}} \tag{34}$$

alakú kifejezés adja, amennyiben a (34) jobb oldalán álló szuprémum a $(0,+\infty]$ intervallumba esik. (A szuprémum egyébként mindig a $[0,+\infty]$ intervallum eleme, de ha értéke 0, akkor az RK-módszer az iménti 2.8. megjegyzést megelőző definíció értelmében nem SSP-tulajdonságú.) A (34)-es formula értelmezése a következő. A szuprémumot a lehetséges Shu–Osher-alakokra vesszük, és adott $(\alpha,\beta)$ mátrixú Shu–Osher-alak mellett az alábbi két kiegészítő megállapodással élünk:

• a $\min_{i,j} \frac{\alpha_{i,j}}{\beta_{i,j}}$ kifejezés definíció szerint 0, ha az $\alpha_{i,j}$, $\beta_{i,j}$ számok között van legalább egy negatív érték, valamint

• ha minden $\alpha_{i,j}$, $\beta_{i,j}$ érték nemnegatív, akkor egy $(i,j)$ indexpárhoz tartozó $\frac{\alpha_{i,j}}{\beta_{i,j}}$ tört értéke $\beta_{i,j}=0$ esetén definíció szerint $+\infty$.

A 2.3.1. szakaszban szereplő *optimális Shu–Osher reprezentáció* a (34) formulára utalt, és a szintén ott említett 2004–2005-ös felfedezés lényege az, hogy – továbbra is feltéve az $(A,b)$ RK-módszer irreducibilitását – fennáll a

$$\mathcal{C} = R(A,b)$$

egyenlőség, ha $R(A,b) \in (0,+\infty]$. Itt az RK-módszer abszolút monotonitási sugarát (azaz Kraaijevanger-együtthatóját) az

$$R(A,b) := \sup\{r \in \mathbb{R} : \forall \varrho \in [0,r]\ \exists (I+\varrho K)^{-1}, \varrho K(I+\varrho K)^{-1} \geq 0, \varrho K(I+\varrho K)^{-1}\mathbb{1} \leq \mathbb{1}\} \tag{35}$$

formulával számolhatjuk ki, ahol

$$K := \begin{pmatrix} A & 0 \\ b^\top & 0 \end{pmatrix}.$$

Figyeljük meg, hogy a definíció miatt $R(A,b) \in [0,+\infty]$ mindig igaz. Ha $R(A,b)=0$, akkor az RK-módszer nem SSP-tulajdonságú.

**2.9. megjegyzés.** *Ha $\|\cdot\|$ félnorma, akkor az 1.3. állítás szerint a* (33) *EE-feltételből a* (4) *probléma $\|\cdot\|$ funkcionálra vett monotonitása is következik.*

**2.10. megjegyzés.** *Az SSP-tulajdonság lényege tehát a következő.* **Ha** *a* (4) *differenciálegyenlet $h_{EE}$ lépésközű EE-módszerrel nyert diszkretizációjakor adódó sorozat teljesíti a diszkrét monotonitási tulajdonságot,* **akkor** *a diszkrét monotonitási tulajdonság a $h \in (0, \mathcal{C}\, h_{EE}]$ lépésközű RK-módszer által generált sorozatra is* átöröklődik.

*A 2.1. szakasz kontextusában a* (21) *PDE térbeli szemidiszkretizációjának tervezésekor az egyik cél éppen az, hogy az ott adódó $f$ függvény teljesítse az EE-feltételt. Ekkor pozitív (és minél nagyobb) SSP-együtthatójú RK-módszer választásával magasabb rendű, hatékony és stabil idődiszkretizációhoz, s ezáltal teljes diszkretizációhoz jutunk.*

**2.11. megjegyzés.** *Az SSP-elmélet sokrétűségét az adja, hogy a $\|\cdot\|$ konvex funkcionál alkalmas megválasztásával sok általános tulajdonság írható át olyan alakúra, mint amilyen az $\|u_{n+1}\| \le \|u_n\|$ diszkrét monotonitási tulajdonságban szerepel.*

**2.12. megjegyzés.** *Ahogyan az 1.1.6. szakaszban láttuk, a stabilitás egyik szokásos megfogalmazása az, hogy a rekurzió során menet közben elkövetett hibák csak mérsékelten halmozódnak fel. Ez gyakran egy*

$$\|u_n\| \le \mu \|u_0\| \tag{36}$$

*alakú egyenlőtlenséggel fejezhető ki, valamilyen $\mu \ge 1$ ($n$-től független) állandóval. Az SSP-módszereket azért nevezték el* erős stabilitást *megőrző módszereknek, mert az $\|u_{n+1}\| \le \|u_n\|$ ($n \in \mathbb{N}$) diszkrét monotonitási tulajdonság maga után vonja a* (36) *egyenlőtlenséget a $\mu = 1$ választással.*

### 2.3.4. Az SSP LT-módszerek definíciója

Az előző pontban a technikai nehézségek főleg abból eredtek, hogy egy adott RK-módszert sokféle alakban lehet reprezentálni; ilyen problémák az LT-módszerek esetében nem lépnek fel.

Tegyük fel ismét, hogy a (4) problémában szereplő $f$ vektormező kielégíti a (33) EE-feltételt egy $h_{\mathrm{EE}}$ konstanssal és adott $\|\cdot\|$ konvex funkcionállal. Tekintsünk egy $k$-lépéses LT-módszert ($k \ge 2$), amelyet az $\alpha_j, \beta_j$ együtthatók határoznak meg. Az LT-módszerről e részben mindvégig feltesszük, hogy konzisztens, 0-stabil és irreducibilis – a gyakorlatban használt módszerek mind teljesítik ezeket a természetes követelményeket. Valamely $h > 0$ lépésköz mellett jelölje az $u_0, u_1, \ldots, u_{k-1}$ kezdőértékekből indított és az LT-módszer által generált sorozatot $u_n$. Azt mondjuk, hogy ez az LT-módszer *SSP-tulajdonságú*, ha létezik olyan $c > 0$ konstans, hogy

$$\forall\, h \in (0, c \cdot h_{\mathrm{EE}}] \text{ és } \forall\, n \in \mathbb{N},\ n \ge k \text{ esetén } \|u_n\| \le \max_{0 \le j \le k-1} \|u_j\|. \tag{37}$$

A legnagyobb ilyen tulajdonságú $c$ számot az LT-módszer *SSP-együtthatójának* nevezzük, és $\mathcal{C}$-vel jelöljük.

**2.13. megjegyzés.** *Itt is fontos hangsúlyozni, hogy az SSP-együttható csak az LT-módszertől függ, és nem függ például a* (4) *problémától, sem a numerikus módszer $u_0, u_1, \ldots, u_{k-1}$ kezdőértékeitől.*

Egy LT-módszer pontosan akkor SSP-tulajdonságú, ha

- minden $1 \le j \le k$ esetén $\alpha_j \ge 0$, és (38a)
- minden $0 \le j \le k$ esetén $\beta_j \ge 0$, és (38b)
- ha valamely $1 \le j \le k$ esetén $\beta_j > 0$, akkor $\alpha_j > 0$. (38c)

SSP-tulajdonságú LT-módszer SSP-együtthatóját a

$$\mathcal{C} = \min_{1 \le j \le k} \frac{\alpha_j}{\beta_j} \tag{39}$$

formula alapján számíthatjuk ki: hasonlóan az RK-módszerek esetéhez, itt is megállapodunk abban, hogy $\alpha_j \ge 0$ és $\beta_j = 0$ esetén $\frac{\alpha_j}{\beta_j} := +\infty$.

## 2.4. RK- és LT-módszerek lépésköz-együtthatói

Egy RK vagy LT numerikus módszer esetében az SSP-tulajdonság nagyon erős megkötés, hiszen ez a tulajdonság garantálja, hogy minden, az EE-módszerre nézve diszkrét monotonitási tulajdonsággal bíró $f$ függvény, minden, elegendően kicsiny $h > 0$ lépésköz, és minden kezdőérték esetén a diszkrét monotonitási tulajdonság átöröklődik a numerikus módszerre. Emiatt gyakran azt tapasztaljuk, hogy a gyakorlatban egyébként hasznos módszerek nem SSP-tulajdonságúak, vagy ha igen, akkor az SSP-együtthatójuk túl kicsi pozitív szám.

Számos matematikus, többek között I. Higueras, W. Hundsdorfer[4], A. Mozartova, S. J. Ruuth, M. N. Spijker és R. J. Spiteri munkája alapján (lásd például [15], [22, 23], [32]–[38]) világossá vált, hogy érdemes az SSP-tulajdonságban szereplő szigorú követelményeken enyhíteni. Így az SSP-elmélet különféle kiterjesztéseit kapjuk, melyekben az SSP-együttható szerepét különféle lépésköz-együtthatók veszik át (vö. (24)). Egy adott numerikus módszerhez tartozó legnagyobb lépésköz-együtthatót ebben a szakaszban a $\gamma_{\sup} \in (0, +\infty]$ szimbólummal jelöljük. A *lépésköz-együttható* elnevezés oka, illetve a megfogalmazott tételek szerkezete a következő.

*Rögzítsünk egy $\mathcal{F}$ függvényosztályt. Tegyük fel, hogy a* (4) *problémában szereplő $f \in \mathcal{F}$ függvény teljesíti a* (33) *EE-feltételt egy $h_{\mathrm{EE}}$ konstanssal valamely $\|\cdot\|$ funkcionálra nézve. Ekkor $\forall\, h \in (0, \gamma_{\sup} \cdot h_{\mathrm{EE}}]$ lépésköz és $\forall\, n$ index esetén fennáll egy $\mathcal{E}$ egyenlőtlenség az $\mathcal{N}$ osztályból vett numerikus módszer által generált $u_n$ sorozatra.*

Az $(\mathcal{F}, \|\cdot\|, \mathcal{E}, \mathcal{N})$ objektumok megválasztásától függően például az alábbi $\gamma_{\sup}$ lépésköz-együtthatókat nyerjük (melyek közül bizonyosak korábban már szerepeltek ebben a dolgozatban, míg más együtthatók definíciója és tulajdonságai csak az utóbbi pár évben kristályosodott ki). Hangsúlyozzuk, hogy az alábbi felsorolás az áttekinthetőség kedvéért egyszerűsítéseket tartalmaz – a kiegészítő technikai feltevések (például a különféle irreducibilitási feltételek) említésétől eltekintünk.

• RK-módszer **lineáris monotonitási** lépésköz-együtthatója (*stepsize coefficient for linear monotonicity*, *threshold factor*, *contractivity radius*):
az $\mathcal{F} = \{$lineáris operátorok$\}$ (vagyis az általános (4) helyett az (5) alakú rendszereket tekintve), $\|\cdot\| =$ konvex funkcionál, $\mathcal{E} : \|u_{n+1}\| \le \|u_n\|$, $\mathcal{N} = \{$RK-módszerek$\}$ esetben $\gamma_{\sup} = R(\psi)$ (lásd (2), (16), (26), illetve [25]).

• RK-módszer **általános külső monotonitási** lépésköz-együtthatója (*stepsize coefficient for external monotonicity*, Kraaijevanger-együttható, SSP-együttható, TVD-tulajdonság):
az $\mathcal{F} = \{$lineáris vagy nemlineáris függvények$\}$ (vagyis az általános (4) alakú rendszereket tekintve), $\|\cdot\| =$ félnorma, $\mathcal{E} : \|u_{n+1}\| \le \|u_n\|$, $\mathcal{N} = \{$RK-módszerek$\}$ esetben $\gamma_{\sup} = R(A, b)$ (lásd (35), illetve [19]).

• RK-módszer **általános belső monotonitási** lépésköz-együtthatója (*stepsize coefficient for internal monotonicity*):
az $\mathcal{F} = \{$lineáris vagy nemlineáris függvények$\}$, $\|\cdot\| =$ félnorma, $\mathcal{E} : \|y_i\| \le \|u_n\|$, $\mathcal{N} = \{$RK-módszerek$\}$ esetben $\gamma_{\sup}$ egy olyan kifejezés, amely a (35) Kraaijevanger-együtthatóéval „rokon” definícióval rendelkezik (lásd (9), illetve [19]).

• RK-módszer **általános külső korlátossági** lépésköz-együtthatója (*stepsize coefficient for external boundedness*, TVB-tulajdonság):
az $\mathcal{F} = \{$lineáris vagy nemlineáris függvények$\}$, $\|\cdot\| =$ félnorma, $\mathcal{E} : \|u_n\| \le \mu\|u_0\|$, $\mathcal{N} = \{$RK-módszerek$\}$ esetben $\gamma_{\sup} = R(A, b)$ (lásd 2.12. megjegyzés, (35), illetve [19]), ahol $\mu \ge 1$ egy konstans, amely csak az RK-módszer Butcher-táblájától függ (tehát nem függ $n$-től, $f \in \mathcal{F}$-től, vagy a $\|\cdot\|$ félnormától).

• RK-módszer **általános belső korlátossági** lépésköz-együtthatója (*stepsize coefficient for internal boundedness*):
az $\mathcal{F} = \{$lineáris vagy nemlineáris függvények$\}$, $\|\cdot\| =$ félnorma, $\mathcal{E} : \|y_i\| \le \mu\|u_0\|$, $\mathcal{N} = \{$RK-módszerek$\}$

[4]Willem Hundsdorfer, a numerikus analízis kiváló kutatója, matematikusi erejének teljében 63 éves korában, 2017. november 10-én elhunyt

esetben $\gamma_{\sup} = R(A,b)$ (lásd (35), illetve [19]), ahol $\mu \geq 1$ egy konstans, amely csak az RK-módszer Butcher-táblájától függ.

• RK-módszer külső- és belső monotonitási lépésköz-együtthatóját [15] vizsgálja abban az $\mathcal{F}$ függvényosztályban, amikor a (4)-beli $f(t,u(t))$ kifejezés helyett $g(t,u(t))Lu(t)$ alakú „**pszeudolineáris**" függvények szerepelnek (korlátos, nemnegatív, skalárértékű $g$ függvényekkel és konstans $L$ mátrixokkal).

• LT-módszer **lineáris monotonitási** lépésköz-együtthatója (*stepsize coefficient for linear monotonicity*):
az $\mathcal{F} = \{\text{lineáris operátorok}\}$, $\|\cdot\| = $ félnorma, $\mathcal{E} : \|u_n\| \leq \max_{0\leq j\leq k-1} \|u_j\|$, $\mathcal{N} = \{\text{LT-módszerek}\}$ esettel kapcsolatban [23]-ben találunk hivatkozásokat.

• LT-módszer **lineáris korlátossági** lépésköz-együtthatója (*stepsize coefficient for linear boundedness*):
az $\mathcal{F} = \{\text{lineáris operátorok}\}$, $\|\cdot\| = $ félnorma, $\mathcal{E} : \|u_n\| \leq \mu \max_{0\leq j\leq k-1} \|u_j\|$, $\mathcal{N} = \{\text{LT-módszerek}\}$ eset tárgyalása [23]-ben szerepel. Itt $\mu \geq 1$ egy konstans, amely csak az LT-módszer együtthatóitól függ.

• LT-módszer **általános monotonitási** lépésköz-együtthatója (*stepsize coefficient for general monotonicity*, *SSP-coefficient*, TVD-tulajdonság):
az $\mathcal{F} = \{\text{lineáris vagy nemlineáris függvények}\}$, $\|\cdot\| = $ félnorma, $\mathcal{E} : \|u_n\| \leq \max_{0\leq j\leq k-1} \|u_j\|$, $\mathcal{N} = \{\text{LT-módszerek}\}$ esetben $\gamma_{\sup} = \mathcal{C}$ (lásd (39)).

• LT-módszer **általános korlátossági** lépésköz-együtthatója (*stepsize coefficient for general boundedness*, TVB-tulajdonság):
az $\mathcal{F} = \{\text{lineáris vagy nemlineáris függvények}\}$, $\|\cdot\| = $ félnorma, $\mathcal{E} : \|u_n\| \leq \mu \max_{0\leq j\leq k-1} \|u_j\|$ (alkalmas $\mu \geq 1$ konstanssal), $\mathcal{N} = \{\text{LT-módszerek}\}$ esetben a $\gamma_{\sup}$ együtthatót [22] vizsgálja.

A lépésköz-együtthatók nagysága a kvalitatív tulajdonságokat megőrző numerikus módszerek hatékonyságával kapcsolatos. Általában nem nyilvánvaló feladat annak eldöntése, hogy egy *adott módszernek* létezik-e (nemnulla) lépésköz-együtthatója, illetve ha igen, akkor mekkora a maximális lépésköz-együttható. Szintén érdekes kérdés annak a vizsgálata, hogy módszerek *valamely osztályában* (például a 3-lépcsős, 3-adrendű ERK-módszerek osztályában) mekkora lehet a legnagyobb lépésköz-együttható értéke.

A fent említett lépésköz-együtthatókon kívül természetesen más $(\mathcal{F}, \|\cdot\|, \mathcal{E}, \mathcal{N})$ adatokhoz tartozó lépésköz-együtthatók is definiálhatók volnának, melyek vizsgálata általában egy-egy új kutatási témát jelöl ki.

# 3. A tudományos eredmények bemutatása (2007–2017)

## 3.1. Numerikus strukturális stabilitás bifurkációs pontok környezetében

A numerikus dinamika (*numerical dynamics*) elméletében dinamikai rendszerek és diszkretizáltjaik kvantitatív és kvalitatív tulajdonságait hasonlítjuk össze (lásd például A. M. Stuart és A. R. Humphries 1998-as monográfiáját).

A (4) egyenlet speciális eseteként tekintsünk egy $u'(t) = f(u(t))$ alakú *autonóm* közönséges differenciálegyenletet, és rögzítsünk egy kicsiny $h > 0$ paramétert. Ha $u(t; x_0)$ jelöli a differenciálegyenlet $x_0$ pontból kiinduló megoldását $t$ idő elteltével, akkor a $\Phi(h, x) := u(h; x)$ formula egy $\Phi(h, \cdot) : \mathbb{R}^m \to \mathbb{R}^m$ leképezést értelmez, amelyet a differenciálegyenlet *h-idejű megoldóoperátorának* (*time-h map*) nevezünk. Ennek segítségével a fázistérben a lehetséges kezdeti értékekből induló összes megoldás hosszú távú viselkedése egyszerre vizsgálható. Másrészt tekintsük a KDE valamely egylépéses, $h$-lépésközű és $p$-edrendű $\varphi$ diszkretizációját ($p \in \mathbb{N}^+$), amelyik szintén $\varphi(h, \cdot) : \mathbb{R}^m \to \mathbb{R}^m$ alakú leképezés – gondoljunk itt például egy RK-módszerre. Ekkor a $\Phi(h, \cdot)$ és a $\varphi(h, \cdot)$ leképezések iteráltjai egy-egy diszkrét idejű dinamikai rendszert határoznak meg: a $\varphi$ által generált dinamikai rendszer a $\Phi$ által generált rendszer perturbáltja, és kis $h > 0$ lépésköz esetén e két dinamikai rendszer „közel" van egymáshoz. Természetes kérdés, hogy az eredeti fáziskép mely tulajdonságai és hogyan őrződnek meg a diszkretizáció során; vagy megfordítva, a rendelkezésre álló diszkretizált megoldásokból az (ismeretlen) eredeti megoldás mely tulajdonságaira következtethetünk. Azt mondjuk, hogy a $\Phi$ által meghatározott dinamikai rendszer *numerikusan strukturálisan stabil*, ha a rendszer „ekvivalens" egy tetszőleges, hozzá „közeli" $\varphi$ diszkretizáció által meghatározott dinamikai rendszerrel. Nemegyensúlyi helyzet környezetében, illetve általános – például különféle hiperbolicitási – feltételek mellett a numerikus strukturális stabilitás kérdését az 1990-es években tisztázták (lásd például W.-J. Beyn, Garay Barnabás vagy M. C. Li dolgozatait).

Felmerül a kérdés, hogy mi a helyzet numerikus strukturális stabilitás szempontjából akkor, ha a hiperbolicitási feltétel nem teljesül. Tekintsük például közönséges differenciálegyenletek egy olyan $u'(t) = f(u(t), \alpha)$ egyparaméteres családját, amelyben a hiperbolicitás az $\alpha$ paraméter valamely értékénél (mondjuk $\alpha = 0$-nál) megsérül, vagyis *bifurkáció* lép fel. Általában nem igaz, hogy egy bifurkációs pontban a $\Phi$ és $\varphi$ leképezések által meghatározott fázisképek „ugyanolyanok" lennének (amint azt például egy síkbeli centrum EE-módszerrel történő diszkretizációja egyszerűen mutatja). Bizonyos bifurkációs pontok környezetében azonban a fázisképek szerkezete megőrződhet: Farkas Gyula[5], illetve M. C. Li a 2000-es évek elején megmutatta, hogy nyeregcsomó-bifurkációs pontok körül a fázisképek topológiailag azonosak, azaz *konjugáltak*. A konjugációkra ebben a kontextusban nemlineáris koordináta-transzformációkként gondolhatunk.

A [12] cikkben a fenti, nyeregcsomó-bifurkációs pontra vonatkozó eredményt terjesztjük ki: noha $m = 1$ dimenzióban dolgozunk, a két szinguláris paraméter, $h \to 0^+$ és $\alpha \to 0$ jelenléte miatt a kvantitatív becslések végigszámolása nehezebbé válik. Nyeregcsomó-bifurkáció esetén az $\alpha \leq 0$ esetben jelen lévő két egyensúlyi helyzet $\alpha = 0$-nál összeolvad, míg $\alpha > 0$ esetén nincs egyensúlyi pont. Jelölje $J(h, \cdot, \alpha)$ a két leképezés, $\Phi(h, \cdot, \alpha)$ és $\varphi(h, \cdot, \alpha)$ között megkonstruált konjugáló homeomorfizmusok kétparaméteres családját. Ekkor az $\alpha \leq 0$ esetben igazoljuk, hogy egy alkalmas pozitív *const* konstanssal minden, elegendően kicsiny $h > 0$, $|x|$ és $|\alpha|$ mellett fennáll a $h$-ban optimális $|J(h, x, \alpha) - x| \leq const \cdot h^p$ közelségi becslés. Az $\alpha > 0$ esetben $\alpha$-ban szinguláris közelségi becslést bizonyítottunk: továbbra is nyitott kérdés, hogy vajon konstruálható-e olyan konjugáció, melyre a $J$ leképezés és az identitásfüggvény távolsága $\mathcal{O}(h^p)$ nagyságrendű – mindenesetre a konkrétan megadott konstrukció esetén a *Mathematica* segítségével elvégzett szimulációk kimutatták a $|J(h, x, \alpha) - x|$ kifejezés szinguláris voltát az $\alpha \to 0^+$ határesetben.

A [6] cikkben transzkritikus- (*transcritical*) és villa-bifurkációs (*pitchfork*) pontok közelében szintén konjugációk megkonstruálásával bizonyítjuk be, hogy $m = 1$ dimenzióban egylépéses diszkretizációk alkalmasan megszorított családjában (amely tartalmazza például az összes RK-módszert) az eredeti és a diszkretizált fáziskép topológiailag azonos, valamint a konjugáció és az identitásfüggvény között optimá-

[5] Farkas Gyula (1972–2002) fiatalon elhunyt kitűnő matematikus, doktori témavezetője Garay Barnabás

lis $\mathcal{O}(h^p)$ közelségi becslések állnak fenn. A közelségi becslésekhez kétparaméteres nemlineáris rekurziók konvergenciasebességének vizsgálatára van szükség: itt kulcsszerephez jutnak a T. Hüls által korábban megkonstruált paraméteres modellfüggvények, melyekhez tartozó nemlineáris rekurziók *zárt alakban* felírhatók. Ebből a *Mathematica* szimbolikus erejének segítségével lehetett $(i)$ következtetni az általános esetben fellépő rekurziók konvergenciasebességét leíró formulák szerkezetére a paraméterek függvényében, illetve $(ii)$ a formulák alakjának megsejtése után az aktuális bizonyításokat elvégezni.

A [11] dolgozat nem tartalmaz konjugációs eredményeket, viszont általános, $m$-dimenziós rendszerek elemzésével foglalkozik, igazolva az RK-módszerek alábbi *megőrzési tulajdonságait*. Megmutatjuk, hogy folytonos idejű rendszerek nyeregcsomó-, csúcs- (*cusp*), illetve Bogdanov–Takens-bifurkációs pontjai RK-diszkretizációk során a nekik megfelelő diszkrét bifurkációs pontokba mennek át, továbbá képleteket adunk a diszkretizált normálformák együtthatóira, illetve az általánosított sajátértékekre.

A [10] cikk, részben a fenti eredményekre építve, általános betekintést nyújt a folytonos dinamikai rendszerek egy-, illetve két kodimenziós bifurkációs pontok körüli egylépéses diszkretizációinak elméletébe, a hangsúlyt a numerikus strukturális stabilitás és a megőrzési tulajdonságok vizsgálatára helyezve.

### 3.2. Runge–Kutta-módszerek belső hibáinak terjedése

RK-módszerek alkalmazásakor a belső lépcsők meghatározására szolgáló (9) egyenletrendszert általában nem lehet egzaktul megoldani: a gyakorlatban $y_i$ helyett ezen mennyiségek $\widetilde{y}_i$ perturbáltjai állnak csak rendelkezésre. Az ilyen hibák forrását illetően gondoljunk például arra, hogy

• nemlineáris $f$ függvény és implicit RK-módszer esetén az $y_i$ mennyiségeket valamely iterációs módszerrel (mondjuk Newton-módszerrel) véges sok lépésben közelítjük;

• bizonyos SSP RK-módszerek, extrapolációs RK-módszerek, illetve Runge–Kutta–Csebisev-módszerek esetén a lépcsők $s$ száma nagy lehet, így (akár explicit) módszer esetén is felerősödhet a kerekítési hibák hatása;

• a mögöttes PDE szemidiszkretizációjából származó KDE-rendszer eleve hibákkal terhelt, amelyek a térbeli diszkretizáció hibáiból vagy a peremfeltételek diszkretizálásából erednek.

Az ily módon keletkező és felhalmozódó hibák analízisét, vagyis az RK-módszerek *belső stabilitását* (*internal stability*) elemezzük a [8] cikkben. A publikált változatban terjedelmi korlátok miatt nem közölt számítások és becslések a [39] linken érhetők el.

A cikkben ERK-módszerek analízisével foglalkozunk. A vizsgálatok során kiderül, hogy a belső hibák terjedése függ az RK-módszer konkrét implementációjától – emlékezzünk arra, hogy egy konkrét Butcher-táblával meghatározott RK-módszerhez végtelen sok Shu–Osher-reprezentáció tartozhat. A reprezentációtól függő hibaterjedés leírásához tetszőleges $(\alpha, \beta)$ Shu–Osher-alak esetén képletet adunk a *belső stabilitási polinomokra*. Ezen $Q_j$ polinomok hasonló szerepet töltenek be, mint a (14) egyenletre alkalmazott ERK-módszer (16)-beli $\psi$ stabilitási polinomja: a klasszikus stabilitási polinommal az egymás utáni *lépések* hibái, míg a belső stabilitási polinomokkal az egymás utáni *lépcsők* hibái mérhetők. Egy $s$-lépcsős ERK-módszer $Q_j$ $(1 \le j \le s)$ belső stabilitási polinomjai esetén az

$$\mathcal{M} \equiv \mathcal{M}(\alpha, \beta, \mathcal{S}) := \max_{1 \le j \le s} \sup_{z \in \mathcal{S}} |Q_j(z)| \tag{40}$$

definícióval bevezetjük a *belső hibák maximális terjedését mérő tényezőt* (*maximum internal amplification factor*), ahol tehát $(\alpha, \beta)$ a Shu–Osher-alakot, míg $\mathcal{S} \subset \mathbb{C}$ a (17) abszolút stabilitási tartományt jelöli. Ezek után az $\mathcal{M}$ mennyiségre néhány gyakran használt soklépcsős ERK-módszer esetében becsléseket adunk. Ezen becslések közül itt az alábbiakat emeljük ki.

• $(\boldsymbol{i})$ Az SSP-együttható szempontjából optimális másodrendű $s$-lépcsős SSP ERK-módszerek családjában megmutatjuk, hogy

$$\mathcal{M}_s \le \frac{s+1}{s}.$$

• $(\boldsymbol{ii})$ Az SSP-együttható szempontjából optimális harmadrendű $s = n^2$-lépcsős $(2 \le n \in \mathbb{N}^+)$ SSP ERK-módszerek esetén egzaktul meghatározzuk $\mathcal{M}_{n^2}$ értékét $2 \le n \le 10$ esetén, valamint tetszőleges

$n \geq 9$ mellett belátjuk az alábbi kétoldali becsléseket:

$$\frac{9}{10}\sqrt{\frac{n}{\ln(n)}} < \left(1 + \frac{\ln(n)}{n^2} - \frac{\ln(\ln(n))}{n^2}\right)^{(n^2-n)/2} < \mathcal{M}_{n^2} < \left(1 + \frac{\ln(n)}{n^2} - \frac{\ln(\ln(n))}{8n^2}\right)^{(n^2-n)/2} < \frac{\sqrt{n}}{\sqrt[16]{\ln(n)}}. \tag{41}$$

• (***iii***) Az EE-módszerre épülő extrapolációs ERK-módszerek esetén, illetve (***iv***) az explicit középponti szabályra épülő extrapolációs ERK-módszerek esetén a módszer $p$ rendjének függvényében részletesen vizsgáljuk az $\mathcal{M}_p(\alpha, \beta, \mathcal{H})$ kifejezések nagyságrendjét, ahol a (40) definícióban $\sup_{z\in\mathcal{S}}$ helyett $\sup_{z\in\mathcal{H}}$ szerepel. Itt a $\mathcal{H}$ halmaz az $\mathcal{S}$ abszolút stabilitási tartomány alkalmazások szempontjából érdekes részhalmaza: $\mathcal{H} = \mathcal{S}$, $\mathcal{H} = \mathcal{S} \cap \{z \in \mathbb{C} : \mathrm{Re}(z) \leq 0\}$, illetve $\mathcal{H} = \{0\}$ (ez utóbbi eset azért fontos, mert így a belső stabilitást nagyon kis lépésközök esetén lehet mérni).

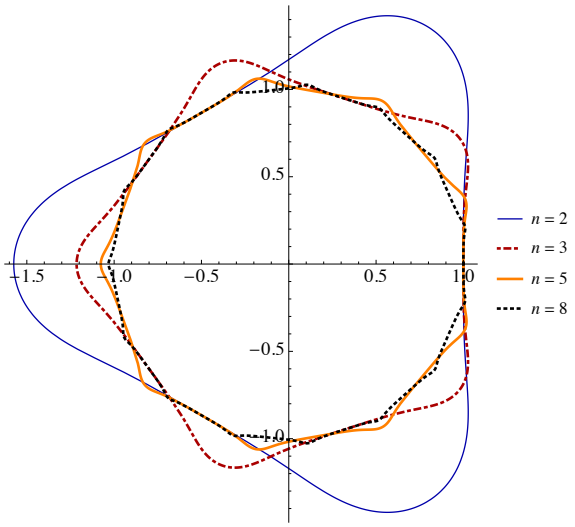
A fenti eredményekhez néhány megjegyzést fűzünk.

Az $(i)$–$(ii)$ pontok azt mutatják, hogy az optimális másod- és harmadrendű SSP-módszerek belső hibáinak terjedése nagyon kedvező, vagyis az $\mathcal{M}$ tényező értéke nagy lépcsőszám esetén is mérsékelt. Az $(i)$ esetben felírt becslésből ez azonnal látszik, míg a $(ii)$ esetben ezt (41) miatt az $\mathcal{M}_s < \sqrt[4]{s}$ ($s = n^2$, $n \geq 4$) egyenlőtlenség mutatja.

Rögzített $n \geq 9$ esetén a (41) becslés igazolásához először az $\mathcal{S}$ tartományt egy skálázással és eltolással normáljuk: néhány ilyen halmaz határgörbéjét tünteti fel az 1. ábra. Ezen halmazok mindegyike tartalmazza a komplex egységkörlapot, és kiderül, hogy az $\mathcal{M}_{n^2}$ konstans felső becsléséhez elegendő a $|Q_j|$ kifejezések szuprémumát az egységkörön kívülre eső „virágszirmokon" megbecsülni. Megmutatjuk, hogy ez a szuprémum egyenlő az

$$[1, +\infty) \ni x \mapsto -1 - \frac{n\, x^{(n-1)^2}\left(1 - \left(1 - \frac{1}{n}\right) x^{2n-1}\right)}{2n-1}$$

polinom egyetlen (1-nél nem kisebb valós) gyökének $(n^2 - n)/2$-edik hatványával. Ennek a gyöknek az alsó- és felső becsléséből adódik (41).




1. ábra: a 3.2. szakasz $(ii)$-es esetéhez tartozó néhány abszolút stabilitási tartomány határgörbéje (átskálázás és eltolás után) a komplex síkon; rögzített $n \geq 2$ esetén a határgörbe implicit egyenlete $\left\{\nu \in \mathbb{C} : \left|\frac{n-1}{2n-1}\nu^{n^2} + \frac{n}{2n-1}\nu^{(n-1)^2}\right| = 1\right\}$.

A $(iii)$–$(iv)$ extrapolációs módszerek esetén azt tapasztaljuk, hogy az $\mathcal{M}$ tényező a $p$ rend függvényében exponenciálisan nő: a hibahalmozódás káros hatását konkrét gyakorlati példában egy soklépcsős 12-edrendű ERK-módszer segítségével demonstráljuk. A $(iii)$ és $(iv)$ módszer közül mindenesetre $(iv)$ rendelkezik jobb belső stabilitási tulajdonságokkal.

Az extrapolációs módszerek lényege a következő. Egy választott alacsonyrendű alapmódszer (esetünkben az EE-módszer $(iii)$-ban, illetve az explicit középponti szabály $(iv)$-ben) és egy lépésszám-sorozat (például $n_j := j \in \mathbb{N}^+$) segítségével a KDE megoldásának $t_{n+1}$-beli értékét a megoldás $t_n$-beli értékének segítségével több menetben közelítjük: az alapmódszert a $[t_n, t_{n+1}]$ intervallumban először $h = (t_{n+1} - t_n)/n_1$ lépésközzel $n_1$-szer, ezután $h = (t_{n+1} - t_n)/n_2$ lépésközzel $n_2$-ször alkalmazzuk, és így tovább. Ezen alacsonyrendű közelítésekből az Aitken–Neville-féle formula alkalmazásával a megoldás magasabb rendű közelítése konstruálható meg.

**3.1. megjegyzés.** *Az extrapolációs módszerek RK-módszerekként is felfoghatók. Mivel az extrapolációs módszerben $p$ tetszőlegesen nagy lehet, ez egyúttal azt is mutatja, hogy léteznek tetszőlegesen magas rendű RK-módszerek.*

Az extrapolációs módszerek esetén – az $(i)$–$(ii)$ SSP-módszerekhez képest – az $\mathcal{M}$ konstans nagyságának meghatározása nehezebb feladatnak bizonyult. Egyrészt az adott alapmódszerhez tartozó $Q_j$ belső stabilitási polinomok az extrapolációs algoritmus alkalmazásakor eleve csak rekurzív módon adottak: az Aitken–Neville-féle formula, mint parciális *differenciaegyenlet* explicit formában való megoldásával azonban a $Q_j$ polinomok explicit alakban megkaphatók. Másrészt ahhoz, hogy a (40) formula alapján e stabilitási polinomok abszolút értékét megbecsüljük a módszer $p$ rendjének függvényében, az $\mathcal{S}_p$ abszolút stabilitási tartományok alakjának minél pontosabb ismerete szükséges – erre a kérdésre a 3.5.1. szakaszban még részleteiben visszatérünk. Ezek után a lépések után az $\mathcal{M}_p$ hibaterjedési konstansok különféle alsó-, felső-, illetve $p \to +\infty$ aszimptotikus becsléseit adjuk meg. Ezek közül egy tipikus számolás például az alábbi. Rögzített $2 \leq p \in \mathbb{N}$ esetén vizsgálandó a

$$\max_{m \in [1,p] \cap \mathbb{N}} \frac{m^p}{(p-m)!m!}$$

kifejezés. Megmutatható, hogy a fenti tört $m$ függvényében először nő, majd csökken, és ha $m_p^*$ jelöli a tört maximumának helyét az $[1, p]$ intervallumban, akkor

$$\lim_{p \to +\infty} \frac{m_p^*}{p} = \frac{1}{1 + W(1/e)} \approx 0.782188.$$

Itt a Lambert-féle $W$-függvényt – melyet a *Mathematica* a `ProductLog` szimbólummal jelöl, és melyre $W : [-1/e, +\infty) \to [-1, +\infty)$ – az

$$x = W(x) \exp(W(x)) \tag{42}$$

egyenlet definiálja.

## 3.3. Új SSP-módszerek

### 3.3.1. Lineáris többlépéses SSP-módszerek változó lépésközzel

Az állandó $h$ lépésközű módszereknél a gyakorlatban sok esetben hatékonyabbnak bizonyulnak a $h_n > 0$ változó lépésközű (adaptív) módszerek. Változó lépésközű LT-módszerekről részletes összefoglaló található például a [20, 21] monográfiákban.

Az [5] dolgozatban az irodalomban először adunk meg változó lépésközű *SSP-tulajdonságú* explicit LT-módszereket. Az ilyen módszerek megkonstruálásakor a fő nehézség az, hogy az aktuális maximális $h_n$ lépésközt (lásd (37)) az ehhez a lépéshez tartozó $\mathcal{C}_n$ SSP-együttható szabja meg, amely együttható viszont függ a választott lépésköz nagyságától. Az [5] dolgozatban

• éles becsléseket adunk változó lépésközű SSP LT-módszerek SSP-együtthatóira – ezek az eredmények H. W. J. Lenferink 1989-es, az állandó lépésközű esetre vonatkozó eredményeit általánosítják;

• optimális SSP-együtthatójú változó lépésközű másod- és harmadrendű módszereket konstruálunk;

• megmutatjuk tetszőlegesen magas rendű változó lépésközű explicit SSP-módszerek létezését;

• elemezzük a $h_n$ lépésközsorozat megválasztására vonatkozó stratégiákat;

• megvizsgáljuk az optimális módszerek stabilitását és konvergenciáját.

Az alábbiakban az [5] cikkből két új – a gyakorlat szempontjából hasznosnak ígérkező – változó lépésközű módszert emelünk ki (autonóm differenciálegyenlet esetére megfogalmazva). Ehhez vezessünk be néhány jelölést. Rögzített $k \geq 2$ egész mellett jelölje $h_n > 0$ a $k$-lépéses LT-módszer $n$-edik lépéséhez tartozó lépésközt (a $h_n$ sorozat megadása is része a módszer definíciójának): ha $t_n$ jelöli a $[t_0, t_0+T]$ időintervallum egy felosztását, akkor tehát $h_n = t_n - t_{n-1}$. Legyen

$$\begin{cases} \Omega_{0,n} := 0, \\ \Omega_{j,n} := \displaystyle\sum_{i=1}^{j} \omega_{i,n} \quad (1 \leq j \leq k), \end{cases}$$

ahol $1 \leq j \leq k$ mellett a lépésközök hányadosait

$$\omega_{j,n} := \frac{h_{n-k+j}}{h_n}$$

jelöli. Az alábbiakban szereplő $\mu_n$ mennyiség a (37) (illetve (33)) formulában szereplő $h_{\mathrm{EE}}$ konstanssal analóg szerepet tölt be (pontos definíciójától itt eltekintünk).

• A $k = 3$ lépéses, optimális másodrendű SSP LT-módszer képlete

$$u_n = \frac{\Omega_{2,n}^2 - 1}{\Omega_{2,n}^2} \left( u_{n-1} + \frac{\Omega_{2,n}}{\Omega_{2,n} - 1}\, h_n\, f(u_{n-1}) \right) + \frac{1}{\Omega_{2,n}^2}\, u_{n-3},$$

ahol a lépésköznek teljesítenie kell a

$$0 < h_n \boxed{\leq} \frac{h_{n-2} + h_{n-1}}{h_{n-2} + h_{n-1} + \mu_n} \cdot \mu_n \tag{43}$$

SSP-feltételt.

• A $k = 4$ lépéses, optimális harmadrendű SSP LT-módszer képlete

$$\begin{aligned} u_n = & \frac{(\Omega_{3,n}+1)^2(\Omega_{3,n}-2)}{\Omega_{3,n}^3} \left( u_{n-1} + \frac{\Omega_{3,n}}{\Omega_{3,n}-2}\, h_n\, f(u_{n-1}) \right) + \\ & \frac{3\Omega_{3,n}+2}{\Omega_{3,n}^3} \left( u_{n-4} + \frac{\Omega_{3,n}(\Omega_{3,n}+1)}{3\Omega_{3,n}+2}\, h_n\, f(u_{n-4}) \right), \end{aligned}$$

ahol a lépésköznek teljesítenie kell a

$$0 < h_n \boxed{\leq} \frac{\sum_{j=1}^{3} h_{n-j}}{\left(\sum_{j=1}^{3} h_{n-j}\right) + 2\mu_n} \cdot \mu_n \tag{44}$$

SSP-feltételt.

**3.2. megjegyzés.** *Az SSP-együttható szempontjából* optimális *3-adrendű k-lépéses formulák pontos leírása meglehetősen technikai; a bizonyítások a lineáris programozásból ismert dualitástételeket, a Farkas-lemmát, valamint paramétertől függő harmadfokú polinomok gyökeinek vizsgálatát tartalmazzák.*

**3.3. megjegyzés.** *Vegyük észre, hogy állandó $h_n = h$ lépésköz esetén $\Omega_{j,n} = j$, továbbá ilyenkor a fenti, változó lépésközű módszerek átmennek a – megfelelő rendű $k$-lépéses – állandó lépésközű „klasszikus" SSP-módszerekbe.*

Az [5] cikkből e rész lezárásaként a változó lépésközű numerikus módszerek alábbi sajátosságát emeljük ki, ami az állandó lépésközű esetben nem fordulhat elő: ha a $h_n$ lépésközsorozat $n \to +\infty$ mellett 0-hoz tart, akkor a numerikus módszer elvben „elakadhat", hiszen ilyenkor nincs garancia arra, hogy véges sok lépésben elérjük a rögzített hosszúságú időintervallum végét. Megmutatjuk viszont, hogy ha a (43)–(44)-beli $\boxed{\leq}$ relációk helyett egyenlőséget írunk, és ily módon *definiáljuk* a $h_n$ lépésközöket (vagyis a lépésközt *mohó algoritmussal* választjuk meg), akkor a $\mu_n$ mennyiségekre tett természetes feltevések mellett a $h_n$ sorozat nem tarthat 0-hoz; látható, hogy ez a lépésközválasztás $k$-lépéses racionális rekurziókra vezet. Az ilyen rekurziók konvergenciája nem nyilvánvaló, szerencsére azonban jól használható általános feltételek állnak rendelkezésre az irodalomban erre vonatkozóan (lásd például a [40] könyvet); a nehézség az, hogy a szokásos monotonitási érvelések nem alkalmazhatók közvetlenül, mert e sorozatok hosszú kezdőszeletei viselkedhetnek nem monoton módon. Illusztrációképpen tekintsük a

$$\tau_1 := 1, \quad \tau_2 := \frac{1}{200}, \quad \tau_3 := \frac{95638788642}{100000000000} \quad \text{és}$$

$$\tau_n := \frac{\tau_{n-1} + \tau_{n-2} + \tau_{n-3}}{1 + \tau_{n-1} + \tau_{n-2} + \tau_{n-3}} \qquad (n \geq 4)$$

képlettel megadott racionális rekurziót. Ennél azt tapasztaljuk, hogy a sorozat első 84 tagja „oszcillál" (azaz ezen tagok között nincs 3-nál hosszabb monoton részsorozat), majd innentől kezdve válik csak a sorozat monoton csökkenővé (legalábbis az első 1000 tag vizsgálata alapján).

Az [5] cikket részletes numerikus tesztek zárják, melyekben egy-, illetve kétdimenziós nemlineáris parciális differenciálegyenletek szemidiszkretizációjával nyert közönséges differenciálegyenlet-rendszerek megoldását approximáljuk változó lépésközű SSP-módszerekkel, és demonstráljuk a változó lépésközű módszerek számítási hatékonyságát az állandó lépésközű módszerekhez képest.

### 3.3.2. Egylépéses SSP-módszerek folytonos kiterjesztése

Az 1.2–1.3. szakaszban ismertetett numerikus módszerek a differenciálegyenlet megoldását a diszkrét $t_0$, $t_1, \ldots$ időpontokban approximálják. Bizonyos fizikai jelenségeket modellező differenciálegyenletek diszkretizációjakor egyes részintervallumokon szükség lehet arra, hogy a $t_0$, $t_1, \ldots$ alappontokhoz képest sokkal sűrűbben ismerjük a megoldás közelítését. Hasonló elvárás fogalmazható meg a numerikus megoldások grafikus megjelenítésekor is, ahol az a cél, hogy az approximáció folytonos görbedarabokból álljon. Ilyen esetekben a numerikus módszer egy jóval kisebb lépésközzel persze újraindítható az adott intervallumon, de számítási szempontból ehelyett sokszor érdemes a numerikus módszer *folytonos kiterjesztésére* vonatkozó formulák alkalmazását megfontolni; ezek angol neve gyakran *dense output* vagy *continuous extension*. A folytonos kiterjesztések lényege az, hogy a $t_n, t_{n+1}$ időpillanatokról úgy terjesszük ki az approximációt a $[t_n, t_{n+1}]$ intervallumra, hogy közben a KDE-ben szereplő $f$ függvényt *ne kelljen újra kiértékelni*. RK-módszerek esetén Hermite-interpoláció segítségével például folytonos kiterjesztések konstruálhatók, amelyek rendje legalább három (lásd [20, Section 2.6]). Felmerül a kérdés, hogy mi mondható az olyan folytonos kiterjesztésekről, amelyek SSP-tulajdonsággal rendelkeznek.

A [4] cikkben SSP RK-módszerek folytonos kiterjesztését vizsgáljuk. Megmutatjuk, hogy – természetes feltevések mellett – *nem létezik* harmad- vagy magasabb rendű SSP-tulajdonságú folytonos kiterjesztés. Pozitív eredményként első- és másodrendű, *egyszerű szerkezetű* feltételeket adunk folytonos kiterjesztésekre az általános esetben, illetve konkrét, optimális SSP-módszerek esetén. Az SSP-tulajdonságú folytonos kiterjesztések létezése sokváltozós polinomiális egyenlőtlenségrendszer megoldhatóságával fogalmazható meg. Ezen polinomiális rendszerek megoldását (vagy megoldhatatlanságát) a *Mathematica* segítségével szimbolikusan, illetve a bonyolultabb esetekben numerikusan térképeztük fel. A cikkben szereplő állítások és bizonyítások e szisztematikus kísérletek eredményeként születtek meg.

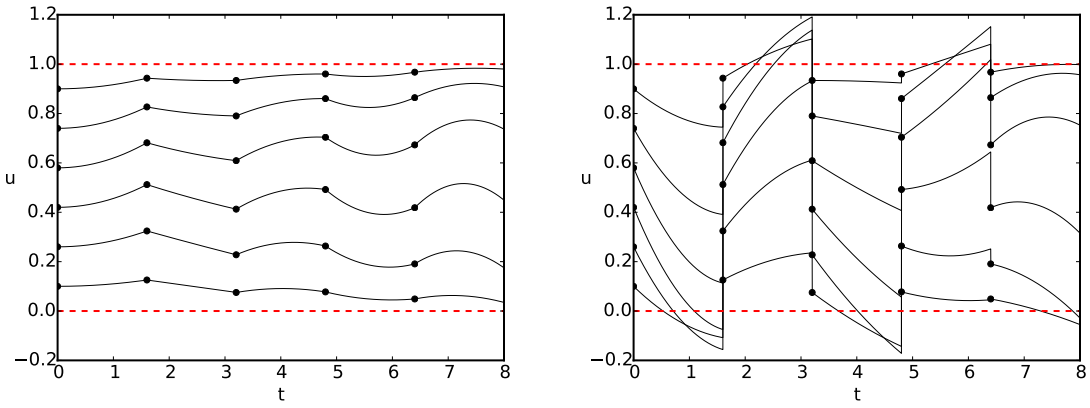
A cikkbeli konstrukciók megvilágítása céljából tekintsük a 2.3.2. szakaszban említett differenciálegyenletet, amelyet ott SSP RK-módszerrel diszkretizáltunk, itt pedig annak az SSP RK-módszernek két folytonos kiterjesztését hasonlítjuk össze. Mindkét folytonos kiterjesztés másodrendű, és a $\theta \in [0,1]$ folytonos paraméter valósítja meg az interpolációt a $t_n$ és $t_{n+1}$ pontok között: $u_{n+\theta}$ az $u(t_n+\theta h)$ pontos megoldás egy közelítése. Az egyik formula

$$u_{n+\theta} = \left(1 - 2\theta + \frac{4}{3}\theta^2\right) u_n + 2(\theta - \theta^2)\left(y_1 + \frac{h}{2} f(t_n, y_1)\right) + \frac{2}{3}\theta^2 \left(y_3 + \frac{h}{2} f(t_n + h, y_3)\right), \qquad (45)$$

míg a másik képlet

$$u_{n+\theta} = u_n + h\left((2\theta - \theta^2) f(t_n, y_1) + (-2\theta + \theta^2) f(t_n + h/2, y_2) + \theta f(t_n + h, y_3)\right). \qquad (46)$$

A (28) egyenlet különböző $u_0 \in \mathcal{I}_1 = [0,1]$ kezdeti feltételekhez tartozó megoldásainak a $h = 16/10$ lépésközű (32) módszerrel nyert diszkrét approximációit a 2. ábrán szereplő fekete pontok halmaza jelöli – a két ponthalmaz a bal és a jobb ábrán azonos. A folytonos görbedarabok úgy adódnak, hogy a (45), illetve (46) formulákat a (28) egyenletre alkalmazzuk, szintén a $h = 16/10$ választással. Ahogyan az ábra mutatja, az SSP-tulajdonságú (45) folytonos kiterjesztés megőrzi az $u_{n+\theta} \in \mathcal{I}_1$ relációt, míg a nem SSP-tulajdonságú (46) módszer esetén ez az invarianciatulajdonság nem áll fenn. A (45) módszer SSP-tulajdonságot megőrző lépésközének maximális nagyságával kapcsolatban egyébként elméletileg azt tudjuk garantálni (amit a numerikus tesztek is megerősítenek), hogy a képlet által szolgáltatott folytonos kiterjesztés az $\mathcal{I}_1$ intervallumban marad, amennyiben a lépésközre fennáll a $h \in [0, 2h_{\mathrm{EE}}]$ (itt $h_{\mathrm{EE}} = 1$) megszorítás.




(a) SSP-módszer (b) nem SSP-módszer

2. ábra: a (28) egyenlet numerikus megoldása a folytonosan kiterjesztett SSP-tulajdonságú (45), illetve a nem SSP-tulajdonságú (46) módszerrel.

### 3.3.3. Konvekciós egyenletek egylépéses SSP-módszerei

A 2.4. szakaszban láttuk, hogy ha a KDE jobb oldalán szereplő $f$ függvény szerkezetéről bizonyos információval rendelkezünk, akkor különféle $\gamma_{\mathrm{sup}}$ lépésköz-együtthatók definiálhatók, amelyek a kvalitatív tulajdonságokat megtartó numerikus módszer hatékonyságát jellemzik.

A [3] dolgozatban speciális alakú $f$ függvények esetén vizsgáljuk a (4) egyenletrendszer RK-diszkretizációit a nemnegativitás megőrzésének szempontjából: azt tesszük fel, hogy az eredeti KDE-rendszer alakja

$$u_k'(t) \;=\; q_k(u(t), t)\, \frac{u_{k-1}(t) - u_k(t)}{\Delta x} \quad (k = 1, \ldots, N). \qquad (47)$$

Itt $\Delta x > 0$ egy rögzített konstans, a *nemnegatív* és korlátos $q_k$ $(k = 1, \dots, N)$ függvények megfelelően simák, az $u = (u_1, \dots, u_N) : [t_0, t_0 + T] \to \mathbb{R}^N$ függvény pedig a KDE megoldását jelöli, szintén *nemnegatív* $u_k(t_0) \geq 0$ $(k = 1, \dots, N)$ kezdeti feltételek mellett. Megmutatható, hogy ezen feltevések mellett a (47) rendszer pozitív, azaz $u(t) \geq 0$ $(t \geq t_0)$. E speciális, (47) alakú KDE-rendszerek azért fontosak, mert ezek közé tartoznak például az egydimenziós skaláris hiperbolikus parciális differenciálegyenletek TVD-tulajdonságú szemidiszkretizációi (emlékezzünk rá korábbról, hogy eredetileg ehhez a függvényosztályhoz fejlesztették ki az SSP-módszereket), de ide sorolhatók az egydimenziós hővezetési (parabolikus) PDE bizonyos szemidiszkretizációi is; ekkor $\Delta x > 0$ a térbeli diszkretizációs paramétert jelöli. A cikkben megmutatjuk, hogy a (47) rendszer esetében számos ERK-módszer nagyobb $\gamma$ lépésköz-együtthatóval rendelkezik, mint amilyen együtthatók az általános SSP-elméletből (azaz általános nemlineáris $f$ esetén) következnek; más szavakkal, hatékony ERK-módszereket konstruálunk egy, a gyakorlatban fontos függvényosztály elemeinek nemnegativitást őrző idődiszkretizációjához. A fő kérdés konkrétabb megfogalmazása a következő: a (47) rendszer diszkretizálásakor mely RK-módszerek milyen lépésköz-együtthatóval fognak nemnegatív $u_0$ kezdővektorból indítva nemnegatív $u_n$ approximációt generálni?

**3.4. megjegyzés.** *A diszkrét pozitivitási tulajdonság mellett a cikkben egy általánosabb tulajdonságot, az értékkészlet korlátosságát is elemezzük (lásd korábban az 1.1.6. szakaszt).*

A [41] dolgozat a fenti kérdést az ERK$(2, 2)$ osztályban vizsgálta, amely osztály a (11) formula szerint egyparaméteres. A [3] cikkben – követve [41] alapgondolatát – elsősorban az ERK$(3, 3)$ családot vesszük szemügyre. A [41, 3] cikkek közvetlenül elemzik a teljes diszkretizációt, tehát a vizsgálati módszer eltér a hagyományos SSP-elmélettől.

Kiderül, hogy a (47) KDE-rendszerre alkalmazott $(A, b)$ Butcher-táblájú $s$-lépcsős RK-módszer (ahol $A \in \mathbb{R}^{s\times s}$ és $b \in \mathbb{R}^s$) bizonyos többváltozós $P_i$ polinomokkal írható le: az RK-módszerhez $s + 1$ darab, külön-külön $s(s + 1)/2$ változós polinom tartozik. Ezután az RK-módszer $\gamma$ *pozitivitási lépésköz-együtthatóját* a

$$\gamma = \gamma(A, b) := \sup\left\{\delta \geq 0 : P_i(\xi) \geq 0 \text{ minden } 0 \leq i \leq s \text{ és minden } \xi \in [0, \delta]^{s(s+1)/2} \text{ esetén}\right\} \tag{48}$$

képlettel értelmezzük: geometriailag $\gamma(A, b)$ tehát az első ortánsban fekvő legnagyobb olyan hiperkockának az élhossza, amelyben az összes $P_i$ polinom nemnegatív.

A (12) képlettel leírt, ERK$(3, 3)$-beli kétparaméteres alosztály esetén a $P_i$ polinomok például az alábbi alakot öltik:

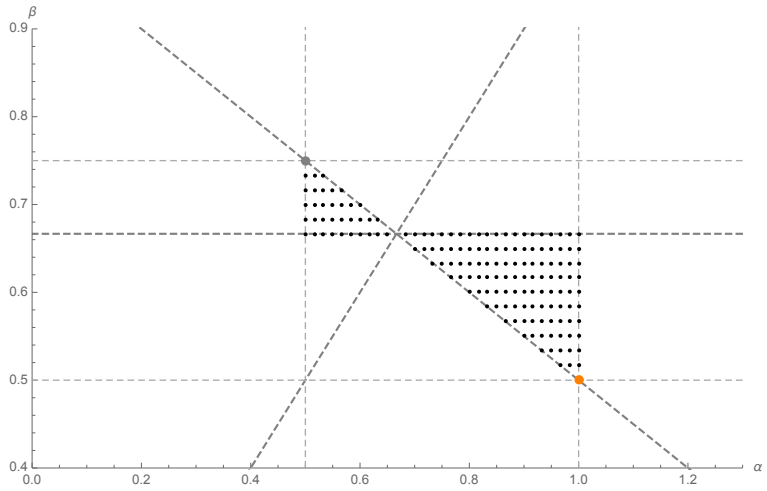
$$\begin{aligned} P_0(x, y, z, u, v, w) = \frac{1}{6\alpha\beta(\alpha - \beta)}\Big[&-\alpha^2\beta(vwz - 3wz + 6z - 6) + \alpha\beta^2(vwz - 3vz + 6z - 6) + \\ &\alpha\beta(vw + 2vz - 3wz) - \beta^2(w - 3)(v - z) - 2\beta(v - z) - 3\alpha^2(w - z) + 2\alpha(w - z)\Big], \end{aligned}$$

$$\begin{aligned} P_1(x, y, z, u, v, w) = \frac{1}{6\alpha\beta(\alpha - \beta)}\Big[&\alpha^2\beta(uwy + vwy + vwz - 3wy - 3wz + 6z) - \\ &\alpha\beta^2[uwy + v(w - 3)(y + z) + 6z] - \alpha\beta(uw + vw + 2vy + 2vz - 3wy - 3wz) + \\ &\beta^2[uw + v(w - 3) - wy - wz + 3z] + 2\beta(v - z) + 3\alpha^2(w - z) - 2\alpha(w - z)\Big], \end{aligned}$$

$$\begin{aligned} P_2(x, y, z, u, v, w) = \frac{1}{6\alpha(\alpha - \beta)}\Big[&-\alpha^2 w[u(x + y) + (v - 3)y] + \\ &\alpha\beta[uw(x + y) + vy(w - 3)] + \alpha(uw + 2vy - 3wy) + \beta w(y - u)\Big], \end{aligned}$$

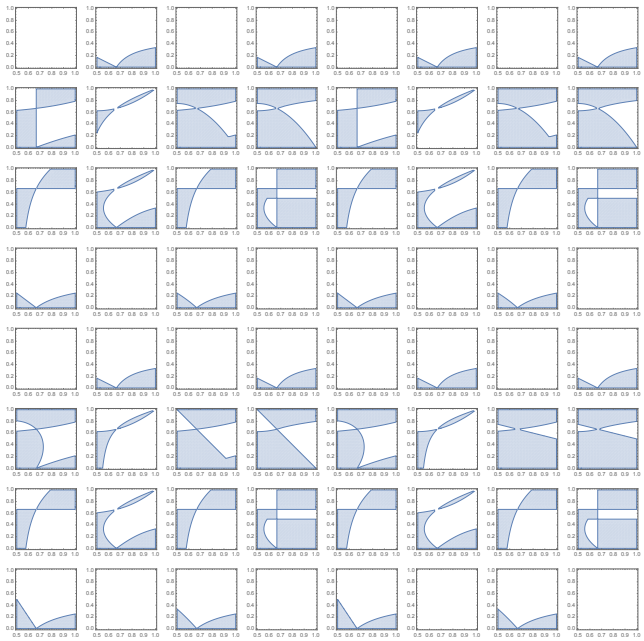
$$P_3(x,y,z,u,v,w) = \frac{1}{6}xuw.$$

Ebben az alosztályban a nemnegativitás-őrzés szempontjából optimális lépésköz-együtthatóval rendelkező RK-módszer(ek) megkeresése az alábbit jelenti: adjuk meg a legnagyobb olyan $\gamma > 0$ értéket és a hozzá tartozó $(\alpha, \beta) \in \mathbb{R}^2$ paraméterpárok halmazát, melyek esetén a fenti $P_{0,1,2,3}$ polinomok mindegyike nemnegatív a $[0,\gamma]^6$ hiperkockában. A cikkben megmutatjuk, hogy ebben az alosztályban $\gamma \leq 1$, továbbá a $\gamma = 1$ optimális lépésköz-együtthatóval rendelkező RK-módszereket az $(\alpha, \beta)$ paramétersíkon a 3. ábrán látható két, pontokkal kitöltött háromszög uniója (egy „csokornyakkendő") írja le. A 3. ábra „csokornyakkendőjének" bal felső sarkában látható szürke pont az az egyetlen RK-módszer, melynek SSP-együtthatója a teljes ERK(3, 3) osztályban az optimális $\mathcal{C} = 1$ értékkel bír; ezt a módszert az SSP-elmélet keretében korábban azonosították. A 3. ábra azt demonstrálja, hogy számos más, ERK(3, 3)-beli módszer rendelkezik $\gamma = 1$ pozitivitási lépésköz-együtthatóval. Ezek között kitüntetett módszer a „csokornyakkendő" jobb alsó sarkában látható narancssárga pont, amely az ERK(3, 3)-beli (egyetlen) Ralston-módszert reprezentálja: ismeretes (lásd [42]), hogy ez a módszer a teljes ERK(3, 3) osztályban minimális hibakonstanssal rendelkezik. Eredményeinkből tehát az is következik, hogy a Ralston-módszer kettős optimalitási tulajdonsággal bír: hibakonstansa a lehető legkisebb, míg pozitivitási lépésköz-együtthatója a lehető legnagyobb az ERK(3, 3) osztályban.



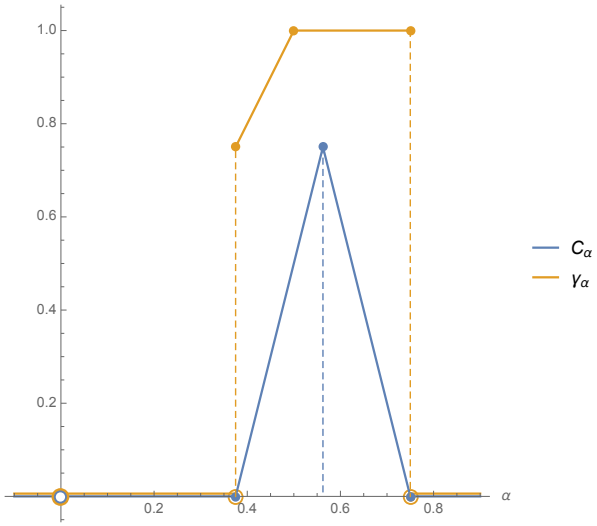

3. ábra: a „csokornyakkendő", vagyis azon $(\alpha, \beta)$ párok halmaza a (12) képlettel leírt ERK(3, 3)-beli kétparaméteres alosztályban, melyekre a (48) lépésköz-együttható értéke 1. A jobb alsó narancssárga pont koordinátái $(\alpha, \beta) = (1, 1/2)$, míg a bal felső szürke pont által reprezentált módszerre $(\alpha, \beta) = (1/2, 3/4)$.

**3.5. megjegyzés.** *Az $(\alpha, \beta)$ paraméterektől függő $P_i$ polinomok $[0,1]^6$ hiperkockában megkövetelt szimultán nemnegativitásának vizsgálata nem nyilvánvaló feladat. Itt fő eszközként a Mathematica* `Reduce` *és* `FindInstance` *parancsát használtuk, melyekkel a hiperkocka mind a 64 csúcsát végigvizsgálva adódott a 4. ábra. Ezen ábrasor alapján már viszonylag egyszerű volt egy rövid,* hagyományos *bizonyítást (lásd 1.1.1. szakasz) adni arra nézve, hogy a „csokornyakkendőn" kívüli paraméterekre $\gamma < 1$ igaz. Arra az állításra, hogy a „csokornyakkendőhöz" tartozó összes (végtelen sok) ERK-módszerre $\gamma = 1$, jelenleg nincs se számítógépes, se hagyományos bizonyításunk – ez tehát továbbra is csak sejtés. A sejtést viszont alátámasztja az alábbi eredmény: a „csokornyakkendőben" feltüntetett (véges sok) fekete ponthoz tartozó RK-módszerre* számítógépes *bizonyítás mutatja, hogy $\gamma = 1$; a számunkra legérdekesebb narancs ponthoz tartozó RK-módszer esetén pedig néhány soros* hagyományos *bizonyítást sikerült adni a $\gamma = 1$ állításra.*



4. ábra: a $[0,1]^6$ hiperkocka 64 csúcsát végiglátogatva a satírozott tartományok jelképezik azon $(\alpha, \beta)$ párokat, melyekre legalább az egyik $P_i$ polinom negatív. Ezeket az ábrákat összesítve született a 3. ábra „csokornyakkendője”.

A [3] cikkben azt is igazoljuk, hogy a *teljes* ERK(3, 3) osztályban igaz a $\gamma \leq 1$ egyenlőtlenség. Ehhez részletesen megvizsgáljuk az ERK(3, 3) osztály másik két, egyparaméteres alosztályát. Az egyik ilyen, $\alpha$-val paraméterezett alosztályban az 5. ábrán hasonlíthatjuk össze a klasszikus $\mathcal{C}_\alpha$ SSP-együttható és az újonnan bevezetett $\gamma_\alpha$ pozitivitási lépésköz-együttható egymáshoz való viszonyát.




5. ábra: az SSP-elmélet által garantált $\mathcal{C}$ SSP-együttható (kék szín) és a $\gamma$ lépésköz-együttható (sárga szín) az ERK(3, 3) osztály egyik egyparaméteres alosztályában, az $\alpha$ paraméter függvényében. Az ábra azt illusztrálja, hogy speciális alakú KDE-rendszerek esetén hogyan lehet megjavítani az általános SSP-elmélet eredményeit.

**3.6. megjegyzés.** *Az 5. ábrán szereplő sárga grafikon legmagasabban fekvő vízszintes szakasza 3 darab hatváltozós és 1 paramétertől függő polinom nemnegativitását mutatja a* $[0,1]^6$ *hiperkockában. Ezt az állítást a Mathematica néhány tizedmásodperc alatt igazolja (számítógépes bizonyítás). Az ennek megfelelő, és a [3] cikk egyik bírálója által utólag kért hagyományos bizonyítás megalkotása mindössze néhány óráig tartott (természetesen a Mathematica segítségével); e másfél oldalas és teljesen elemi hagyományos bizonyítás csak a polinomok értékkészletét becsüli meg, és nem használja a derivált fogalmát.*

**3.7. megjegyzés.** *A 3. ábra „csokornyakkendőjének" létezése, vagy például az 5. ábra több korábbi cikkben leírt megfigyelést is megmagyaráz. Numerikus kísérletekben sokszor azt tapasztalták, hogy az SSP-elméletből adódó, az RK-módszer lépésközére vonatkozó megszorítás túl pesszimista: a konkrét módszer az SSP-elmélet által jósoltnál nagyobb lépésköz esetén is pozitivitásőrző. Ha figyelembe vesszük, hogy a példákban használt tesztegyenletek szerkezete meglehetősen speciális, akkor erre a függvényosztályra az SSP-elmélet helyett például a [3] cikkben leírt elméletet alkalmazva az SSP-együtthatónál gyakran nagyobb pozitivitási lépésköz-együtthatókat kapunk.*

Az ERK$(3,3)$ család után magasabb rendű ($p \geq 4$) és többlépcsős ($s \geq 4$) RK-módszereket is szemügyre veszünk: numerikus és szimbolikus technikákkal sem sikerült azonban ezekben az osztályokban pozitív $\gamma > 0$ pozitivitási lépésköz-együtthatójú ERK-módszert találni – az ilyen módszerek létezése tehát továbbra is nyitott kérdés maradt.

**3.8. megjegyzés.** *A 4-lépéses, 4-lépcsős explicit Runge–Kutta-módszerek vizsgálatakor az alábbi meglepő karakterizációt kaptuk meg: az ERK(4,4) családból vett módszer* $(A,b)$ *Butcher-táblájára* $A \geq 0$ *és* $b \geq 0$ *(elemenként) akkor és csak akkor áll fenn, ha a módszer a* (13) *klasszikus negyedrendű RK-módszer (RK44). Később kiderült, hogy ez az eredmény korábban már implicit módon ismert volt, az állítás ugyanis kiolvasható bizonyos, stabilitással kapcsolatos tételek (például [27, Theorem 9.6]) technikai jellegű bizonyításából – tudomásunk szerint azonban explicit módon még sehol sem fogalmazták meg az RK44-módszer fenti jellemzését a Butcher-mátrix nemnegativitásával.*

**3.9. megjegyzés.** *A [43] cikk fő eredménye azt állítja, hogy a* (13) *klasszikus negyedrendű RK-módszer pozitivitási lépésköz-együtthatójára* $\gamma = 1$ *teljesül. A [3] dolgozatban ezt az állítást ellenpéldával cáfoljuk meg. Az elemzésünk kideríti, hogy az RK44-módszerhez tartozó 5 darab 10 változós* $P_i$ *polinom legalább egyike negatív értéket vesz fel a* $[0,\varepsilon]^{10}$ *hiperkockában,* tetszőleges $\varepsilon > 0$ *esetén; úgy tűnik, hogy a [43] cikkben csak a* $[0,1]^{10}$ *hiperkocka csúcsaiban vizsgálták e polinomok nemnegativitását.*

## 3.4. Egy- és többlépéses módszerek monotonitása és korlátossága

### 3.4.1. Racionális törtfüggvények abszolút monotonitása

A 2.4. szakaszban azt láttuk, hogy

($i$) az általános, lineáris vagy nemlineáris, (4) alakú rendszerek;

($ii$) a lineáris, (5) alakú rendszerek

RK-diszkretizációja esetén a numerikus módszer SSP-tulajdonságát biztosító $\gamma_{\text{sup}}$ maximális lépésköz-együttható (vagyis SSP-együttható) bizonyos irreducibilitási feltételek mellett

- az ($i$) esetben megegyezik a (35) képlettel definiált $R(A,b)$ Kraaijevanger-együtthatóval;
- a ($ii$) esetben megegyezik az $R(\psi_{A,b})$ mennyiséggel, vagyis az RK-módszer (16) stabilitási függvényének (2) abszolút monotonitási sugarával.

A gyakorlat szempontjából fontos annak az ismerete, hogy RK-módszerek valamely rögzített osztályában mennyi az $R(A,b)$ és $R(\psi_{A,b})$ együtthatók optimális értéke, illetve melyek az ezekhez tartozó optimális módszerek. Ismert, hogy egy RK-módszer esetén

$$0 \le R(A,b) \le R(\psi_{A,b}). \tag{49}$$

*Explicit* $p$-edrendű $s$-lépcsős ($p \ge 1$, $s \ge 1$) RK-módszerekre bebizonyították, hogy $R(\psi_{A,b}) \le s$, tehát $R(A,b) \le s$ is fennáll. *Implicit* RK-módszerek esetén sokkal kevesebbet tudunk az $R(A,b)$ és $R(\psi_{A,b})$ együtthatók optimális értékéről. Ezen eredmények ismertetéséhez vezessünk be néhány jelölést.

Az 1.2. szakaszban láttuk, hogy ERK-módszerek esetén a $\psi_{A,b}$ stabilitási függvény polinom, míg IRK-módszerek esetén racionális törtfüggvény. Pozitív egész $m$ és $n$ esetén azt mondjuk, hogy egy racionális törtfüggvény $(m,n)$-típusú, ha a számlálója $m$-edfokú, míg nevezője $n$-edfokú polinom. A $\Pi_{m/n,p}$ függvényosztály elemei olyan $(m,n)$-típusú racionális törtfüggvények, amelyek az exponenciális függvényt az origóban legalább $p$-edrendben közelítik. A $\widehat{\Pi}_{m/n,p}$ osztály részhalmaza $\Pi_{m/n,p}$-nek: a $\widehat{\Pi}_{m/n,p}$-beli racionális törtfüggvények nevezőjében szereplő polinom teljes $n$-edik hatvány, vagyis a nevező $(1-az)^n$ alakú ($a \in \mathbb{R}$). Egy $p$-edrendű $s$-lépcsős IRK-módszerre $\psi_{A,b} \in \Pi_{s/s,p}$, míg egy SDIRK-módszerre $\psi_{A,b} \in \widehat{\Pi}_{s/s,p}$. Jelölje végül a $\Pi_{m/n,p}$ és a $\widehat{\Pi}_{m/n,p}$ osztályban elérhető legnagyobb abszolút monotonitási sugarat rendre

$$R_{m/n,p} := \sup\{R(\psi) : \psi \in \Pi_{m/n,p}\},$$
$$\widehat{R}_{m/n,p} := \sup\{R(\psi) : \psi \in \widehat{\Pi}_{m/n,p}\};$$

amennyiben valamelyik halmaz üres, a szuprémumot 0-nak definiáljuk. Az IRK-módszerekkel kapcsolatos legfontosabb sejtések az alábbiak. (A zárójelben álló évszám a sejtés megfogalmazásának évét jelöli.)

**3.10. sejtés (1986).** *Legyen $m,n \in \mathbb{N}^+$. Ekkor $R_{m/n,2} = m + \sqrt{mn}$.*

Ebből speciális esetként nyilvánvalóan adódik az alábbi sejtés.

**3.11. sejtés.** *Speciálisan, ha $m = n = s \in \mathbb{N}^+$, akkor $R_{s/s,2} = 2s$.*

**3.12. sejtés (2008).** *Tekintsünk egy $(A,b)$ Butcher-táblázattal megadott $p$-edrendű $s$-lépcsős SDIRK-módszert, ahol $p \ge 2$ és $s \ge 1$. Ekkor $R(A,b) \le 2s$.*

**3.13. sejtés (2009).** *Tekintsünk egy $(A,b)$ Butcher-táblázattal megadott $p$-edrendű $s$-lépcsős IRK-módszert, ahol $p \ge 2$ és $s \ge 1$. Ekkor $R(A,b) \le 2s$.*

A 3.10. sejtés igazsága $n \in \{1,2\}$ és minden $m \ge 1$ esetén ismert. A 3.12. és 3.13. sejtést $s \in \{1,2\}$ és $p = 2$ esetén bebizonyították: a legérdekesebb a $p = 2$ eset, hiszen a legalább harmadrendű RK-módszerek automatikusan teljesítik a $p = 2$ rendfeltételeket is. A 3.13. sejtés az SSP RK-módszerek *legfontosabb, jelenleg is nyitott kérdésének* tekinthető.

**3.14. megjegyzés.** *A definíciókat összehasonlítva azt látjuk, hogy egy adott $(s,p)$ pár esetén az $\widehat{R}_{s/s,p}$ konstans értékének meghatározásánál nehezebb feladat $R_{s/s,p}$ értékének meghatározása, és ennél is jóval nehezebb az optimális $R(A,b)$ értékének kiszámítása:*

- *még egy konkrét (azaz paramétertől nem függő) $\psi$ racionális törtfüggvény esetén sem nyilvánvaló az $R(\psi)$ abszolút monotonitási sugár kiszámítása;*
- *a 3.13. sejtésben szereplő paraméterek száma az $s$ lépcsőszám növekedésével négyzetesen nő. Ha például $p = s = 2$, akkor olyan optimumszámítási feladattal állunk szemben, amely 9 polinomiális egyenlőtlenséget és 2 egyenletet tartalmaz 7 változóban.*

**3.15. megjegyzés.** *Egy több lépcsővel rendelkező RK-módszer esetén több számítási munkát kell végezni, de $\mathcal{C}$ SSP-együtthatója is nagyobb lehet. Hogy e két tényezőt egyszerre lehessen figyelembe venni, érdemes*

*bevezetni az* effektív SSP-együttható *fogalmát, amelyet a $\mathcal{C}/s$ hányadossal definiálunk. A fentiek fényében tudjuk, hogy ERK-módszerek effektív SSP-együtthatója legfeljebb 1, és sejtjük, hogy IRK-módszerek effektív SSP-együtthatója legfeljebb 2.*

*Az LT-módszerekre vonatkozó analóg eredményeket H. W. J. Lenferink már az 1990-es években bebizonyította: explicit LT-módszer SSP-együtthatója legfeljebb 1, míg implicit LT-módszer SSP-együtthatója legfeljebb 2. Vajon az RK- és LT-módszereket is magukban foglaló explicit és implicit* általános lineáris *módszerek SSP-együtthatói esetén is fel lehet majd fedezni ezt az* $1:2$ *arányt?*

Az SDIRK $\subset$ IRK tartalmazás miatt a 3.13. sejtés általánosabb a 3.12. sejtésnél, mégis, a részletes numerikus kísérletek szerint (rögzített $p$ és $s$ mellett) az $R(A,b)$ együttható szempontjából optimális IRK-módszer egyúttal az SDIRK-osztály eleme is. Úgy véljük, hogy e két sejtésben $p=2$ és rögzített $s\geq 1$ mellett az *egyértelmű optimális* racionális törtfüggvény a

$$\psi^*_{A,b}(z) := \frac{\left(1+\frac{z}{2s}\right)^s}{\left(1-\frac{z}{2s}\right)^s} \tag{50}$$

függvény, melyhez egyszerű Butcher-táblájú RK-módszer tartozik, és melyre fennáll, hogy

$$\psi^*_{A,b}\in\widehat{\Pi}_{s/s,2}\subset\Pi_{s/s,2}\quad\text{és}\quad R(A,b)=R(\psi^*_{A,b})=2s.$$

Ezek után figyelembe véve, hogy rögzített $s$-re a $p\mapsto R_{s/s,p}$ leképezés nemnövekvő, a (49) egyenlőtlenség alapján *a 3.11. sejtés igazsága maga után vonná a 3.13. sejtés megoldását.*

Sajnálatos módon azonban kiderül, hogy az előző mondatban feltételes módban megfogalmazott reményteljes út nem járható: a [9] cikkben bebizonyítjuk, hogy a 3.11. sejtés $s=3$ esetén nem igaz. Ezt például a

$$\psi(z)=\frac{\frac{1246}{384649}z^3+\frac{2289}{34970}z^2+\frac{119}{269}z+1}{-\frac{4}{327}z^3+\frac{8}{65}z^2-\frac{150}{269}z+1}$$

racionális törtfüggvény mutatja, melynek $R(\psi)$ abszolút monotonitási sugarát a

$$\{43572620,-880461561,5950520030,-13451175530\}$$

harmadfokú polinom egyetlen valós gyöke adja meg (a jelöléssel kapcsolatban lásd (1)); e gyök értéke pedig körülbelül $6.7783>2\cdot 3$.

Numerikus kísérletek alapján a cikkben további sejtéseket fogalmazunk meg, melyek közül itt az alábbi kettőt idézzük.

**3.16. sejtés.** *Tetszőleges $s\geq 3$ esetén $\widehat{R}_{s/s,2}=2s$ (az állítás $s\in\{1,2\}$ esetén is igaz, ezek bizonyítása ismert).*

**3.17. sejtés.** *Tetszőleges $s\geq 2$ esetén $\widehat{R}_{s/s,3}=s-1+\sqrt{s^2-1}$.*

Pozitív eredményként számos $(s,p)$ pár esetén meghatározzuk az $R_{s/s,p}$ és $\widehat{R}_{s/s,p}$ konstansok *egzakt* értékét, melyeket az 1. és 2. táblázatban tekinthetünk át. A rövidség kedvéért ($R_{4/4,7}$ kivételével) itt nem közöljük a táblázatban szereplő magasabb fokú algebrai számok explicit alakját, csak numerikus approximációikat.

| $(s,p)$ | $R_{s/s,p}$ | $\deg(R_{s/s,p})$ |
|---|---|---|
| $(2,2)$ | $= 4^{\dagger}$ | 1 |
| $(2,3)$ | $\approx 2.732050^{\ddagger}$ | 2 |
| $(2,4)$ | $= 0^{\dagger}$ | 1 |
| $(3,5)$ | $\approx 2.301322$ | 6 |
| $(3,6)$ | $\approx 2.207606^{\ddagger}$ | 3 |
| $(4,7)$ | $\approx 2.743911$ | 30 |
| $(4,8)$ | $= 0^{\dagger}$ | 1 |

1. táblázat: a [9] cikkben meghatározott optimális $R_{s/s,p}$ értékek, illetve numerikus approximációik. A felső indexben található $^{\dagger}$, illetve $^{\ddagger}$ szimbólumok rendre azt jelzik, hogy a megfelelő konstans értékét korábban már egzakt módszerekkel, illetve numerikusan meghatározták. A deg oszlopban található értékek az $R_{s/s,p}$ algebrai szám fokát adják meg.

| $(s,p)$ | $\widehat{R}_{s/s,p}$ | $\deg(\widehat{R}_{s/s,p})$ |
|---|---|---|
| $(2,2)$ | $= 4^{\dagger}$ | 1 |
| $(3,2)$ | $= 6$ | 1 |
| $(4,2)$ | $= 8$ | 1 |
| $(2,3)$ | $\approx 2.732050^{\ddagger}$ | 2 |
| $(3,3)$ | $\approx 4.828427$ | 2 |
| $(4,3)$ | $\approx 6.872983$ | 2 |
| $(3,4)$ | $\approx 3.287278$ | 9 |
| $(4,4)$ | $\approx 5.167265$ | 9 |
| $(4,5)$ | $\approx 3.743299$ | 12 |

2. táblázat: az optimális $\widehat{R}_{s/s,p}$ értékek (a jelmagyarázat megegyezik az 1. táblázatéval)

A [9] dolgozat nagyban épít a [26] cikkben megfogalmazott eredményekre; [26] egyébként több más $R_{m/n,p}$ értéket is tartalmaz $m \neq n$ esetén, ám ezen értékek mindegyike (irracionális számok esetén) csak *közelítő* numerikus érték. A [9] cikk eredményei többek között

- $s \in \{3,4\}$ esetén megoldják a 3.12. és a 3.16. sejtést;
- $s \in \{2,3,4\}$ esetén a 3.17. sejtést;
- $s \in \{2,3,4\}$ és $p = 2$ esetén igazolják az abszolút monotonitási sugár szempontjából optimális függvény unicitásának kérdését a $\widehat{\Pi}_{s/s,2}$ osztályban (lásd (50)), továbbá
- teljes leírást adnak az $\widehat{R}_{s/s,p}$ konstansokról mindazon $(s,p)$ párokra, melyekre $s \leq 4$ és $p \geq 2$.

Az alkalmazott technikák és a felmerülő nehézségek illusztrálása céljából e rész lezárásaként az alábbi két eredményt emeljük ki a [9] cikkből.

1. A táblázatokban szereplő legmagasabb fokú algebrai szám az $R_{4/4,7} \approx 2.743911$ konstans, amely a

$\{4191472, 370695456, 15701398968, 423572490288, 8166117227943, 119697694352106, 1385540391042992, 12982888147808790,$
$100093600309232610, 641253042735937920, 3429070908298155495, 15292307228951973150, 56488604555080263600,$
$170258230386661619700, 405691319555093173950, 698620766882164002000, 475967180133964116000,$
$-276873748560736884000 0, -191713640990944618440 00, -831778993836939153600 00, -273285810570616506720000,$
$-658873655023431060000000, -107607729252613818600 0000, -103942980631680422400 0000, -402385174364630880000000,$
$163249595900852832000 0000, 151710362848315152000 0000, 634752788122252800000 00000, 141035850392839718400000000,$
$165407760061818624000000000, 819124663931118720000 00000\}$

30-adfokú polinom gyöke. Ez a polinom az alábbi meggondolásokból származtatható. A $\Pi_{4/4,7}$ osztály elemei alkalmas $a \in \mathbb{R}$ paraméterrel a

$$\psi_a(z) = \frac{\frac{1}{840}(7a+4)z^4 + \frac{1}{210}(21a+13)z^3 + \left(-\frac{1}{14}(7a+2) + a + \frac{1}{2}\right)z^2 + (a+1)z + 1}{-\frac{1}{840}(7a+3)z^4 + \frac{1}{210}(21a+8)z^3 - \frac{1}{14}(7a+2)z^2 + az + 1}$$

alakban írhatók fel. A feladat tehát annak az $a$ paraméterértéknek a megkeresése, amelyhez tartozó $\psi_a$ függvény összes deriváltja nemnegatív egy $[-r, 0]$ intervallumon, és az $r > 0$ érték a lehető legnagyobb. A [26] cikk (természetes normálási és irreducibilitási feltételek mellett) kétféle típusú felső korlátot fogalmaz meg egy konkrét racionális törtfüggvény $r$ abszolút monotonitási sugarának nagyságára:

(∗) $r \le B$, ahol a $B \ge 0$ (itt nem definiált) mennyiség csak a racionális törtfüggvény pólusainak geometriai elhelyezkedésétől függ a komplex síkon, illetve

(∗∗) $r \le \varrho$, ahol $-\varrho \le 0$ gyöke a racionális törtfüggvény *valamelyik* deriváltjának.

Esetünkben ezen állításoknak az $a$ paramétertől függő változatait alkalmaztuk. Kiderül, hogy a $\Pi_{4/4,7}$ osztályban az $a \mapsto B(a)$ valós függvény maximális értéke dönt. A $B(a)$ kifejezés esetszétválasztással van megadva, és paramétertől függő polinomok gyökeivel írható le; a teljes értelmezési tartományon a $B$ függvény csak szakaszonként monoton, deriválható, vagy konvex/konkáv. Itt a részfeladatok számítógépes bizonyításai már jóval nehezebbek: a *Mathematica* számára több másodpercig tartottak (vö. a 3.6. megjegyzésben szereplő néhány tizedmásodperccel). Amikor ezeknek a részállításoknak a bizonyításait elemi lépésekre lebontva ember által olvasható formában reprodukáltuk, az így adódó PDF fájl majdnem 300 oldalt foglalt el. Ez a méret annak is köszönhető, hogy a közbenső számításokban fellépő polinomok fokszáma gyakran több száz volt, együtthatóinak nagyságrendje pedig gyakran a $10^{310}$ tartományba esett. Az igazán meglepő dolog tehát az, hogy az $R_{4/4,7} \approx 2.743911$ algebrai szám fokszáma mindössze 30: a számításokban fellépő magas fokú polinomokat szerencsés módon szorzattá lehetett bontani.

2. A cikkben kidolgozott legbonyolultabb eset az $\widehat{R}_{4/4,2} = 8$ optimális érték bizonyítása. A $\widehat{\Pi}_{4/4,2}$ osztálybeli függvények a

$$\psi_{a,c,d}(z) = \frac{1 + \left(1 - a\binom{4}{1}\right)z + \left(\frac{1}{2} - a\binom{4}{1} + a^2\binom{4}{2}\right)z^2 + cz^3 + dz^4}{(1-az)^4} \tag{51}$$

alakban írhatók fel, alkalmas $a, c, d \in \mathbb{R}$ paraméterekkel. Kiderül, hogy ebben a háromparaméteres családban a maximális $r = 8$ abszolút monotonitási sugarat nem az előző pont (∗) részében említett $B$ függvény, hanem (∗∗) alapján a $\psi_{a,c,d}$ racionális törtfüggvény alkalmas deriváltjának egyik gyöke jelöli ki. A bizonyítás során konkrétan azt mutatjuk meg, hogy ha minden $k \in \mathbb{N}$ esetén

$$\psi^{(k)}_{a,c,d}(-8) \ge 0, \tag{52}$$

akkor

$$a = \frac{1}{8}, \quad c = \frac{1}{128} \quad \text{és} \quad d = \frac{1}{4096} \tag{53}$$

kell fennálljon – vagyis az $r = 8$ abszolút monotonitási sugárral bíró racionális törtfüggvény ebben a családban éppen az az egyértelműen meghatározott függvény, amelyet $s = 4$-re az (50) képlet szolgáltat. Az unicitás igazolásakor az (52) – végtelen sok tagú – egyenlőtlenségrendszerből azt a szükséges feltételt kapjuk, hogy minden $k \in \mathbb{N}^+$ mellett a

$$\begin{aligned}
&k^3\left(6a^4 - 6a^3 + a^2 + 2ac + 2d\right) + \\
&k^2\left(-384a^5 + 324a^4 - 42a^3 - 144a^2c - 6ac - 192ad - 12d\right) + \\
&k\left(4608a^6 - 3072a^5 + 618a^4 + 2304a^3c - 36a^3 + 144a^2c + 4608a^2d - \right. \\
&\qquad \left. a^2 + 4ac + 576ad + 22d\right) + \\
&4608a^6 - 2688a^5 - 6144a^4c + 300a^4 - 24576a^3d - 4608a^2d - 384ad - 12d
\end{aligned} \tag{54}$$

polinomnak nemnegatívnak kell lennie. A részszámítások során az $a \in \mathbb{R}$ paraméteregyenest alkalmas részekre bontjuk fel, és lépésről lépésre minden $a \in \mathbb{R} \setminus \{\frac{1}{8}\}$ értéket kizárunk. Például az $a \in \left[\frac{1}{24}\left(9+2\sqrt{6}\right), \frac{22}{10}\right] \approx [0.579, 2.2]$ paraméterintervallum kizárásához azt mutatjuk meg, hogy alkalmas $k \geq 4$ egész esetén az (54) polinom negatív. Meglepő módon (a $c, d$ paraméterekre fennálló bizonyos megszorítások mellett) a $k \in [4, 1000] \cap \mathbb{N}$ intervallumban az (54) polinom egyedül $k = 54$ esetén negatív. Fontos hangsúlyozni, hogy a fenti szükséges feltételben $k$ csak egész értéket vehet fel: az optimális (53) esetben ugyanis az (54) polinom bizonyos pozitív *nemegész* $k$ értékekre *negatív* értéket is felvesz. Az (54) polinom nemnegativitására vonatkozó, kezdetben 30 oldalas bizonyítást fokozatos egyszerűsítések után a cikk publikált változatában sikerült mindössze 7 oldalon leírni.

**3.18. megjegyzés.** *Még egyszer kiemeljük, hogy a cikkben nagy gondot fordítottunk arra, hogy a legfontosabb állítások számítógépes bizonyításait hagyományos bizonyítások formájába átírjuk. Ez bizonyításonként gyakran több hetes munkával járt. Az ilyen átírások során néhány alkalommal hibát (*bug*) fedeztünk fel bizonyos Mathematica-parancsok működésében – például paraméterektől függő többváltozós polinomiális egyenletrendszerek megoldása során speciális paraméterkombinációk esetén (tehát nagyon kis valószínűséggel) a korábbi számításainkkal nem összeegyeztethető kimenetet kaptunk. Ezeket a hibákat a fejlesztőknek jeleztük, akik készségesen, 1–2 napon belül javítókódokkal reagáltak megkeresésünkre – a javítások pedig a Mathematica következő verzióiba már véglegesen bekerültek.*

### 3.4.2. Többlépéses módszerek korlátossági lépésköz-együtthatóinak egzakt optimális értéke

A 2.4. szakaszban definiáltuk egy LT-módszer SSP-együtthatóját, azaz általános *monotonitási* lépésköz-együtthatóját. Szintén ott szerepelt ennek általánosítása, az általános *korlátossági* lépésköz-együttható, amely lépésköz-megszorítás az LT-módszer által generált sorozat korlátosságát garantálja az általános (lineáris vagy nemlineáris) (4) problémaosztályban. Ebben a szakaszban az általános korlátossági lépésköz-együtthatóra (angol neve alapján) az *SCB-együttható* rövidítéssel hivatkozunk. A gyakorlat szempontjából fontos kérdés, hogy adott LT-módszer esetén hogyan lehet

(*i*) eldönteni, hogy létezik-e pozitív $\gamma > 0$ SCB-együttható;

(*ii*) eldönteni, hogy egy adott pozitív szám SCB-együttható-e;

(*iii*) meghatározni a legnagyobb pozitív $\gamma_{\text{sup}}$ SCB-együtthatót.

Ahogyan látni fogjuk, az SCB-együtthatóra vonatkozó fenti kérdések jóval nehezebbek az SSP-együtthatóra vonatkozó analóg kérdéseknél, hiszen ez utóbbiakat a (38)–(39) formulák egyszerűen megválaszolják.

A [38] cikk az SCB-együtthatókra vonatkozó fenti (*ii*) pontra ad választ; a megfogalmazott karakterizáció lényege az alábbi. Tegyük fel, hogy az LT-módszer teljesít néhány természetes feltételt (például konzisztencia és irreducibilitás), és legyen adott egy pozitív $\gamma > 0$ konstans. Ekkor $\gamma$ pontosan akkor SCB-együttható, ha

$$-\gamma \in \text{int}(\mathcal{S}), \text{ és } \mu_n(\gamma) \geq 0 \text{ minden } n \in \mathbb{N}^+ \text{ mellett}, \tag{55}$$

ahol $\text{int}(\mathcal{S})$ jelöli az LT-módszer abszolút stabilitási tartományának belsejét (lásd korábban az 1.3. szakaszban a (19) képletet), és a $\mu_n(\gamma)$ sorozatot az LT-módszer (18)-beli $(\alpha_j, \beta_j)$ együtthatóival a

$$\mu_n(\gamma) := \begin{cases} 0 & \text{ha } n < 0, \\ \beta_n - \gamma\,\beta_0\mu_n(\gamma) + \sum_{j=1}^{k}(\alpha_j - \gamma\,\beta_j)\mu_{n-j}(\gamma) & \text{ha } 0 \leq n \leq k, \\ -\gamma\,\beta_0\mu_n(\gamma) + \sum_{j=1}^{k}(\alpha_j - \gamma\,\beta_j)\mu_{n-j}(\gamma) & \text{ha } n > k \end{cases} \tag{56}$$

rekurzióval értelmezzük.

Sajnos az (55) kritérium elvi szempontból nem tűnik alkalmasnak arra, hogy vele a fenti (*i*) pontot is megválaszoljuk, hiszen ha például *nem* létezik pozitív $\gamma$ SCB-együttható, akkor ennek kimutatásához az

(55) feltétel végtelen sokszori alkalmazására lenne szükség (egy $0^+$-hoz tartó $\gamma$ sorozat mentén), ráadásul a kritérium minden egyes alkalmazásakor (azaz rögzített $\gamma$-ra) külön-külön végtelen sok egyenlőtlenséget kellene ellenőrizni ($\mu_n(\gamma) \geq 0$ minden $n$-re).

Ezt a problémát orvosolja a [22] cikkben szereplő módszer, amely az LT-módszer (18) együtthatóinak segítségével egy $\tau_n$ rekurziót definiál, melynek pozitivitásával lehet lényegében eldönteni az $(i)$ kérdést. A $\tau_n$ rekurzió előnye, hogy nem függ a $\gamma$ paramétertől, így a $\mu_n(\gamma)$ sorozatnál egyszerűbben vizsgálható és kezelhető. Ez alapján a [22] cikk LT-módszerek több fontos osztályában – például az Adams–Moulton, Adams–Bashforth, BDF, extrapolált BDF, Milne–Simpson vagy a Nyström-módszerek esetén – megvizsgálja az SCB-együttható létezését. Itt többször azt tapasztaljuk, hogy bizonyos LT-módszereknek nincs pozitív SSP-együtthatója, de van pozitív SCB-együtthatója, azaz általános monotonitással nem, viszont általános korlátossági tulajdonsággal rendelkeznek ezek a módszerek.

Korábban kevés LT-módszer esetén volt ismert a $\gamma$ SCB-együtthatók lehetséges értéke. A [2] dolgozatban – építve a [9] cikkben kifejlesztett technikákra – a fenti $(iii)$ kérdés megválaszolására fektetjük a hangsúlyt

- a BDF-módszerek (implicit), illetve
- az Adams–Bashforth-módszerek (explicit)

családjában, egzaktul, algebrai számként megadva a lehető legnagyobb $\gamma_{\text{sup}}$ SCB-együtthatót. (Ezen kívül formálisan bizonyítunk néhány állítást a [22] cikkből, melyeket ott csak numerikus kísérletek támasztottak alá.) Adott LT-módszer esetén az (55) karakterizáció értelmében célunk tehát lényegében a következő: megkeresni a legnagyobb $\gamma_{\text{sup}} > 0$ értéket úgy, hogy minden $\gamma \in (0, \gamma_{\text{sup}})$ és minden $n \in \mathbb{N}$ mellett az (56) rekurzió tagjaira $\mu_n(\gamma) \geq 0$ teljesüljön.

**3.19. megjegyzés.** *Egyszerűen látható, hogy explicit LT-módszer és rögzített $n$ esetén a $\mu_n(\cdot)$ függvények polinomok, míg implicit LT-módszer esetén racionális törtfüggvények.*

**3.20. megjegyzés.** *A magasabb rendű lineáris rekurziók nemnegativitásának eldöntése az általános esetben nehéz és megoldatlan feladat, amely mély számelméleti állítások (különféle diofantikus approximációk) igazságán múlna. Az ezzel kapcsolatos frissebb eredményeket a [44, 45, 46, 47] publikációk foglalják össze.*

A [2] cikkbeli eredmények illusztrálásához tekintsük először például a $k = 5$ lépéses BDF-módszert (lásd a (20) formulát az 1.3. szakaszban). Ekkor az (56) rekurzió az alábbi alakot ölti:

$$(60\gamma + 137)\mu_n(\gamma) - 300\mu_{n-1}(\gamma) + 300\mu_{n-2}(\gamma) - 200\mu_{n-3}(\gamma) +$$

$$75\mu_{n-4}(\gamma) - 12\mu_{n-5}(\gamma) = 0 \qquad (n \geq 5),$$

a megfelelő kezdőértékek pedig

$$\mu_0(\gamma) = \frac{60}{60\gamma + 137}, \quad \mu_1(\gamma) = \frac{18000}{(60\gamma + 137)^2}, \quad \mu_2(\gamma) = \frac{18000(-60\gamma + 163)}{(60\gamma + 137)^3},$$

$$\mu_3(\gamma) = \frac{12000\left(3600\gamma^2 - 37560\gamma + 30469\right)}{(60\gamma + 137)^4} \quad \text{és}$$

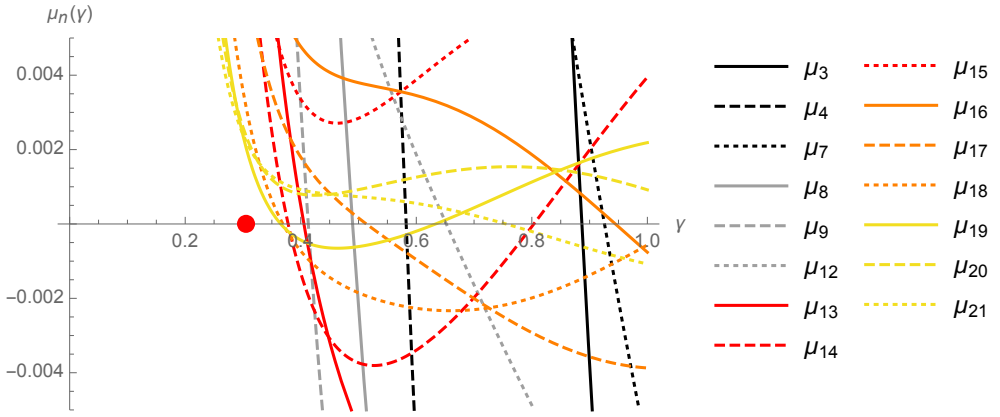
$$\mu_4(\gamma) = \frac{4500\left(-216000\gamma^3 + 8600400\gamma^2 - 22146420\gamma + 10021847\right)}{(60\gamma + 137)^5}.$$

A 6. ábrán az első néhány $\mu_n(\cdot)$ függvény grafikonját látjuk. A 7. ábra analóg szituációt mutat, csak más „nézőpontból”: a 6-lépéses BDF-módszer optimális paraméterértékéhez tartozó $\mu_n(\gamma_{\text{sup}})$ sorozatát tünteti fel $n$ függvényében; ez a $\gamma_{\text{sup}} \approx 0.131359$ konstans egyébként a

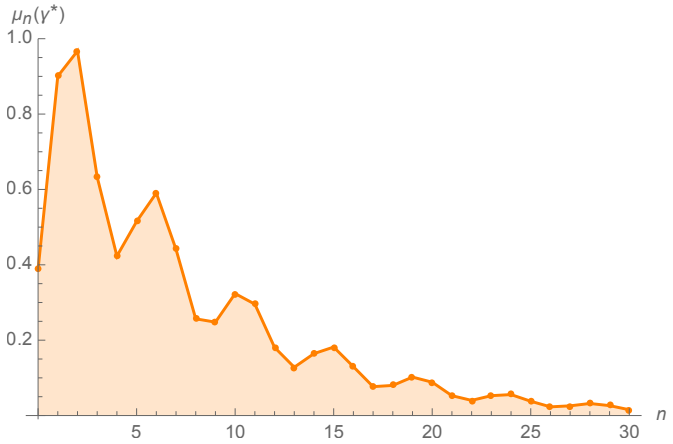
{3014991538380452755283116032000000000, 1226395855345048398189452014387200000000,
384963168041618344234237602954215424000000, 2754957003308188522312802320744458485760000,
6883218301719049493344792020881093689344000000, −384146941872396676115776998321179378948505600,

1148435884877509023231036682498035997863051264 00, −1006269459507863531788997342497299304467812843520,
5587246198359348966734174906666273788289332150272, −1742994479585896501088299686807315532951483940864 0,
3595911414144309586488624075051788478749789743104 0, −5335782722513254244314532744202925053609886368768 0,
5877907847072023567714364851996852450433631890560 0, −4811713104065419274087788780168854930357866871206 4,
2880915319585617372631296769697616863391766202424 0, −1215853010152056609922124822634701943275606226224 0,
3383327891741061214240426918034255832010259451480, −5413708008781257125916105851451946595223788968 80,
3332809264118625455076024766116814876826293706 7}

18-adfokú polinom egyik gyöke.




6. ábra: az 5-lépéses BDF-módszerhez tartozó, és az (56) rekurzió által meghatározott $\gamma \mapsto \mu_n(\gamma)$ racionális törtfüggvények $1 \le n \le 21$ esetén – az $n \in \{1, 2, 5, 6, 10, 11\}$ indexekhez tartozó grafikonok az ábrán kívülre esnek. A piros pont az optimális lépésköz-együttható $\gamma_{\text{sup}} \approx 0.30421$ értékét mutatja; ennek az algebrai számnak a foka 10. A $\gamma \in (0, \gamma_{\text{sup}})$ intervallumban minden $n \in \mathbb{N}$ mellett $\mu_n(\gamma) \ge 0$, és a $(0, \gamma_{\text{sup}})$ intervallum ebből a szempontból maximális.




7. ábra: a 6-lépéses BDF-módszerhez tartozó, és az (56) rekurzió által meghatározott $n \mapsto \mu_n(\gamma^*)$ sorozat (lineáris interpoláltja) a $\gamma^* = \gamma_{\text{sup}} \approx 0.131359$ optimális lépésköz-együttható esetén, amely egy 18-adfokú algebrai szám. Az optimalitás itt is azt jelenti, hogy tetszőleges $\varepsilon > 0$ mellett alkalmas $n \in \mathbb{N}$ indexszel $\mu_n(\gamma_{\text{sup}} + \varepsilon) < 0$.

A bizonyítások során az egyes $\mu_n$ rekurziók tagjait „explicit” formában írjuk fel az (56) rekurzió $\mathcal{P}(\cdot, \gamma)$ karakterisztikus polinomjának gyökeivel – természetesen minden kifejezés a $\gamma > 0$ paraméter függvénye –, majd e karakterisztikus gyökök „útvonalát” követjük nyomon a komplex síkon, ahogyan $\gamma$ értéke változik. A vizsgálatok során például az alábbi két elemi állítás bizonyult hasznosnak.

**3.21. állítás.** *Rögzítsünk egy egységnyi abszolút értékű $z \in \mathbb{C} \setminus \mathbb{R}, |z| = 1$ számot, egy $w \in \mathbb{C} \setminus \{0\}$ számot, és egy* valós $\nu_n \to 0$ *($n \to +\infty$) nullsorozatot. Ekkor végtelen sok $n \in \mathbb{N}$ esetén $wz^n + \bar{w}(\bar{z})^n + \nu_n < 0$.*

**3.22. állítás.** *Ha valamely $n \in \mathbb{N}^+$ és $\gamma^* > 0$ esetén $\mu_n(\gamma^*) = 0$ és $\mu_n'(\gamma^*) \in \mathbb{R} \setminus \{0\}$ (itt a vessző a $\mu_n(\cdot)$ függvény deriváltját jelöli), akkor $\gamma_{\sup} \leq \gamma^*$.*

A [2] cikk megállapításait az alábbiakban foglaljuk össze. A BDF- és az Adams–Bashforth-családban az optimális $\gamma_{\sup}$ lépésköz-együtthatókat – a fenti állításoknak megfelelő – alábbi két feltétel valamelyike karakterizálja:

**a)** a $\mu_n(\gamma)$ rekurzió karakterisztikus polinomjának domináns (azaz maximális abszolút értékű) gyöke $\gamma \in (0, \gamma_{\sup})$ mellett *pozitív* valós, de e gyök domináns tulajdonsága elvész, amint $\gamma$ átlépi $\gamma_{\sup}$-ot;

**b)** van olyan $n_0 \in \mathbb{N}$ index, hogy $\gamma_{\sup}$ a $\mu_{n_0}(\cdot)$ függvény egyszeres gyöke.

Kiderül, hogy $\gamma_{\sup}$ értékét

- a $k$-lépéses BDF-családban $k \in \{2, 4, 5, 6\}$ esetén az **a)** feltétel;
- a $k$-lépéses BDF-családban $k = 3$ esetén $n_0 = 6$-tal a **b)** feltétel;
- a $k$-lépéses Adams–Bashforth-családban $k \in \{1, 2, 3\}$ esetén pedig $n_0 = 2$-vel a **b)** feltétel

határozza meg.

**3.23. megjegyzés.** *A fenti **a)** és **b)** feltételek a 3.4.1. szakasz végén, az* 1. *-ben ismertetett (*) és (**) feltételek analogonjainak tekinthetők.*

**3.24. megjegyzés.** *A $k = 3$ lépéses BDF-módszerre $\gamma = 5/6 \approx 0.83333$ esetén az (56) rekurzió $\mathcal{P}_3(\cdot, \gamma)$ harmadfokú karakterisztikus polinomjának egyik gyöke $\varrho_1(\gamma) > 0$ pozitív valós, másik két gyöke $\varrho_{2,3}(\gamma)$ konjugált komplex, és $|\varrho_1(\gamma)| = |\varrho_2(\gamma)| = |\varrho_3(\gamma)|$; a pozitív valós gyök ennél a paraméterértéknél veszti el domináns tulajdonságát. Ebből a 3.21. állítással a $\gamma_{\sup} \leq 5/6$ felső becslést kapjuk, mégsem ez lesz $\gamma_{\sup}$ pontos értéke. Ahogyan fent említettük, a pontos értéket itt a 3.22. állítás adja $n_0 = 6$-tal, vagyis $\gamma_{\sup} = \gamma^*$, ahol $\gamma^* \approx 0.83126$ (egy 4-edfokú algebrai szám) a $\mu_6(\cdot) = 0$ egyenlet megoldása. Ennél a paraméterértéknél a $\mathcal{P}_3(\cdot, \gamma^*)$ polinom gyökeire fennáll, hogy $\varrho_1(\gamma^*) \approx 0.500518$ és $|\varrho_{2,3}(\gamma^*)| \approx 0.499935$ (vagyis a domináns gyök még pozitív valós, a másik két gyök komplex), de abszolút értékben $\varrho_1$ és $\varrho_{2,3}$ közel vannak egymáshoz. Ez a közelség $\gamma^*$ optimalitásának bizonyítása során 1-hez közeli kvóciensű mértani sorozatokat eredményez: a $\mu_n(\gamma^*) > 0$ egyenlőtlenség természetes módon következik a*

$$2 \cdot \frac{2777}{10000} \left( \frac{9989}{10000} \right)^n < \frac{50155}{100\,000}$$

*elégséges feltételből, amely csak $n \geq 93$ mellett teljesül. A $\mu_n(\gamma^*) > 0$ egyenlőtlenségeket tehát $n \in \{0, 1, \ldots, 92\} \setminus \{6\}$ esetén külön ellenőrizni kell – az utolsó indexre például $\mu_{92}(\gamma^*) \approx 1.585176 \cdot 10^{-28}$.*

**3.25. megjegyzés.** *A $k = 4$ lépéses BDF-módszer esetén az optimális $\gamma_{\sup} \approx 0.48622$ érték megtalálása (amely egy 5-ödfokú algebrai szám) nem bizonyult egyszerűnek, ugyanis a munkának ebben a fázisában még nem kristályosodott ki a 3.21. állítás. Különböző $\gamma$ értékekkel tesztelve a $\mu_n(\gamma)$ sorozat nemnegativitását, $\gamma_{\sup}$ felső becsléseit fokozatosan lejjebb és lejjebb szorítottuk. A $\gamma_{\sup} < 0.48625$ korlát belátásához például az $1 \leq n \leq 27000$ tartományban vizsgálódtunk, és azt kaptuk, hogy ezekre az $n$ értékekre*

$$\mu_n(48625/100000) < 0 \Longleftrightarrow n \in \{26814, 26875, 26886, 26936, 26947, 26997\}.$$

*A fenti 6 index megtalálásához a $\mu_n(48625/100000)$ sorozatot nagy pontossággal kellett kiértékelni, melyhez a Mathematica tetszőleges pontosságú és adaptív aritmetikáját használtuk: tagonként 15000 tizedesjegy pontosság nem volt elegendő, de 16000 már igen. Valójában az ilyen és ehhez hasonló* numerikus kísérletekből *született a 3.21. állítás.*

## 3.5. Egy- és többlépéses diszkretizációk stabilitási tartománya

Ebben a szakaszban bizonyos RK-módszerek abszolút stabilitási tartományát vizsgáljuk, a [7] cikk főbb gondolatait ismertetve. Analóg kérdések vethetők fel LT-módszerek abszolút stabilitási tartományai alakjának pontosabb meghatározásával kapcsolatban (például az $A(\alpha)$-stabilitással kapcsolatban) – az elmúlt hónapokban született ezirányú érdekes eredmények publikálása folyamatban van.

### 3.5.1. Az exponenciális függvény Taylor-sorának részletösszegei

Ahogyan azt a 3.2. szakaszban említettük, az extrapolációs módszerek stabilitásának vizsgálata az ott definiált belső stabilitási polinomok abszolút értékének alsó-, illetve felső becslését igényli. Ezek a becslések a módszer abszolút stabilitási tartományán, illetve annak bizonyos részhalmazain (például a bal komplex félsíkba eső részeken) érdekesek. Kiderül, hogy a szóban forgó abszolút stabilitási tartományokat az exponenciális függvény origó közepű Taylor-sorának részletösszegei határozzák meg: $n \in \mathbb{N}^+$ esetén jelölje ezeket a halmazokat

$$\mathcal{U}_n := \left\{ z \in \mathbb{C} : \left| \sum_{k=0}^{n} \frac{z^k}{k!} \right| \leq 1 \right\}, \tag{57}$$

lásd a 8. ábrát. A fenti *nem skálázott* (*unscaled*) részletösszegek helyett néha kényelmes a nekik megfelelő *skálázott* (*scaled*)

$$\mathcal{S}_n := \left\{ z \in \mathbb{C} : \left| \sum_{k=0}^{n} \frac{(nz)^k}{k!} \right| \leq 1 \right\} \tag{58}$$

halmazokat vizsgálni, lásd a 9. ábrát. Az $\mathcal{U}$ és $\mathcal{S}$ halmazok egyébként szoros kapcsolatban állnak a részletösszeg-polinomok zérushelyeinek

$$\mathcal{Z}_n := \left\{ z \in \mathbb{C} : \sum_{k=0}^{n} \frac{(nz)^k}{k!} = 0 \right\} \tag{59}$$

halmazával is, melyek aszimptotikus leírásakor felbukkan a

$$\Sigma_1 := \{ z \in \mathbb{C} : |ze^{1-z}| \leq 1 \} \cap D_1 \tag{60}$$

kompakt halmaz; itt és a továbbiakban $D_\varrho$ jelöli a komplex sík origó középpontú és $\varrho > 0$ sugarú zárt körlapját. A $\Sigma_1$ halmaz $\partial\Sigma_1$ határgörbéjét Szegő-görbének[6] hívják, lásd a 10. ábrát. Az $\mathcal{U}_n$, $\mathcal{S}_n$ vagy a $\mathcal{Z}_n$ a halmazok – a numerikus analízistől függetlenül – természetesen önmagukban is fontos és sokat tanulmányozott klasszikus objektumok; az elmúlt száz évben több tucat szerző vizsgálta őket, többek között Szegő 1924-ben [48], Buckholtz az 1960-as években [49], Jeltsch és Nevanlinna az 1980-as években [50], vagy Varga[7] a 2000-es években. A további eredmények ismertetéséhez szükségünk lesz még néhány egyszerű jelölésre: a zárt bal komplex félsíkot, illetve a képzetes tengelyt jelölje rendre $\{\mathrm{Re} \leq 0\}$, illetve $\{\mathrm{Re} = 0\}$; a $\{\mathrm{Re} < 0\}$ és $\{\mathrm{Re} > 0\}$ szimbólumok jelentése magától értetődő.

Az [50] cikkben a szerzők az $\mathcal{S}_n$ halmazok aszimptotikus viselkedését vizsgálva például belátják, hogy

$$\mathcal{S}_\infty = \big( D_{1/e} \cap \{\mathrm{Re} \leq 0\} \big) \cup \partial\Sigma_1, \tag{61}$$

lásd ismét a 10. ábrát, ahol az $\mathcal{S}_\infty$ halmazra gondoljunk úgy, mint az $\mathcal{S}_n$ halmazok „határértékére" (itt ezt a fogalmat nem definiáljuk pontosabban). A numerikus analízis gyakorlati alkalmazásaiban a (61) által szolgáltatott jellemzés azonban nem praktikus, hiszen a képlet nem mond semmit az $\mathcal{S}_n$ halmazok alakjáról *konkrét*, *kis* $n \in \mathbb{N}^+$ értékek esetén – a 3.2. szakaszban pedig pont ilyen eredményekre van szükség.

[6] Szegő Gábor (1895–1985)

[7] Richard Steven Varga (1928–)

A fentiek által motiválva a [7] dolgozatban a *Mathematica* segítségével egzakt algebrai számmal kifejezve megmérjük

• az $\mathcal{S}_n$ halmazok origótól vett maximális távolságát $1 \le n \le 20$ esetén, lásd a 3. táblázatot;

• az $\mathcal{S}_n$ halmazok $\{\mathrm{Re} \le 0\}$ bal félsíkba eső részének origótól vett maximális távolságát $1 \le n \le 20$ esetén, lásd a 4. táblázatot.

Ezek a táblázatok tehát – legalábbis $1 \le n \le 20$ esetén – megadják az

$$\mathcal{S}_n \subset D_{\varrho_n}, \text{ illetve az } \mathcal{S}_n \cap \{\mathrm{Re} \le 0\} \subset D_{\widetilde{\varrho}_n} \cap \{\mathrm{Re} \le 0\} \tag{62}$$

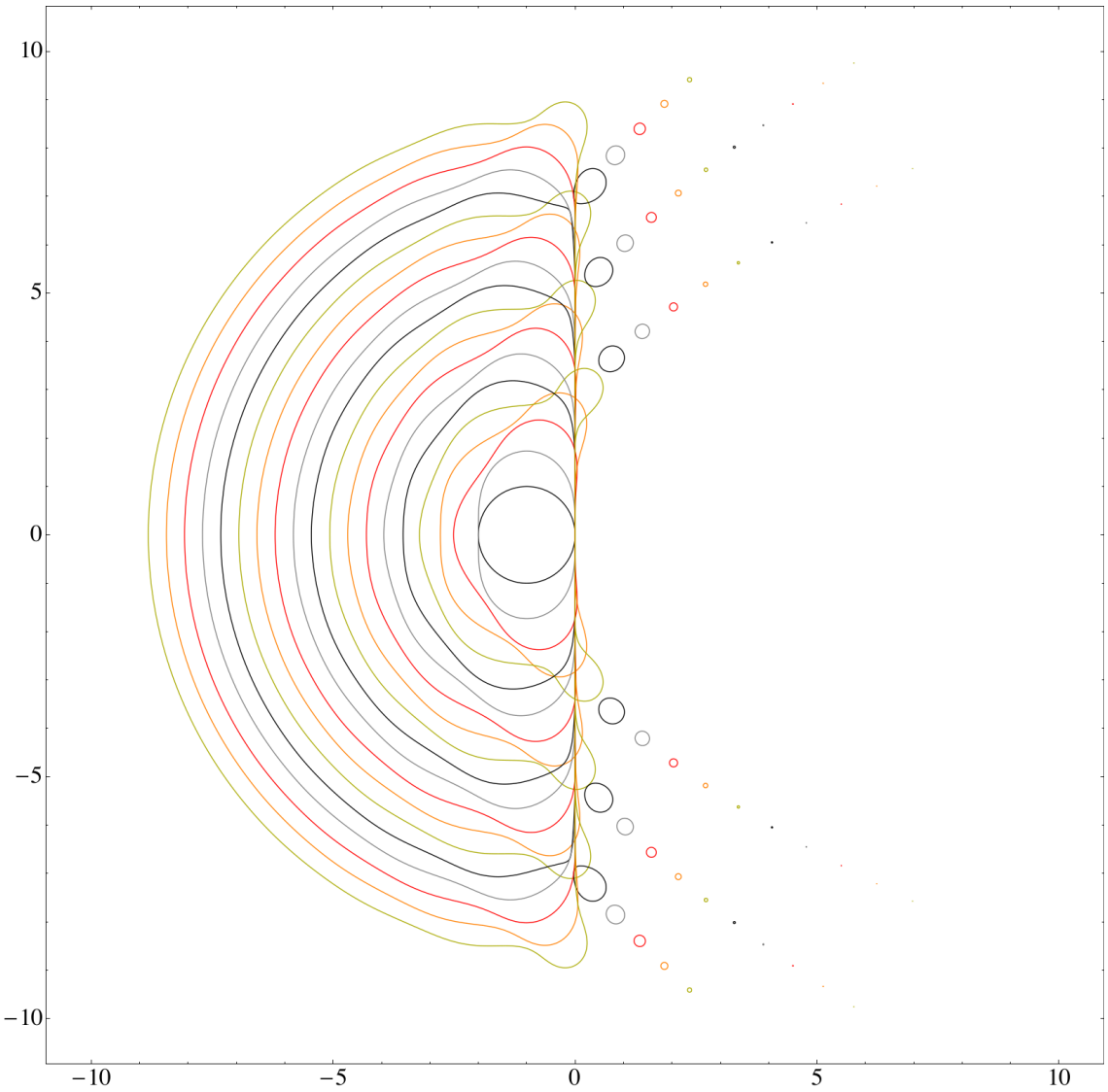
tartalmazásokban fellépő legkisebb lehetséges $\varrho_n, \widetilde{\varrho}_n > 0$ sugarak értékét.

| $n$ | $\max_{z\in\mathcal{S}_n} \lvert z\rvert$ | $n$ | $\max_{z\in\mathcal{S}_n} \lvert z\rvert$ |
|---|---|---|---|
| 1 | 2 | 11 | 0.664 |
| 2 | $\frac{1}{2}\sqrt{2\left(1+\sqrt{2}\right)} \approx 1.099$ | 12 | 0.670 |
| 3 | 0.847 | 13 | 0.676 |
| 4 | 0.741 | 14 | 0.682 |
| 5 | 0.690 | 15 | 0.687 |
| 6 | 0.665 | 16 | 0.692 |
| 7 | 0.6546 | 17 | 0.697 |
| 8 | 0.6523 | 18 | 0.702 |
| 9 | 0.6542 | 19 | 0.707 |
| 10 | 0.659 | 20 | 0.711 |

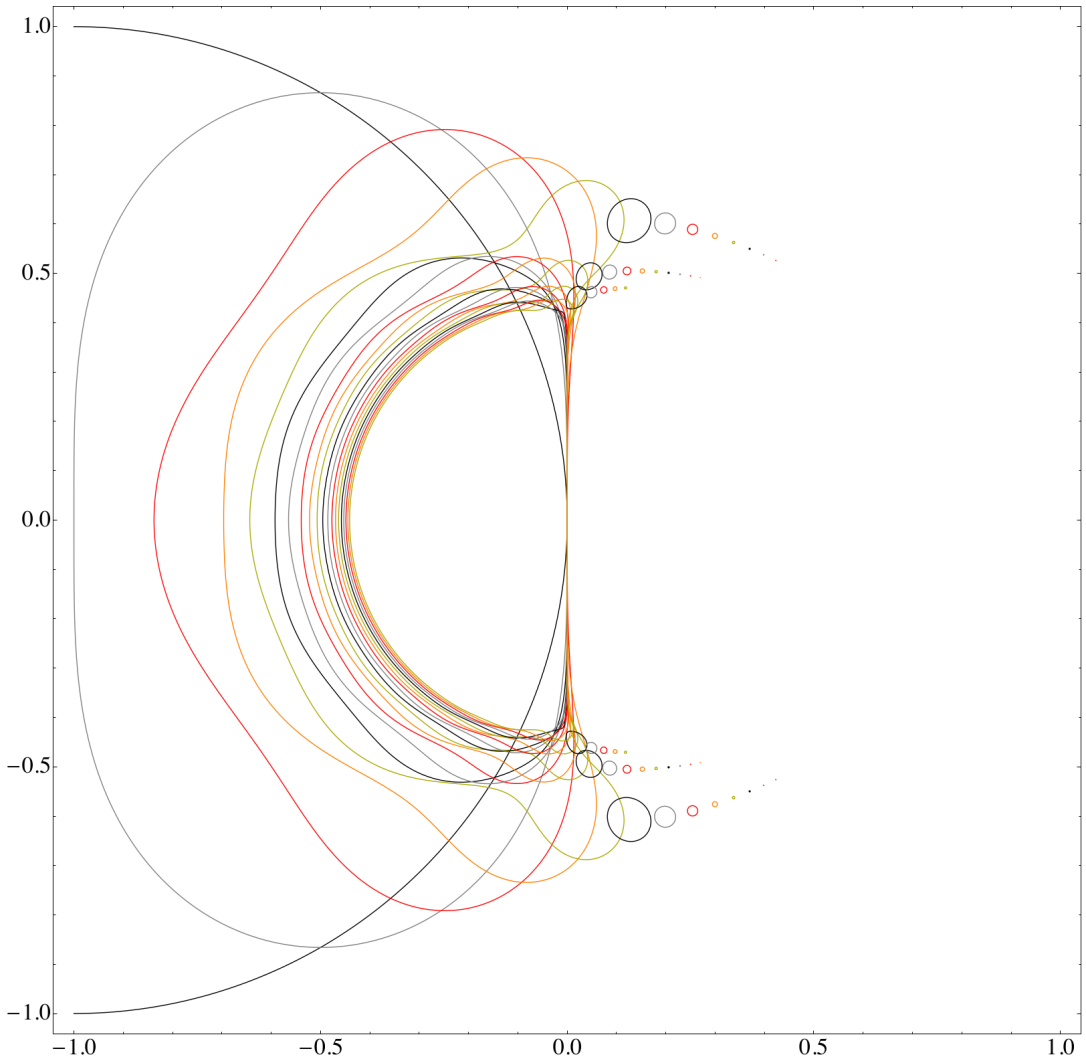
3. táblázat: az első néhány $\mathcal{S}_n$ halmaz origótól legtávolabbi pontjának távolsága. Az algebrai számok egzakt értékét $n \ge 3$-ra felfelé kerekítettük; az $n = 20$-hoz tartozó konstans például egy 760-adfokú polinom gyöke, és ezt a polinomot a *Mathematica* (laptopon futtatva) mindössze 40 másodperc alatt találta meg. Az $1 \le n \le 20$ intervallumban a táblázatbeli értékek minimuma $n = 8$-nál található; a maximum helye $1 \le n \le 4$ esetén a $\{\mathrm{Re} < 0\}$, míg $5 \le n \le 20$ esetén a $\{\mathrm{Re} > 0\}$ halmazba esik. Figyeljük meg, ahogyan a sorozat a (61) aszimptotikus ($n \to +\infty$) formula értelmében 1-hez közeledik.

| $n$ | $\max_{z\in\mathcal{S}_n\cap\{\mathrm{Re}\le 0\}} \lvert z\rvert$ | $n$ | $\max_{z\in\mathcal{S}_n\cap\{\mathrm{Re}\le 0\}} \lvert z\rvert$ |
|---|---|---|---|
| 1 | 2 | 11 | 0.496 |
| 2 | $\frac{1}{2}\sqrt{2\left(1+\sqrt{2}\right)} \approx 1.099$ | 12 | 0.486 |
| 3 | 0.847 | 13 | 0.480 |
| 4 | 0.741 | 14 | 0.476 |
| 5 | 0.680 | 15 | 0.474 |
| 6 | 0.597 | 16 | 0.458 |
| 7 | 0.566 | 17 | 0.453 |
| 8 | 0.546 | 18 | 0.450 |
| 9 | 0.534 | 19 | 0.449 |
| 10 | 0.527 | 20 | 0.448 |

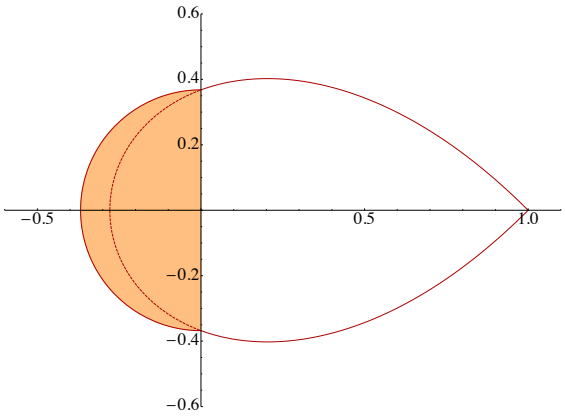
4. táblázat: néhány $\mathcal{S}_n$ halmaz bal félsíkba eső részének maximális távolsága az origótól. Az egzakt algebrai számokat $n \ge 3$ esetén itt is felfelé kerekítettük. A táblázatbeli értékek a (61) aszimptotikus ($n \to +\infty$) formula értelmében $1/e$-hez közelednek, ám itt monoton csökkenő módon.




8. ábra: az (57) képlettel definiált $\mathcal{U}_n$ halmazok határgörbéi $1 \leq n \leq 20$ esetén, a $-10 \leq \mathrm{Re}(z) \leq 10$ és $-10 \leq \mathrm{Im}(z) \leq 10$ ablakban. Az ábrán növekvő $n$ értékek mellett ciklikusan 5 színt használtunk (ez a 12. ábrán megfigyelhető bizonyos 5 vagy 6 hosszúságú részsorozatokkal van kapcsolatban). Általában nem ismert, hogy rögzített $n$-re az $\mathcal{U}_n$ halmaz hány összefüggő komponensből áll: a numerikus kísérletek szerint az $\mathcal{U}_n$ halmaz $1 \leq n \leq 5$ esetén összefüggő, $6 \leq n \leq 10$ esetén pedig 3, míg $n = 11$-re 5 komponensből áll. Kis $n$ értékek mellett ezek az ábrák a numerikus analízissel foglalkozó könyvekben megtalálhatók.




9. ábra: az (58) képlettel definiált $\mathcal{S}_n$ halmazok határgörbéi $1 \leq n \leq 20$ esetén a $-1 \leq \mathrm{Re}(z) \leq 1$, $-1 \leq \mathrm{Im}(z) \leq 1$ négyzetben. Rögzített $n$ esetén a görbék színe megegyezik a 8. ábra megfelelő görbéjének színével.




10. ábra: az itt látható „csepp” alakú görbe a (60) képlettel definiált $\Sigma_1$ halmaz $\partial\Sigma_1$ határgörbéje, a Szegő-görbe. A bal oldali narancssárga, $1/e$ sugarú félkörlappal együtt e két halmaz alkotja a (61) képletben szereplő $\mathcal{S}_\infty$ halmazt, amely az $\mathcal{S}_n$ halmazok „határértéke”.

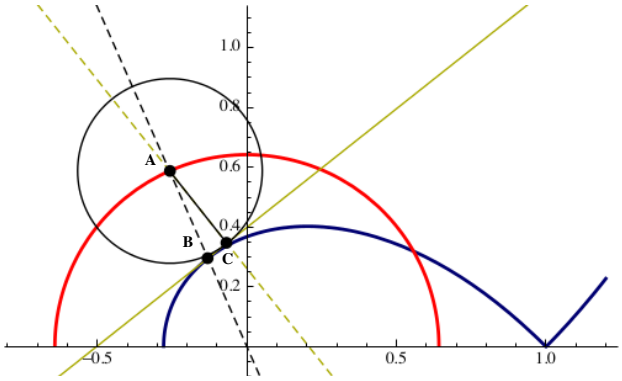
Természetes módon felmerül a kérdés, hogy mi mondható nagyobb $n$-ek esetén a (62)-beli tartalmazásokról. Rögzített $n \in \mathbb{N}^+$ esetén az (59) definícióban szereplő skálázott Taylor-polinom együtthatóinak monotonitási tulajdonsága miatt az Eneström–Kakeya-tétel értelmében e polinom összes gyöke a $D_1$ körlapban fekszik. Ebből az állításból egyszerűen adódik az alábbi tétel.

**3.26. tétel.** *Tetszőleges $n \in \mathbb{N}^+$ esetén $\mathcal{S}_n \subset D_2$.*

Buckholtz [49] az (59)-beli skálázott Taylor-polinom egy (itt nem definiált) $T_n$ transzformáltjára differenciálegyenletet írt fel, melynek segítségével *felülről* megbecsülte a $T_n$ polinomok nagyságát a $D_1$ halmaz komplementerén, illetve a (60)-beli $\Sigma_1$ Szegő-tartomány komplementerén. Ezt a technikát a komplex analízis elemi Cauchy-egyenlőtlenségével ötvözve, a $T_n$ polinomok *alsó* becslése adódik. A bizonyítás során a (60) által *implicit* módon megadott Szegő-tartományt a Lambert-féle $W$-függvénnyel (lásd (42) a 3.2. szakaszban) *expliciten* reprezentáljuk, és ennek segítségével becsüljük meg bizonyos körívek és a Szegő-görbe távolságát, lásd például a 11. ábrát. Ezekből kapjuk a következő effektív tartalmazási relációkat.

**3.27. tétel.** *Minden $\varepsilon \in (0,1)$ és $n \geq n_0(\varepsilon) := \left(\frac{1.0085\, e}{\varepsilon}\right)^2$ esetén $\mathcal{S}_n \subset D_{1+\varepsilon}$.*

**3.28. megjegyzés.** *A 3.27. tétel a (61) formula miatt aszimptotikusan optimális, vagyis az 1 konstans a $D_{1+\varepsilon}$ képletben nem helyettesíthető kisebb számmal. Másrészt $3 \leq n \leq 20$ esetén a 3. táblázatból tudjuk, hogy $\mathcal{S}_n \subset D_1$. Vajon igaz-e, hogy tetszőleges $n \geq 21$ esetén is fennáll az $\mathcal{S}_n \subset D_1$ tartalmazás?*




11. ábra: a 3.27. tétel bizonyításában többek között szükségünk van bizonyos (piros) körívdarabok és a (kék) Szegő-görbe távolságának alsó becslésére

Ezek után a 3. táblázat eredményeit kombinálva a 3.27. tétellel, a 3.26. tétel alábbi javítását nyerjük.

**3.29. tétel.** *Tetszőleges $n \geq 2$ egész mellett $\mathcal{S}_n \subset D_{1.6}$.*

Az $\mathcal{S}_n$ halmazok bal komplex félsíkba eső részének félkörlapokba foglalását az alábbi tartalmazásokkal írjuk le.

**3.30. tétel.** *Tetszőleges $\varepsilon > 0$ számhoz van olyan $n_0(\varepsilon) \in \mathbb{N}^+$ index, hogy $n \geq n_0(\varepsilon)$ esetén*

$$\mathcal{S}_n \cap \{\mathrm{Re} \leq 0\} \subset D_{1/e+\varepsilon} \cap \{\mathrm{Re} \leq 0\}.$$

Ezen állítás egy effektív változata az alábbi aszimptotikusan optimális eredmény.

**3.31. tétel.** *Rögzítsünk egy tetszőleges $\delta > 0$ számot és legyen $\widetilde{\varrho}_n := \frac{1}{e} + \frac{(2+\delta)\sqrt{e^2+1}}{\sqrt{n}}$. Ekkor tetszőleges $n \in \mathbb{N}^+$ esetén*

$$\mathcal{S}_n \cap \{\mathrm{Re} \leq 0\} \subset D_{\widetilde{\varrho}_n} \cap \{\mathrm{Re} \leq 0\}.$$

A 3.29. tétel $\{\mathrm{Re} \le 0\}$-ra vonatkozó változata az alábbi állítás.

**3.32. tétel.** *Tetszőleges $n \ge 3$ mellett*

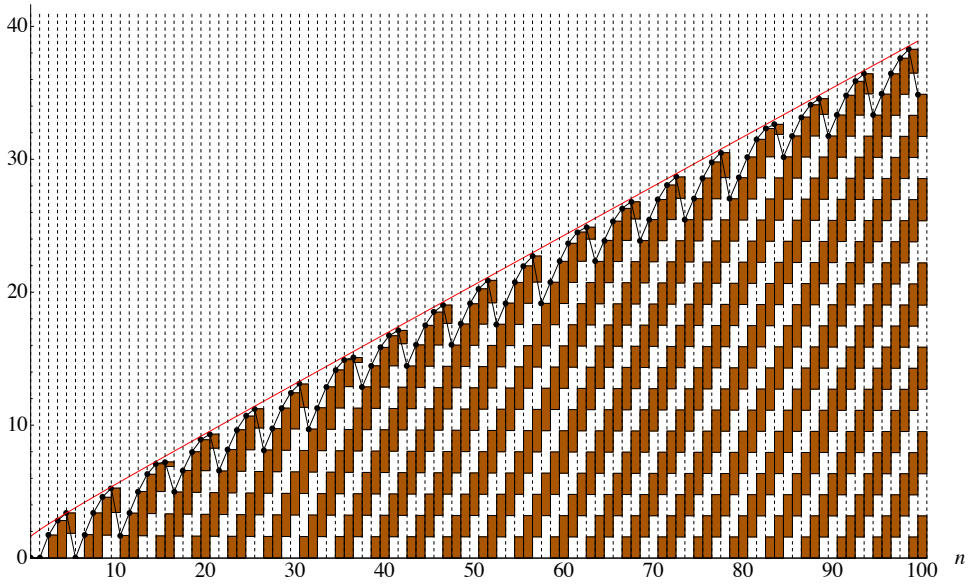
$$\mathcal{S}_n \cap \{\mathrm{Re} \le 0\} \subset D_{0.95} \cap \{\mathrm{Re} \le 0\}.$$

**3.33. megjegyzés.** *A tételben szereplő 0.95 konstans csökkenthető volna, ha feltesszük, hogy $n \ge n_0$, ám ilyenkor az $n_0$ kezdőindex lényegesen nagyobbá válna a 4. táblázatban kiszámolt utolsó ($n = 20$) esethez képest.*

A [7] cikk második részében az $\mathcal{U}_n$ halmazok alakját vizsgáljuk a képzetes tengely közelében. (Itt az $\mathcal{S}_n$ halmazok helyett esztétikai okokból tekintjük az $\mathcal{U}_n$ halmazokat.) A *Mathematica* tetszőleges pontosságú és adaptív aritmetikáját kihasználva felfedeztük, hogy $n$ növekedésével $\mathcal{U}_n$ határgörbéjének „függőleges" része *nagyon* közel kerül a képzetes tengelyhez: „aszimptotikusan reguláris" oszcillációkat mutat „mikroszkopikus" amplitúdóval. Ezen tulajdonságok jellemzéséhez tekintsük az $\mathcal{U}_n$ halmazoknak a képzetes tengely felső félegyenesével vett

$$\mathcal{V}_n^+ := \left\{ \mathrm{Im}(z) : z \in \mathbb{C},\ \mathrm{Re}(z) = 0,\ \mathrm{Im}(z) \ge 0,\ \left| \sum_{k=0}^{n} \frac{z^k}{k!} \right| \le 1 \right\} \qquad (n \in \mathbb{N}^+) \tag{63}$$

metszeteit (a $\mathcal{V}$ betű itt a függőleges, *vertical* szóra utal). Az első 100 ilyen halmazt a 12. ábra mutatja.




12. ábra: rögzített $1 \le n \le 100$ esetén egy (63)-beli $\mathcal{V}_n^+$ halmaz tipikusan több intervallum uniója, melyeket a jobb láthatóság érdekében az aktuális $n$ értéknek megfelelő oszlopban egy-egy vékony barna téglalappal reprezentálunk. Természetesen minden barna intervallum végpontja egzakt algebrai számként áll rendelkezésre. A (töröttvonallal összekötött) fekete pontok a $\max(\mathcal{V}_n^+)$ sorozat elemei, melyek az $1 \le n \le 100$ tartományban 5-ös vagy 6-os nemcsökkenő blokkokba rendeződnek (a „visszaeső" fekete pontok a 8. ábrán annak felelnek meg, amikor az ottani „buborékok" a fő alakzatról „leválnak"). A felső piros görbe az $n \mapsto \frac{n}{e} + \frac{\ln n}{2e} + 1.2604$ függvény grafikonja, mely görbe az $1 \le n \le 100$ tartományon egy felső becslése az extremális fekete pontoknak.

A 12. ábra megfigyeléseit elméleti szempontból kielégítő módon megmagyarázzuk – lényegében a Hurwitz-tétel Rouché-tétel segítségével történő bizonyításának megismétlésével. Az előbb említett „aszimptotikus regularitás" annak felel meg, hogy az ábrán (az alsó- és felső pozíciótól eltekintve, azaz) a „teljes",

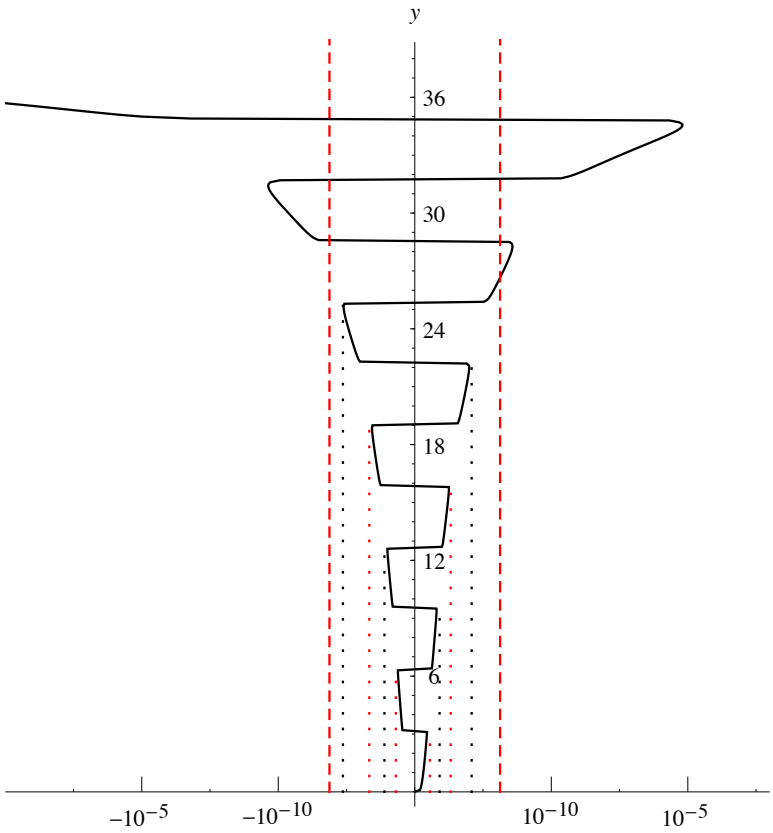
„nem csonka” barna intervallumok hossza $n \to +\infty$ esetén $\pi$-hez tart; pontosabban a barna intervallumok végpontjai $n$ paritásától függően a $\{\pi\ell : \ell \in \mathbb{N}\}$, illetve a $\{0\} \cup \{\pi/2 + \pi\ell : \ell \in \mathbb{N}\}$ rácsokhoz konvergálnak. Azt is igazoljuk, hogy a $\mathcal{V}_n^+$ halmaz 0-*t tartalmazó összefüggő komponense* $n \equiv 0 \mod 4$ vagy $n \equiv 3 \mod 4$ esetén egy pozitív hosszúságú intervallum, míg $n \equiv 1 \mod 4$ vagy $n \equiv 2 \mod 4$ esetén az egypontú $\{0\}$ halmaz – ez jelenti a 12. ábra legalsó sorában a „fehér-fehér-barna-barna” 4-es periódust. A $\mathcal{V}_n^+$ halmaz 0-t tartalmazó *legnagyobb* összefüggő komponense egyébként $n = 8$-hoz tartozik, amikor is $\mathcal{V}_8^+$ egyetlen intervallumból áll: $\mathcal{V}_8^+ = [0, y_{8,1}]$, ahol $y_{8,1} \approx 3.39514$.

Az $\mathcal{U}_n$ halmazok határgörbéinek a képzetes tengely közelében futó részei tehát aszimptotikusan reguláris „frekvenciával” oszcillálnak; de mi a helyzet ezen oszcillációk amplitúdójával, vagyis vízszintes irányú kitéréseivel? Ezt figyelhetjük meg például a 13. ábrán.

A [7] cikk harmadik, befejező részében azt vizsgáljuk, hogy mekkora az $\mathcal{S}_n \cap \{\mathrm{Re} \le 0\}$ halmaz által tartalmazott maximális sugarú bal félkörlap, vagyis mi mondható $\varrho_n^* > 0$ nagyságáról $D_{\varrho_n^*} \cap \{\mathrm{Re} \le 0\} \subset \mathcal{S}_n$ esetén. Az első néhány pontos értéket az 5. táblázat tartalmazza. A cikkben végül numerikus kísérletekkel sugárirányú metszeteket készítve elemezzük, hogy az $\mathcal{S}_n \cap \{\mathrm{Re} \le 0\}$ halmaz alakja mennyire tér el a (61)-beli aszimptotikus $D_{1/e} \cap \{\mathrm{Re} \le 0\}$ félkörlaptól.

| $n$ | $\varrho_n^*$ | A $\varrho_n^*$ algebrai szám foka |
|---|---|---|
| 3 | $\sqrt{3}/3 \approx 0.577$ | 2 |
| 7 | $\approx 0.252$ | 6 |
| 11 | $\approx 0.154$ | 10 |
| 15 | $\approx 0.111$ | 14 |
| 19 | $\approx 0.086$ | 18 |
| 4 | $\approx 0.653$ | 24 |
| 8 | $\approx 0.424$ | 6 |
| 12 | $\approx 0.281$ | 10 |
| 16 | $\approx 0.207$ | 14 |
| 20 | $\approx 0.164$ | 18 |

5. táblázat: a $D_{\varrho_n^*} \cap \{\mathrm{Re} \le 0\} \subset \mathcal{S}_n$ tartalmazásokban fellépő maximális $\varrho_n^*$ konstansok (lefelé kerekített) értéke $3 \le n \le 20$ esetén. A fentiekből tudjuk, hogy ilyen $\varrho_n^* > 0$ számok csak akkor léteznek, ha $n \equiv 0 \mod 4$ vagy $n \equiv 3 \mod 4$. A $\varrho_4^*$ konstans fokszám szempontjából kivételesen viselkedik – vajon mi lehet ennek az oka?




13. ábra: az $\mathcal{U}_{100}$ halmaz határgörbéjének azon része a felső komplex félsíkban, amely az origóból indul és a képzetes tengelyhez nagyon közel fut. Az ábrán könnyen beazonosíthatjuk a 12. ábra jobb szélső oszlopában a $\mathcal{V}_{100}^{+}$ halmazt alkotó 6 barna intervallumot, melyek végpontjai $\pi$ többszöröseinek közelében találhatók. A 12. ábra aszimptotikus formulája (az ottani piros görbe) szerint az $\mathcal{U}_{100}$ halmaz határgörbéje $100/e \approx 36.8$ alatt az oszcilláció befejezésével a bal félsíkba lép át, ahol megkezdi leírni a (8. ábrán látottakhoz hasonló aszimptotikus) félkört. Az $\mathcal{U}_{100}$ halmaz határgörbéjét a *Mathematica*-ban paraméteres algebrai görbeként állítjuk elő: az ebben szereplő polinom foka 200, és együtthatóit összesen körülbelül 76000 számjeggyel lehet leírni. Az oszcillációk (vízszintes irányú) amplitúdója olyan kicsi abszolút értékű, hogy láthatóvá tételükhöz az ábrán egy speciális (itt pontosan nem definiált) „reciprok logaritmikus" skálát használunk. Viszonyításképpen a függőleges piros *szaggatott* vonalakat a gépi pontosság szokásos $\pm 10^{-16}$-os határánál helyeztük el. A függőleges (piros vagy fekete) *pontozott* vonalak az amplitúdó lokális szélsőértékeit adják meg: ezek helyzete (balról jobbra haladva) körülbelül $-10^{-19}, -10^{-30}, -10^{-45}, -10^{-72}, +10^{-90}, +10^{-55}, +10^{-38}, +10^{-24}$. A nagyságrendek változását az is érzékelteti például, hogy e határgörbe két további pontja az $(y, x) \approx (1, 10^{-160})$ és az $(y, x) \approx (0.1, 10^{-262})$ pont. Érdekes feladat lenne az ábrán függőlegesen kirajzolódó „burkoló tölcsér" (aszimptotikus) alakjának vizsgálata.

# 4. Az oktatáshoz kapcsolódó publikációk (2007–2017)

• L. Lóczi, N. Sándor, *Informatics to Students of Cognitive Science*, 100 oldal, angol nyelvű interaktív tananyag a *Mathematica* programnyelv oktatásához, a BME TÁMOP 4.1.2.A/1-11/0064 támogatásával, 2013. aug.
A támogatott projektek listája elérhető: `http://tankonyvtar.ttk.bme.hu/uj_tamop/8.html` (*Glimpses of mathematics for students in cognitive science*).

• Lóczi L., *A Fourier-sorfejtés és a Laplace-transzformáció*, 52 oldal, a BME TÁMOP 4.1.2-08/2/A/KMR-2009-0027 pályázatának keretében, 2011. ápr. A jegyzet a *BME Gépészkari Matematika MSc* c. 330 oldalas szabadon letölthető tankönyv egyik fejezetét alkotja.
Elérhető: `http://tankonyvtar.ttk.bme.hu/pdf/23.pdf`

## 4.1. Főbb fordítások és ismeretterjesztő publikációk (2007–2017)

• Könyvismertető *Stephen Wolfram: An Elementary Introduction to the Wolfram Language* c. könyvéről, Érintő Elektronikus Matematikai Lapok (a Bolyai János Matematikai Társulat lapja), **1**, 2016 szeptember.
Elérhető: `http://www.ematlap.hu/index.php/gazda-g-sag/`
`340-a-wolfram-programozasi-nyelv-rovid-tortenete-3`

• A *Középiskolai Matematikai és Fizikai Lapok* (`http://komal.hu`) informatika feladatrovatának angol fordítója 2001 és 2016 között

• *Thomas' Calculus: Second-order Differential Equations* c. könyvfejezet magyarra fordítása, 45 oldal, a BME TÁMOP 4.1.2-08/2/A/KMR-2009-0027 pályázatának keretében, 2011. febr.
Elérhető: `http://tankonyvtar.ttk.bme.hu/pdf/21.pdf`

• *Find the error*, a `http://uni.dougshaw.com/findtheerror/` honlapon szereplő 11 matematikai dialógus magyarra fordítása, 8 oldal, a BME TÁMOP 4.1.2-08/2/A/KMR-2009-0027 pályázatának keretében, 2011. febr.
Elérhető: `http://tankonyvtar.ttk.bme.hu/pdf/49.pdf`

• *D. E. Knuth: The Art of Computer Programming, Vol. 4, Fascile 3. Generating All Combinations and Partitions, pp. 71–96, pp. 135–155* könyvrészletek magyarra fordítása, 2008. júl.–aug.

• A `www.universalcurriculum.com` honlap középiskolásoknak szóló interaktív matematika tananyagának magyarra fordítása a TranzPress Fordítóiroda fordítócsapatának tagjaként, a Nemzeti Fejlesztési Terv egyik alprogramjának keretében, 2007. jún.–aug.
A magyar változat a `http://realika.educatio.hu/` címen érhető el.

# Hivatkozások


[1] `www.wolfram.com`

[2] L. Lóczi, *Exact optimal values of step-size coefficients for boundedness of linear multistep methods*, Numerical Algorithms, **77**, No. 4 (2018), 1093–1116
First online: June 7, 2017
Elérhető: `https://arxiv.org/pdf/1609.07858`
DOI: `http://dx.doi.org/10.1007/s11075-017-0354-5`

[3] I. Fekete, D. I. Ketcheson, L. Lóczi, *Positivity for convective semi-discretizations*, J. Sci. Comp., **74**, No. 1 (2018), 244–266
First online: Apr 19, 2017
Elérhető: https://arxiv.org/abs/1610.00228
DOI: http://dx.doi.org/10.1007/s10915-017-0432-9

[4] D. I. Ketcheson, L. Lóczi, A. Jangabylova, A. Kusmanov, *Dense output for strong stability preserving Runge–Kutta methods*, J. Sci. Comp., **71**, No. 3 (2017), 944–958
Elérhető: http://arxiv.org/abs/1605.02429
DOI: http://dx.doi.org/10.1007/s10915-016-0331-5

[5] Y. Hadjimichael, D. I. Ketcheson, L. Lóczi, A. Németh, *Strong stability preserving explicit linear multistep methods with variable step size*, SIAM J. Num. Anal., **54**, No. 5 (2016), 2799–2832
Elérhető: https://arxiv.org/abs/1504.04107
DOI: http://dx.doi.org/10.1137/15M101717X

[6] L. Lóczi, *Discretizing the transcritical and pitchfork bifurcations—conjugacy results*, J. Diff. Eq. Appl., **21**, No. 3 (2015) 155–196
Elérhető: http://arxiv.org/abs/1411.6252
DOI: http://dx.doi.org/10.1080/10236198.2014.992786

[7] D. I. Ketcheson, T. A. Kocsis, L. Lóczi, *On the absolute stability regions corresponding to Taylor polynomials of the exponential function*, IMA J. Num. Anal., **35**, No. 3 (2015), 1426–1455
Elérhető: http://arxiv.org/abs/1312.0216
DOI: http://dx.doi.org/10.1093/imanum/dru039

[8] D. I. Ketcheson, L. Lóczi, M. Parsani, *Internal error propagation in explicit Runge–Kutta methods*, SIAM J. Num. Anal., **52**, No. 5 (2014), 2227–2249
DOI: http://dx.doi.org/10.1137/130936245

[9] L. Lóczi, D. I. Ketcheson, *Rational functions with maximal radius of absolute monotonicity*, LMS J. Comp. Math., **17**, No. 1 (2014), 159–205
Elérhető: http://arxiv.org/abs/1303.6651
DOI: http://dx.doi.org/10.1112/S1461157013000326

[10] J. Páez Chávez, L. Lóczi, *Various closeness results in discretized bifurcations*, Differ. Equ. Dyn. Syst., **20**, No. 3 (2012) 235–284
DOI: http://dx.doi.org/10.1007/s12591-012-0135-5

[11] L. Lóczi, J. Páez Chávez, *Preservation of bifurcations under Runge–Kutta methods*, Int. J. Qual. Th. Diff. Eq. Appl., **3**, No. 1–2 (2009) 81–98
Előnézet elérhető: http://math.uni-pannon.hu/~hartung/ijqtdea/vol3/loczi.pdf

[12] B. M. Garay, L. Lóczi, *Discretizing the fold bifurcation—a conjugacy result*, Per. Math. Hung., **56**, No. 1 (2008) 37–53
DOI: http://dx.doi.org/10.1007/s10998-008-5037-1

[13] W. Hundsdorfer, J. Verwer, *Numerical Solution of Time-dependent Advection-diffusion-reaction Equations*, Springer Series in Computational Mathematics, **33**, Springer-Verlag, Berlin (2003)

[14] H. L. Smith, *Monotone Dynamical Systems*, Amer. Math. Soc., Providence, RI (1995)

[15] I. Higueras, *Strong stability for Runge–Kutta schemes on a class of nonlinear problems*, J. Sci. Comput., **57**, No. 3 (2013), 518–535

[16] E. Hairer, Ch. Lubich, G. Wanner: *Geometric Numerical Integration*, Springer-Verlag, Berlin (2006)

[17] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*, John Wiley & Sons Ltd., Chichester (2008)

[18] G. Dahlquist, R. Jeltsch, *Reducibility and contractivity of Runge–Kutta methods revisited*, BIT Num. Math., **46** (2006), 567–587

[19] W. Hundsdorfer, M. N. Spijker, *Boundedness and strong stability of Runge–Kutta methods*, Math. Comp., **80**, No. 274 (2011), 863–886

[20] E. Hairer, S. P. Nørsett, G. Wanner, *Solving Ordinary Differential Equations, I. Nonstiff Problems*, Springer, Berlin (2008)

[21] E. Hairer, G. Wanner, *Solving Ordinary Differential Equations II., Stiff and Differential-Algebraic Problems*, Springer, Berlin (2002)

[22] M. N. Spijker, *The existence of stepsize-coefficients for boundedness of linear multistep methods*, Appl. Numer. Math., **63** (2013), 45–57

[23] M. N. Spijker, *Stability and boundedness in the numerical solution of initial value problems*, Math. Comp., **86** No. 308 (2017), 2777–2798

[24] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Basel (1999)

[25] S. Gottlieb, D. Ketcheson, C.-W. Shu, *Strong Stability Preserving Runge–Kutta and Multistep Time Discretizations*, World Scientific Publishing Co., Hackensack, NJ (2011)

[26] J. A. van de Griend, J. F. B. M. Kraaijevanger, *Absolute monotonicity of rational functions occurring in the numerical solution of initial value problems*, Numer. Math., **49**, No. 4 (1986), 413–424

[27] J. F. B. M. Kraaijevanger, *Contractivity of Runge–Kutta methods*, BIT Numer. Math., **31**, No. 3 (1991), 482–528

[28] Z. Horváth, *Positivity of Runge–Kutta and diagonally split Runge–Kutta methods*, Appl. Numer. Math., **28** (1998), 309–326

[29] P. E. Kloeden, J. Schropp, *Runge–Kutta methods for monotone differential and delay equations*, BIT Num. Math., **43** (2003), 571–586

[30] B. M. Garay, L. Lóczi, *Monotone delay equations and Runge–Kutta discretization*, Special Issue of Funct. Diff. Equ., **11**, No. 1–2 (2004) 59–67
Elérhető:
`http://functionaldifferentialequations.com/index.php/fde/article/view/232`

[31] L. Bonaventura, A. D. Rocca *Unconditionally strong stability preserving extensions of the TR-BDF2 method*, J. Sci. Comp., **70** (2017), 859–895

[32] M. N. Spijker, *Stepsize conditions for general monotonicity in numerical initial value problems*, SIAM J. Numer. Anal., **45**, No. 3 (2007), 1226–1245

[33] W. Hundsdorfer, A. Mozartova, M. N. Spijker, *Stepsize conditions for boundedness in numerical initial value problems*, SIAM J. Numer. Anal., **47**, No. 5 (2009), 3797–3819

[34] W. Hundsdorfer, S. J. Ruuth, R. J. Spiteri, *Monotonicity-preserving linear multistep methods*, SIAM J. Numer. Anal., **41**, No. 2 (2003), 605–623

[35] S. J. Ruuth, W. Hundsdorfer, *High-order linear multistep methods with general monotonicity and boundedness properties*, J. Comput. Phys., **209**, No. 1 (2005), 226–248

[36] W. Hundsdorfer, S. J. Ruuth, *On monotonicity and boundedness properties of linear multistep methods*, Math. Comp., **75**, No. 254 (2006), 655–672

[37] W. Hundsdorfer, A. Mozartova, M. N. Spijker, *Special boundedness properties in numerical initial value problems*, BIT, **51**, No. 4 (2011), 909–936

[38] W. Hundsdorfer, A. Mozartova, M. N. Spijker, *Stepsize restrictions for boundedness and monotonicity of multistep methods*, J. Sci. Comput., **50**, No. 2 (2012), 265–286

[39] D. I. Ketcheson, L. Lóczi, M. Parsani, *Propagation of internal errors in explicit Runge–Kutta methods and internal stability of SSP and extrapolation methods*, Technical report, 57 oldal (2014), `http://arxiv.org/abs/1309.1317`

[40] M. R. S. Kulenović, G. Ladas, *Dynamics of Second Order Rational Difference Equations with Open Problems and Conjectures*, Chapman & Hall/CRC (2002)

[41] M. M. Khalsaraei, *An improvement on the positivity results for 2-stage explicit Runge–Kutta methods*, J. Comput. Appl. Math., **235**, No. 1 (2010), 137–143

[42] A. Ralston, *Runge–Kutta methods with minimum error bounds*, Math. Comp., **16** (1962), 431–437

[43] M. M. Khalsaraei, *Positivity of an explicit Runge–Kutta method*, Ain Shams Engin. Journal, **6**, No. 4 (2015), 1217–1223

[44] J. Ouaknine, J. Worrell, *Decision problems for linear recurrence sequences*, In: Reachability Problems, 6th International Workshop, RP 2012, Bordeaux, France, September 17–19, 2012, `http://dx.doi.org/10.1007/978-3-642-33512-9_3`

[45] J. Ouaknine, J. Worrell, *Positivity problems for low-order linear recurrence sequences*, In: Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, 2014, `http://dx.doi.org/10.1137/1.9781611973402.27`

[46] J. Ouaknine, J. Worrell, *On the positivity problem for simple linear recurrence sequences*, In: Automata, languages, and programming. Part II, 318–329, Springer, Heidelberg, 2014

[47] J. Ouaknine, J. Worrell, *Ultimate positivity is decidable for simple linear recurrence sequences*, In: Automata, languages, and programming. Part II, 330–341, Springer, Heidelberg, 2014

[48] G. Szegő, *Über eine Eigenschaft der Exponentialreihe*, Sitzungsber. Berl. Math. Ges., **23** (1924), 50–64

[49] J. D. Buckholtz, *A characterization of the exponential series*, Amer. Math. Monthly, **73**, No. 4, part II (1966), 121–123

[50] R. Jeltsch, O. Nevanlinna, *Stability and accuracy of time discretizations for initial value problems*, Numer. Math., **40**, No. 2 (1982), 245–296