Eötvös Loránd Tudományegyetem
Természettudományi Kar

Orobej Máté Borisz

Egy explicit nemstandard véges differencia séma-család merev feladatokra

BSc Alkalmazott Matematikus Szakdolgozat

Témavezető:

Dr. Fekete Imre
Alkalmazott Analízis és Számításmatematikai Tanszék

Budapest
2024

Tartalomjegyzék

Köszönetnyilvánítás

Szeretném kifejezni hálámat témavezetőmnek, Dr. Fekete Imrének, hogy megismertette velem a numerikus analízis világát. Az előadásaitól kezdve a személyes konzultációkig, mindig érezhettem motiváló hozzáállását, ami a sok szakmai segítség mellett nagyban hozzájárult ahhoz, hogy lelkes szívvel tekintsek a feladatra.

Továbbá szeretném megköszönni családomnak a rengeteg támogatást, amit az elmúlt évek során kaptam tőlük. Nem feledkezhetek meg barátnőmről sem, aki mindvégig ott volt mellettem és századjára is kikérhettem véleményét, akár ugyanarról a bekezdésről is. Illetve minden barátomat is köszönet illeti, nélkülük feleekkora élmény sem lett volna a BSc képzés.

I. Bevezetés

Folytonosan változó jelenségekkel világunk minden területén találkozhatunk, ezen folyamatok modellezésére született a differenciálegyenletek egyre bővülő eszköztára. Ilyen, matematikailag megfogható modellek a fizika mellett a biológiától kezdve a közgazdaságtanig számtalan tudományág megértésében kulcsfontosságú szerepet töltenek be.

A differenciálegyenletek (röviden DE) és megoldásaik vizsgálatában a legnagyobb akadály, hogy pontos megoldást zárt alakban a felbukkanó problémák csak nagyon speciális osztályaira tudunk adni. Emiatt született az igény, hogy a megoldást egy numerikus módszer segítségével közelítsük. Az elmélet és a technológia jelentős fejlődésének köszönhetően egyre "nehezebb" problémákra tudunk egyre "jobb" közelítő megoldást adni. Ugyanakkor továbbra is fennállnak nehézségek, melyek hatékony orvoslása napjainkban is kutatások tárgya.

Dolgozatom célja, hogy bemutassa a standard módszerek felépítését és ezek bizonyos nehézségeit merev feladatok esetén, majd betekintést nyújtson egy új, alternatív módszervilágba, az úgynevezett *nemstandard véges differencia* (angolul NonStandard Finite Difference - NSFD) metodikába, mely segítségével képesek leszünk áthidalni a megismert akadályokat.

A 2. fejezet [16] és [12, 13] könyvek megfelelő részeit követve a közönséges differenciálegyenletek (KDE) egylépéses módszereinek standard elméletéről szól. Ebben a részben ismertetem az alapfogalmakat és tételeket, majd az explicit Euler továbbfejlesztésével kapott Runge–Kutta (RK) módszer-családot. Példákon keresztül megismerkedhetünk merev feladatok nehézségeivel, majd bevezetésre kerülnek fontos stabilitási fogalmak is.

A 3. fejezet a téma jeles úttörőjének, Mickens monográfiájának KDE részét dolgozza fel [19]. Numerikus instabilitások különböző forrásait fogom bemutatni. Majd konkrét feladatokon keresztül példát láthatunk, hogyan orvosolhatóak a tapasztalt problémák megfelelő NSFD séma alkalmazásával.

Végül a 4. fejezetben az ismertetett metodika elemeit felhasználva a [15] szakcikk eredményét vezetem le és hasonlítom össze más megoldásokkal. Így végeredményül egy olyan explicit NSFD séma-családot kapunk, mely rendelkezik a megismert stabilitási tulajdonságokkal, így képes hatékonyan megoldani merev feladatokat.

II. Standard véges differencia módszerek

Mint a bevezetőben is említettem, a közönséges differenciálegyenletek klasszikus, avagy standard numerikus módszereivel fogunk foglalkozni. Itt két fő kategóriát különböztethetünk meg: azokat, melyek minden lépés során csupán egy kezdőértéket használnak fel (ezeket nevezzük *egylépéses* módszereknek) és azokat, melyek több, a megelőző lépésekben kiszámolt adatpontot is felhasználnak (*többlépéses* módszerek). Említésre méltó Bashforth és Adams 1883-as cikke [1], melyben többek között a sok mai alkalmazás szempontjából is fontos, többlépéses Adams-Bashforth módszert írták le. Továbbá Runge 1895-ös publikációja [26], melyre a modern egylépéses módszerek kiindulópontjaként tekintünk [4]. Munkámban az utóbbi, egylépéses módszerek elméletével fogok foglalkozni. Ugyan a definíciók általános formában kerülnek kimondásra, a könnyebb átláthatóság kedvéért a levezetések a skalár esetet fogják tárgyalni.

Mielőtt numerikus módszereket szerkesztenénk, tisztázzuk az alapfeladatot:

2.1. Definíció *Legyen $G \subset \mathbb{R} \times \mathbb{R}^d$ tartomány (azaz összefüggő és nyílt halmaz), $(t_0, \boldsymbol{u}_0) \in G$ egy adott pont, $\boldsymbol{f} : G \to \mathbb{R}^d$ egy folytonos leképezés. A*

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} &= \boldsymbol{f}(t, \boldsymbol{u}(t)) \\ \boldsymbol{u}(t_0) &= \boldsymbol{u}_0 \end{aligned} \tag{2.1}$$

feladatot Cauchy-feladatnak*, avagy* kezdetiérték-feladatnak *nevezzük és* CF*-el jelöljük.*

2.2. Definíció *Az olyan $\boldsymbol{u} : I \to \mathbb{R}^d$ (I nyílt intervallum) folytonasan differenciálható függvényt, amelyre*

- $\{(t, \boldsymbol{u}(t)) : t \in I\} \subset G$,
- $u'(t) = f(t, \boldsymbol{u}(t))$ *minden* $t \in I$,
- $t_0 \in I$ *és* $\boldsymbol{u}(t_0) = \boldsymbol{u}_0$

teljesül, a (2.1) CF megoldásának *nevezzük.*

Valós problémát modellező feladatnál alapvető elvárás, hogy létezzen egyértelmű megoldás. Ezt a Picard–Lindelöf tétel garantálja [13], amennyiben $\boldsymbol{f}$ teljesíti a következő, *második változóbeli Lipschitz* tulajdonságot:

$$\exists L > 0 : \|\boldsymbol{f}(t, \boldsymbol{u}_1) - \boldsymbol{f}(t, \boldsymbol{u}_2)\| \leq L\|\boldsymbol{u}_1 - \boldsymbol{u}_2\| \quad \forall (t, \boldsymbol{u}_1), (t, \boldsymbol{u}_2) \in G. \tag{2.2}$$

Emiatt a második változóbeli Lipschitz tualjdonságot mindig fel fogjuk tenni. Emellett, hogy a jelölések átláthatóak maradjanak, az I intervallumot $[0, T]$-nek választjuk és a jelöléseket a skalár esethez igazítjuk.

II.1. Egylépéses módszerek alapjai

Először is bevezetünk egy rácshálót, amivel a dolgozat további részében dolgozni fogunk.

2.3. Definíció *A* $[0, T]$ *időintervallum alábbi*

$$\omega_h := \{t_n = nh : n = 0, 1, 2, \ldots, N;\ h = T/N\}$$

egyenletes felosztását ekvidisztáns rácshálónak *nevezzük.*

Ezen t_n pontokban állítjuk elő $u(t_n)$ – a pontos megoldás – közelítését, melyet y_n-el jelölünk. Idézzük fel az első, talán legegyszerűbb numerikus módszert, melyet Euler *Institutiones Calculi Integralis* című 1768-as könyvében publikált [10]. Tekintsük a (2.1) bal oldalán szereplő derivált egyik lehetséges

$$\frac{y_{n+1} - y_n}{h} = f(t_n, y_n), \quad n = 0, 1, \ldots, N - 1, \tag{2.3}$$

diszkretizációját, a haladó véges differencia sémát. Az ebből átrendezéssel nyert

$$y_{n+1} = y_n + hf(t_n, y_n), \quad n = 0, 1, \ldots, N - 1$$

véges differencia módszert $y_0 = u_0$ természetes megválasztása mellett *explicit Euler* (röviden EE) módszernek nevezzük. Az y_n numerikus megoldás $[0, T]$-n való konvergenciája (akár változó lépésköz mellett, egy nem ekvidisztáns rácshálón is) bizonyítható [16], és mint kiderül, a hiba Ch-val felülről becsülhető, ahol C egy h-tól független konstans. Ez [13] II. fejezetének példájával élve azt jelenti, hogyha egy adott feladatra N lépéssel a módszer hibája egy, míg a kívánt pontosság hét tizedesjegyű, akkor $10^6 N$ lépést kell tennünk, hogy ezt garantálni tudjuk.

Ez igencsak motiválja a kérdést, hogy tudunk-e hatékonyabb módszereket konstruálni. Ennek megválaszolásához szükségünk lesz az alábbi fogalmakra, melyeket a következő Φ általános egylépéses numerikus módszerre fogalmazunk meg:

$$y_{n+1} = y_n + h\Phi(h, t_n, y_n, y_{n+1}), \quad n = 0, 1, \ldots, N-1. \tag{2.4}$$

2.4. Definíció *Azokat a módszereket, amelyekre* $\Phi = \Phi(h, t_n, y_n,)$ *(tehát a* Φ *függvény nem függ* y_{n+1}*-től)* explicit *módszernek nevezzük. Amennyiben* Φ y_{n+1}*-től is függ, akkor a módszert* implicitnek *nevezzük.*

2.5. Definíció *Az* $e_h(t_\star) = y_n - u(t_\star)$, *(ahol* $nh = t_\star$*) függvényt a* Φ *numerikus módszer* $t_\star$ *pontbeli* globális diszkretizációs hibafüggvényének *nevezzük. Azt mondjuk, hogy a* Φ *numerikus módszer* konvergens a $t_\star$ pontban*, ha*

$$\lim_{h \to 0} e_h(t_\star) = 0. \tag{2.5}$$

Ha a Φ *numerikus módszer konvergens* $[0, T]$ *minden pontjában, akkor* konvergens módszernek *nevezzük. A* (2.5) *konvergenciájának rendjét pedig a* Φ*-módszer* konvergenciarendjének *nevezzük.*

Mivel a globális hibát sokszor nem könnyű kiszámítani, érdemes a lokális hibát vizsgálni. Ehhez vezessük be az alábbi

$$g_n(h) = -u(t_{n+1}) + u(t_n) + h\Phi(h, t_n, u(t_n), u(t_{n+1})) \tag{2.6}$$

függvényt, mely a pontos megoldásból indított *numerikus lépés* és *pontos lépés* eltérését mutatja.

2.6. Definíció *A* (2.6) $g_n(h)$ *függvényt a* (2.4) *alakú* Φ *numerikus módszer* $t_n \in \omega_h$ *pontbeli* képlethibájának *(más szóval* lokális approximációs hibafüggvényének*) nevezzük. Azt mondjuk, hogy a* Φ *numerikus módszer* p-edrendben konzisztens*, ha*

$$g_n(h) = \mathcal{O}(h^{p+1})$$

valamely $p > 0$ *állandóval minden* $t_n \in \omega_h$ *rácspontban.*

Végül a két hibafogalmat összekötő elvet, a numerikus analízis alaptételét mondjuk ki, miszerint egy numerikus módszer *stabilitása*[1] mellett, p-edrendű konzisztenciából p-edrendű konvergencia következik.

[1] A fogalom a numerikus módszer bemenő adatoktól való folytonos függését, pontosabban a hiba korlátos változását fejezi ki.

2.7. Tétel **[16, 3.24]** *Tegyük fel, hogy a* (2.4) *képlettel definiált* Φ *numerikus módszer*

- *p-edrendben konzisztens;*
- *a* Φ *függvény folytonos, a harmadik és negyedik változójában egyaránt kielégíti a Lipschitz feltételt*[2].

Ekkor a módszer p-edrendben konvergens a $[0, T]$ *intervallumon.*

Ezen fogalmak ismeretében, az explicit Euler egy lépését összehasonlítva a pontos megoldás t_n-körüli Taylor-sorával látható, hogy a módszer valóban elsőrendben konzisztens és f Lipschitz tulajdonságának köszönhetően konvergens is. Ez a [13, 16]-ban részletesen is kifejtett számolás egy standard technikának mondható, így dolgozatomban nem fogom tárgyalni.

A továbbiakban magasabbrendű módszerek szerkesztésével fogunk foglalkozni. Ez nem különösebben nehéz abban az esetben, amikor f csupán t függvénye, ilyenkor a (2.1) CF az alábbi

$$u'(t) = f(t), \quad u(0) = u_0$$

interpolációs feladatra egyszerűsödik, melynek megoldása

$$u(t) = u_0 + \int_0^t f(s)\mathrm{d}s.$$

Ha az integrált a középponti

$$u(t_{n+1}) \approx u(t_n) + hf(t_n + h/2), \quad n = 0, 1, \ldots, N-1$$

kvadratúrával közelítjük, akkor az alábbi

$$\begin{aligned} y_{n+1} &= y_n + hf(t_n + h/2), \quad n = 0, 1, \ldots, N-1, \\ y_0 &= u_0 \end{aligned}$$

numerikus módszert nyerjük. Mivel az alkalmazott kvadratúra rendje kettő, a kapott közelítés hibája $\mathcal{O}(h^2)$, ami az előző példa feltevéseivel élve, a kívánt 7 tizedesjegyű pontosság eléréshez csupán $10^3 N$ lépést követel meg. Ez ezerszer kevesebb lépés, mint az explicit

[2] Azaz léteznek $L_3, L_4 \geq 0$ állandók, amelyek mellett tetszőleges s_1, s_2, p_1 és p_2 számok esetén $|\Phi(h, t_n, p_1, s_1) - \Phi(h, t_n, p_2, s_2)| \leq L_3|p_1 - p_2| + L_4|s_1 - s_2|$ teljesül.

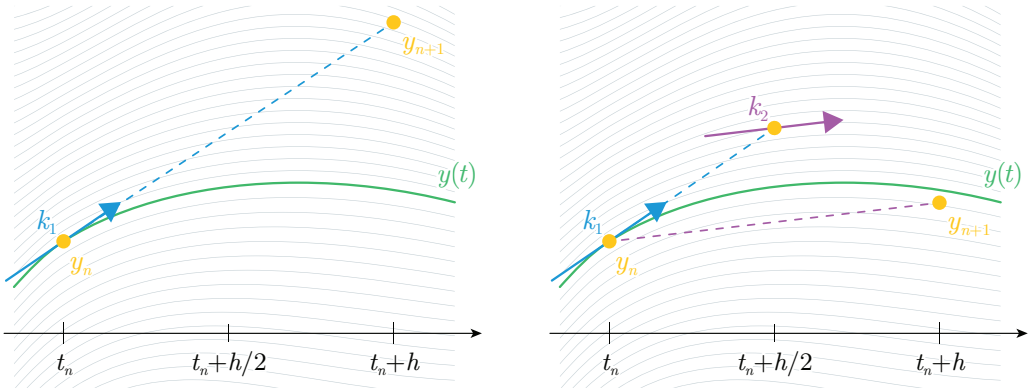
Euler alkalmazásával. Most próbáljuk meg kiterjeszteni a fenti ötletet az általános (2.1) feladat megoldására. Ekkor az alábbi

$$y_{n+1} = y_n + hf(t_n + h/2, y(t_n + h/2))$$

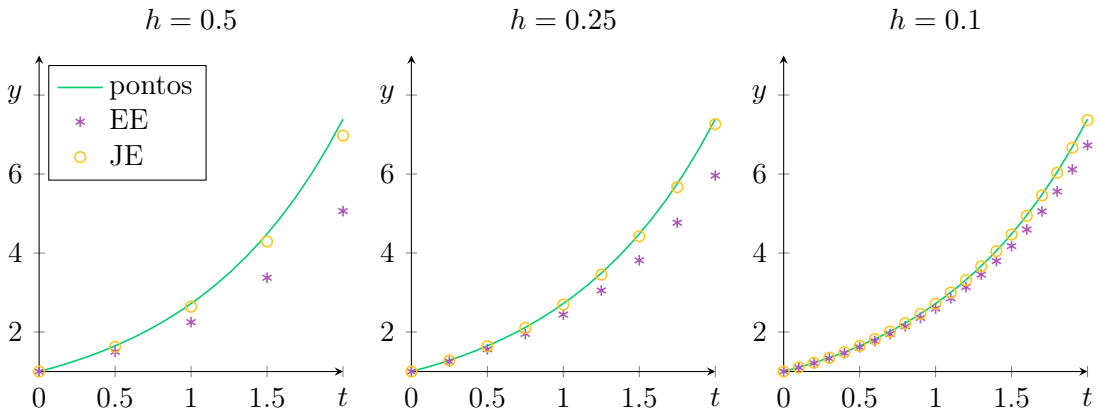
formulát kapjuk az n-edik lépésre. Ebben a képletben az $y(t_n+h/2)$ ismeretlen kiszámítása egy nemlináris egyenlet megoldását követelné, viszont egy $h/2$ hosszú EE lépéssel explicit módon közelíthetjük. Ez pedig a következő

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + h/2, y_n + h/2 \cdot k_1) \\ y_{n+1} &= y_n + hk_2 \end{aligned} \tag{2.7}$$

explicit, másodrendű módszerhez vezet, melyet ezért *javított Euler* módszernek nevezünk és a JE szimbólummal jelöljük. Az eddigi két módszer összehasonlítását mutatja be a 2.1. és 2.2. ábra.



2.1. ábra. Az explicit Euler (bal) és a javított Euler (jobb) egy lépése.



2.2. ábra. Az eddigi módszerek az $u'(t) = u(t),\ u(0) = 1$ feladaton finomodó lépésköz mellett.

II.2. Explicit Runge–Kutta módszerek

Megfigyelhettük, hogy míg az EE egy lépése során csupán egyszer kerül kiértékelésre a jobb oldali f függvény, a JE egy lépésében kétszer is, ezt k_1 és k_2 formájában láthatjuk a módszert definiáló (2.7) képletben. Ezeket *lépcsőknek* hívjuk, így az EE *eglépcsős*, míg a JE *kétlépcsős* eglépéses módszer. Ebben a szakaszban megvizsgáljuk, hogy több lépcső felvételével tudunk-e magasabbrendű módszereket szerkeszteni. Így hát keressük a módszerünket (2.7) általánosításaként a következő formában:

2.8. Definíció *Legyen* $s \in \mathbb{N}$ *adott és* $a_{i,j},\ b_i,\ c_i \in \mathbb{R}$ $(i,j = 1,2,\dots,s)$ *rögzített együtthatók. Ekkor*

$$\begin{aligned} k_1 &= f(t_n + c_1 h, y_n) \\ k_2 &= f(t_n + c_2 h, y_n + h a_{21} k_1) \\ k_3 &= f(t_n + c_3 h, y_n + h(a_{31} k_1 + a_{32} k_2)) \\ &\vdots \\ k_s &= f(t_n + c_s h, y_n + h(a_{s1} k_1 + \dots + a_{s,s-1} k_{s-1})) \\ y_{n+1} &= y_n + h(b_1 k_1 + \dots + b_s k_s) \end{aligned} \tag{2.8}$$

egy s*-lépcsős explicit Runge–Kutta típusú módszer (röviden* ERK*).*

Megfigyelhető, hogy $s = 1$ esetén az EE-t kapjuk, míg $s = 2$ esetén a paraméterek megfelelő választásával a JE-t, így ez a definíció valóban általánosítja az eddigi módszereinket[3]. Bevezetésként tekintsük az s=2 esetet és keressünk egy másik másodrendű módszert. Ekkor egy általános lépés az

$$y_{n+1} = y_n + b_1 h f(t_n, y_n) + b_2 h f(t_n + c_2 h, y_n + h a_{21} f(t_n, y_n))$$

alakot ölti. A pontos megoldást behelyettesítve, majd Taylor-sorba fejtve és h együtthatóit csoportosítva az alábbi egyenlőséget kapjuk:

$$\begin{aligned} u(t_n + h) &= u(t_n) + b_1 h\big(f + c_2 h \partial_1 f + a_{21} h \partial_2 f \partial_2 f\big) + \mathcal{O}(h^3) \\ &= u(t_n) + (b_1 + b_2) h f + h^2\big(c_2 b_2 \partial_1 f + b_2 a_{21} f \partial_2 f\big) + \mathcal{O}(h^3). \end{aligned}$$

Így a másodrendű konzisztenciához az alábbi szükséges és elégséges feltételek adhatók:

$$b_1 + b_2 = 1, \qquad c_2 b_2 = 1/2, \qquad a_{21} b_2 = 1/2. \tag{2.9}$$

[3]Érdemes azt is megfigyelni, hogy amennyiben f teljesíti a (2.2)-es tulajdonságot L Lipschitz-állandóval, akkor az s-lépcsős ERK-t leíró Φ függvény kielégíti a 2.7. Tétel második pontját $L_3 = L$ egy s-edfokú polinomjával.

Az egyenletrendszer általános megoldása a következő $b \neq 0$ paraméterrel írható fel:

$$b_2 = b, \qquad b_1 = 1 - b, \qquad c_2 = a_{21} = \frac{1}{2b}.$$

Felmerülhet a kérdés, hogy a paraméterek ügyesebb megválasztásával vajon kaphatunk-e harmadrendű módszert. A válasz nemleges és könnyen ellenőrizhető az $f(y,t) = y$ lineáris egyenlettel [16, 4.1 fejezet]. A másodrend feltételét, (2.9)-et behelyettesítve megkapjuk, hogy a hibatagban h^3 együtthatója minden mástól függetlenül 1/6.

Nagyobb s értékekre a paraméterek mennyisége miatt nehezen átláthatóvá válhat a definícióban szereplő jelölés. Ezt orvosolandó J. C. Butcher az alábbi

$$\begin{array}{c|ccccc} c_1 & & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ c_s & a_{s,1} & a_{s,2} & \dots & a_{s,s-1} & \\ \hline & b_1 & b_2 & \dots & b_{s-1} & b_s \end{array}$$

könnyen áttekinthető felírási módot vezette be [3], melyet az adott ERK módszer *Butcher-tablójának* nevezzük. Az így kapott szigorú alsó háromszögmátrixot, két vektort és az általuk alkotott Butcher-tablót röviden a következőképp jelöljük:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ a_{21} & 0 & 0 & \dots & 0 \\ a_{31} & a_{32} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{s1} & a_{s2} & \dots & a_{s,s-1} & 0 \end{bmatrix}, \qquad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_s \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_s \end{bmatrix}, \qquad \begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^\top \end{array}.$$

A lenti 2.1. táblázatban az eddigi módszerek mellett Runge és Kutta klasszikus harmadrendű módszereinek Butcher-tablójára láthatunk példát. Ezekben a módszerekben minden c_i éppen az $\mathbf{A}$ együtthatómátrix i-edik sorösszegével egyenlő, azaz $\mathbf{Ae} = \mathbf{c}$. Ez feltétele annak, hogy minden pont, ahol f-et kiértékeljük a pontos megoldás egy elsőrendű közelítése legyen. Ugyan kettő és három lépcsőnél ezen feltétel elhagyásával növekedni fog a paraméterek szabadsági foka (másod-, avagy harmadrend feltételeire nézve), negyed- és magasabbrendű módszerek szerkesztésében már hátrányt jelent [22]. Így ez egy – már Kuttánál is megjelent – alapfeltevés, amiből $c_1 = 0$ következik.

0	0
	1

0		
1/2	1/2	
1	0	1
	1/6	2/3	0	1/6

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

0		
1/2	1/2	
	0	1

2.1. táblázat. EE, JE, harmadrendű Runge és harmadrendű Kutta módszerek Butcher-tablója.

II.2.1. Rendfeltételek

"The reader is now asked to take a deep breath, take five sheets of reversed computer paper, remember the basic rules of differential calculus, and begin the following computations: ..."
(E. Hairer, G. Wanner, S. P. Nørsett [13, II.2])

Mint az az ERK módszereket definiáló képletekből is sejthető, magasabb rend ellenőrzése egyre hosszabb Taylor-sorba fejtéssel jár, ami elméletben egyszerű, de a lépcsők számában igencsak gyorsan növekvő mennyiségű számolást igényel, mely a direkt megközelítést ellehetetleníti. Erre egy lehetséges megoldás a (2.9)-es egyenletekhez hasonló *rendfeltételek* szerkesztése, melyek gyorsan kiértékelhető egyenletek és adott s lépcsőszám esetén, amennyiben teljesülnek, garantálják a kívánt rendű konzisztenciát. Habár ilyen rendfeltételek megalkotása korántsem egyszerű feladat, bizonyos címkézett gyökeres-fák segítségével a formulák kidolgozása kezelhetőbbé válik [12, II.2]. Dolgozatom szempontjából a leveztések kevésbé fontosak, így az eredmények tárgyalása lesz fókuszban. Először is vezessük be a következő jelöléseket:

$$\mathbf{c}^k = [c_1^k, c_2^k, \dots, c_s^k]^\top, \quad \mathbf{C} = \operatorname{diag}(c_1, c_2, \dots, c_s) \in \mathbb{R}^{s\times s}.$$

Így az $\mathbf{Ae} = \mathbf{c}$ feltételt kihasználva az alábbi kompakt módon tudjuk kimondani a rendfeltételeket:

Rend: p	Feltételek
1	$\mathbf{b}^\top e = 1$
2	$\mathbf{b}^\top \mathbf{c} = 1/2$
3	$\mathbf{b}^\top \mathbf{c}^2 = 1/3, \quad \mathbf{b}^\top \mathbf{Ac} = 1/6$
4	$\mathbf{b}^\top \mathbf{c}^3 = 1/4, \quad \mathbf{b}^\top \mathbf{CAc} = 1/8, \quad \mathbf{b}^\top \mathbf{Ac}^2 = 1/12, \quad \mathbf{b}^\top \mathbf{A}^2\mathbf{c} = 1/24$
5	$\mathbf{b}^\top \mathbf{c}^4 = 1/5, \quad \mathbf{b}^\top (\mathbf{C}^\top)^2 \mathbf{Ac} = 1/10, \quad \mathbf{b}^\top (\mathbf{AC})^2 = 1/20,$ $\mathbf{b}^\top \mathbf{CAc}^2 = 1/15, \quad \mathbf{b}^\top \mathbf{Ac}^3 = 1/20, \quad \mathbf{b}^\top \mathbf{CA}^2\mathbf{c} = 1/30,$ $\mathbf{b}^\top \mathbf{ACAc} = 1/40, \quad \mathbf{b}^\top \mathbf{A}^2\mathbf{c}^2 = 1/60, \quad \mathbf{b}^\top \mathbf{A}^3\mathbf{c} = 1/120$

2.2. táblázat. Runge–Kutta típusú módszerek rendfeltételei $p = 5$-ig.

Tehát egy (2.8) alakú módszer, mely teljesíti az egytől p-ig terjedő egyenleteket, egy legalább p-edrendben konzisztens explicit Runge–Kutta típusú módszer.

Ilyen numerikus eljárások egyik lehetséges szerkesztési módja kvadratúra formulákon alapszik[4]. Talán az egyik leghíresebb – RK4 nevet viselő – ERK módszer is levezethető ily módon. Ehhez tekintsük a Simpson-formulát a pontos megoldásra alkalmazva a (t_n, t_{n+1}) intervallumon:

$$h/6 \cdot [f(t_n, u(t_n)) + 4f(t_n + h/2, u(t_n + h/2) + f(t_{n+1}, u(t_{n+1}))] \,.$$

A zárójelben lévő első tag explicit kiszámítható, a többi ismeretlen tagot különböző EE lépésekkel közelítve a következő négylépcsős ERK módszerhez jutunk:

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + 1/2 \cdot h, y_n + h/2 \cdot k_1) \\ k_3 &= f(t_n + 1/2 \cdot h, y_n + h/2 \cdot k_2) \\ k_4 &= f(t_n + h, y_n + hk_3) \\ y_{n+1} &= y_n + h/6 \cdot (k_1 + 2k_2 + 2k_3 + k_4) \end{aligned}$$

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

2.3. táblázat. A Simpson formulából levezetett RK4 és Butcher-tablója.

[4]Emellett sok más technika létezik [13, II.1], melyet dolgozatomban nem tárgyalok.

Felmerülhet a logikus kérdés: minden s-re létezik olyan s-lépcsős ERK, melynek rendje legalább s? A válasz erre a kérdésre nemleges. Ennek magyarázata pedig a kielégítendő feltételek és szabad változók kapcsolatában rejlik. Hogy ezt megértsük, először tekintsük a 2.4. táblázatot.

Rend: p	1	2	3	4	5	6	7	8	9	10
Feltételek száma	1	2	4	8	17	37	85	200	486	1205

2.4. táblázat. Runge–Kutta módszerek rendfeltételeinek száma.

Azaz adott rend eléréséhez ismert a kielégítendő feltételek száma. Így amit vizsgálnunk kell, az a paraméterek megválasztásának szabadsági foka. Az $\mathbf{Ae} = \mathbf{c}$ feltevést figyelembe véve a (2.8)-es definíció alapján kapjuk, hogy az s-lépcsős ERK szabad változóinak száma $\frac{s(s+1)}{2}$. Ami $s = 5$-re csupán 15, míg a kielégítendő egyenletek száma 17. Így ötödrend eléréséhez egy legalább hatlépcsős módszert kell konstruálnunk. A következő táblázatban $s = 10$-ig láthatjuk az adott lépcsőszámmal elérhető maximális konzisztenciarendet:

Lépcsőszám: s	1	2	3	4	5	6	7	8	9	10
Elérhető rend: p	1	2	3	4	4	5	6	6	7	7

2.5. táblázat. ERK módszerek maximális elérhető rendje.

Mikor egy KDE feladat megoldására numerikus eljárást szeretnénk alkalmazni, nem egyszerű eldönteni, hogy melyik az optimális. Figyelembe véve, hogy a lépcsők kiszámításában az f függvény kiértékelése olykor időigényes lehet, a 2.5. táblázat alapján a négylépcsős negyedrendű ERK típusú módszerek tűnhetnek "optimálisnak". Ugyanakkor bizonyos feladatoknál lehetséges, hogy egy negyedrendű módszer csak olyan kis lépésköz mellett képes garantálni az elvárt pontosságot, hogy annak számítógépes kivitelezése túlzottan költségessé válik. Így az, hogy melyik módszer "optimális", mindig az adott feladattól és a közelítő megoldásra tett elvárásainktól függ.

II.2.2. A lépésköz megválasztása

Ezen alfejezet tartalma alkalmazások szempontjából elengedhetetlen fontosságú, ennek érzékeltetésére szolgál a bekezdés. Könnyű elképzelni olyan valós folyamatokat, melyek a számunkra fontos időintervallum egy rövid részén gyorsan, a többin pedig lassan, vagy akár egyáltalán nem változnak. Ekkor, hogy a gyors változást visszaadja a numerikus közelítés, "kis" lépésközt kell választanunk, viszont ezen esetben a lassan változó részen sokkal több számolást végzünk, mint az "szükséges lenne". Ilyen feladatoknál előnyös lenne olyan eljárást alkalmazni, melynek lépésköze képes időben dinamikusan változni. Továbbá valós alkalmazásokban elég a numerikus közelítést egy adott hibahatáron belül keresni. Ez motiválja a feladatot, hogy a lokális hibában, mely egy p-edrendű módszer esetén, kellően kis h-ra Ch^{p+1} alakú, ne csak a kitevőt, hanem az együtthatót, azaz a fentiekben C-vel jelölt, úgynevezett *hibakonstanst* is ismerjük. Amennyiben ezt megbízhatóan becsülni tudjuk, lehetőségünk nyílik *adaptív* numerikus eljárásokat konstruálni.

Két nevezetes *hibabecslési* eljárást fogok ismertetni [16, 4.2.3], az első – a Richardson extrapoláció – tetszőleges ERK módszerre alkalmazható. A kulcsötlet, hogy t_n-ből két különböző módon lépünk t_{n+1}-be. Először az ERK egy h hosszú lépését, majd két egymást követő $h/2$ hosszú lépését számítjuk ki, ezeket rendre jelöljük y_n és $\widehat{y}_n$-el. Az így nyert két közelítő megoldás különbségéből p-edrendű konzisztenciát feltételezve a

$$C \approx \frac{|y_n - \widehat{y}_n|}{(1 - 2^{-p})h^{p+1}}$$

becslést adhatjuk a hibakonstansra. Az ötlet hibabecslés mellett egy új, eggyel magasabbrendű módszer előállításához is használható. Ehhez vegyük azt észre, hogy a két közelítő megoldás ügyes

$$y_n^* = \widehat{y}_n + \frac{|y_n - \widehat{y}_n|}{2^p - 1}$$

kombinálásával kiejthető C, a lokális hiba Ch^{p+1} együtthatója. Így y_n^* lokális hibája $\mathcal{O}(h^{p+2})$, azaz az új módszer konzisztenciarendje eggyel magasabb, mint az eredeti. Továbbá, adott ε hibakorlát függvényében az alábbi

$$h_{\text{új}} = h_{\text{régi}} \cdot \left(\frac{\varepsilon(2^p - 1)}{|y_n - \widehat{y}_n|} \right)^{\frac{1}{p+1}}$$

explicit módon ki tudjuk fejezni a szükséges lépéshosszt, mely garantálja a kívánt pontosságot.

A második technika két különböző módszer által adott eredményt fog felhasználni a hiba, illetve az optimális lépésköz megtalálására. Ehhez két ERK lépést fogunk tenni, melyek rendje tipikusan p, illetve $p+1$[5]. Így a magasabbrendű módszer segítségével képesek leszünk becsülni az alacsonyabb rendű módszer megoldásának hibáját. Hogyha olyan módszereket választunk, melyek $\mathbf{A}$ együtthatómátrixa megegyezik, minimalizálni tudjuk a számítási munkát. Ilyen esetben a két ERK módszert célszerű egy

$\mathbf{c}$	$\mathbf{A}$
	$\mathbf{b}^\top$
	$\widehat{\mathbf{b}}^\top$

kibővített Butcher-tablóban reprezentálni, ahol $\mathbf{b}$ a p-edrendű, $\widehat{\mathbf{b}}$ pedig a $(p+1)$-edrendű módszerhez tartozó súlyokat tartalmazó vektor. Ilyenkor a két módszert együtt *beágyazott* Runge–Kutta típusú *módszerpárnak* nevezzük[6]. Szokás úgy is mondani, hogy az alacsonyabb rendű *beágyazott módszere* a magasabb rendűnek. Most a hibát a két közelítő megoldás eltérésével becsüljük, melyet a $\mathbf{b}-\widehat{\mathbf{b}}$ vektorral az alábbi

$$e_n = h\sum_{i=1}^{s}(b_i-\widehat{b}_i)k_i$$

alakban is felírhatunk. Így az alábbi

$$h_{\text{új}} = h_{\text{régi}}\cdot\left(\frac{\varepsilon}{e_n}\right)^{\frac{1}{p+1}}$$

kompakt képlet adható az új lépésközre. Ezen ötlet alapján *adaptív* módszereket tudunk szerkeszteni, melyek képesek a lépésközt dinamikusan változtatni a hibabecslés alapján. Érdemes megjegyezni, hogy változó lépésközű módszerek optimális implementálása a legegyszerűbb esetben sem triviális, így nem csoda, hogy óriási elmélet épül ennek kutatására [13, II.5] [12, IV.2]. Az alfejezetet egy gyakorlatban kimondottan népszerű adaptív Runge–Kutta típusú módszer, a Dormand–Prince [9] (röviden DOPRI vagy DOPRI54, 2.6. táblázat) ismertetésével és egy alkalmazásával zárom.

A két megoldás rendje 4, illetve 5, emellett az ötödrendű megoldás paraméterei úgy lettek megválasztva, hogy hibájában h^6 együtthatóját minimalizálják. Vegyük észre, hogy az $\mathbf{A}$

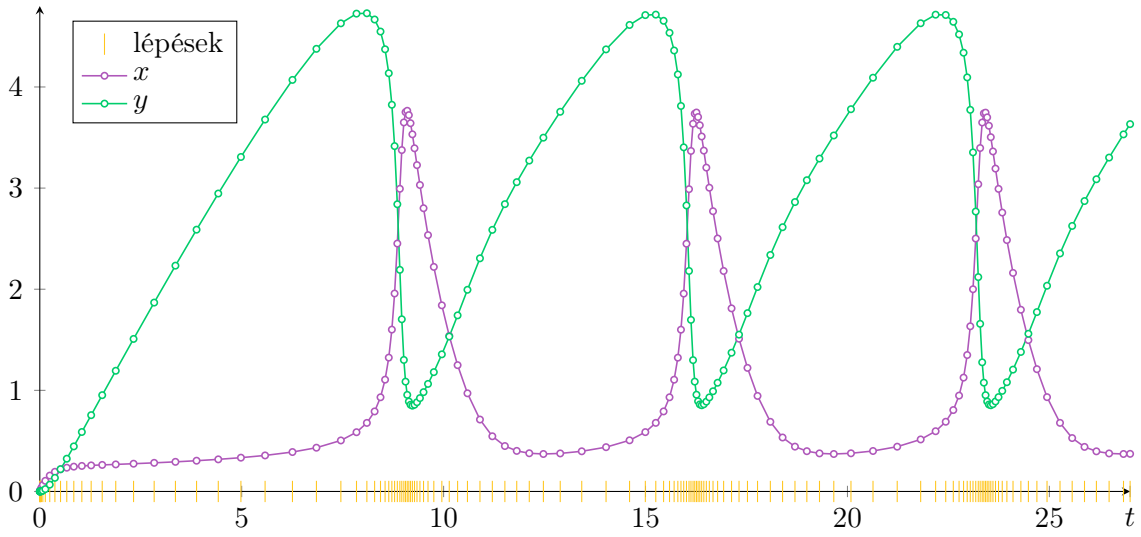
[5] Ezen leggyakoribb választás mellett más, például p, illetve $(p-1)$ rendekkel is szokás dolgozni.

[6] Gyakran csak "módszer"-nek nevezzük a módszerpárt.

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0

2.6. táblázat. Dormand–Prince módszer kibővített Butcher-tablója.

együtthatómátrix utolsó sora megegyezik a $\widehat{\mathbf{b}}$ vektorral, így a következő lépésben f egyik kiértékelését újrahasználhatjuk, azaz a 7 lépcső ellenére csupán 6 függvény-kiértékelés szükséges egy lépéshez. Ezen tulajdonságoknak köszönhetően igen jól alkalmazható valós feladatok megoldásában. A lépésköz adaptivitásának hasznosságát jól mutatja a 2.3. ábrán látható Brusselator egyenlet (Appendix A) numerikus megoldása a DOPRI módszerrel.



2.3. ábra. Brusselator megoldása DOPRI módszerrel $\varepsilon = 10^{-6}$ tolerancia mellett.

Mint az a fenti ábrából is látható, a megoldás rövid, gyorsan változó részeket tartalmaz, és ezt a hibabecslésnek köszönhetően a lépésköz mérete megfelelően követi. Az elmúlt évtizedek során hatalmas fejlődés ment végbe a lépésköz-megválasztás elméletében [29,30, 32], viszont dolgozatom terjedelmére való tekintettel a modern elméletet nem tárgyalom.

II.3. Implicit Runge–Kutta módszerek

Az eddigiekben a Runge–Kutta típusú módszerek közül csak explicit módszerekkel foglalkoztunk. Ezeket egy $\mathbf{A} \in \mathbb{R}^{s\times s}$ szigorú alsó háromszögmátrix és $\mathbf{c}, \mathbf{b} \in \mathbb{R}^s$ vektorok által alkotott Butcher-tablóval tudjuk definiálni. Egy általános Runge–Kutta típusú módszer (RK) annyiban bővíti az ERK fogalmát, hogy az $\mathbf{A}$ együtthatómátrixra nem helyez szigorú alsó háromszög megkötést. Egy olyan RK módszert, melyre $\mathbf{A}$ nem szigorú alsó háromszögmátrix *implicit* Runge–Kutta típusú módszernek (IRK) nevezzük. Mint azt a II.2. fejezetben láttuk, az $\mathbf{Ae} = \mathbf{c}$ feltevés nem jelent kimondott megkötést, így ezt ezentúl a RK típusú módszerek definíciójának részeként fogjuk tekinteni.

Az IRK egy lépése során definíció szerint lesz olyan lépcső, amiben a k_i érték egy implicit egyenlettel lesz megadva, így ennek kiszámítását nem lehet az ERK módszerekhez hasonlóan egymás után, lépcsőnként meghatározni, hanem egy (általában nemlineáris) egyenletrendszert kell megoldanunk. Ehhez az implementáció során valamilyen (tipikusan Newton-féle) iterációt kell alkalmazni.

J. C. Butcher *"Implicit Runge–Kutta processes"* című cikkének eredeti verzióját [2] visszaküldték a szerkesztők, mivel nem volt világos, hogy az IRK egy lépését definiáló egyenletrendszernek létezik-e egyértelmű megoldása[7]. Mint kiderült [13, II.7], ehhez f Lipschitz-folytonosságán kívül h-ra is bizonyos megkötésnek kell teljesülnie. A fejezet hátralévő részére a megoldás létezését és egyértelműségét mindig fel fogom tenni. Ennek realitását és következményeit [12, IV.8, IV.14] részletesen tárgyalja, viszont dolgozatomban nem térek ki rá.

Bevezetésként tekintsük a talán első IRK módszert, melyet Cauchy írt le 1824-ben [5]. Ehhez vegyük a (2.1) feladatot az alábbi

$$u(t_{n+1}) = u(t_n) + \int_{t_n}^{t_{n+1}} f(s, u(s))\mathrm{d}s$$

integrál alakban. A Lagrange-féle középértéktétel szerint valamely $\theta_1, \theta_2 \in (0,1)$-re

$$u(t_{n+1}) = u(t_n) + hf\big(t_n + \theta_1 h, u(t_n) + \theta_2(u(t_{n+1}) - u(t_n))\big)$$

egyenlőség teljesül. Ekkor a $\theta_1 = \theta_2 = \theta$ paraméterválasztás az

$$y_{n+1} = y_n + hf\big(t_n + \theta h, y_n + \theta(y_{n+1} - y_n)\big)$$

[7] Ennek feloldását a publikáció appendixe tartalmazza.

numerikus módszerhez vezet, melyet *θ-módszernek* nevezzük. Megfigyelhető, hogy míg a $\theta = 0$ eset az EE módszer, a $\theta = 1$ választás az eddigiektől eltérő

$$k_1 = f(t_n, y_{n+1})$$
$$y_{n+1} = y_n + hk_1$$

implicit módszert eredményezi. Sorbafejtéssel belátható, hogy az így kapott, úgynevezett *implicit Euler* (röviden IE) is elsőrendű. Most tekintsünk két érdekes IRK módszert, az első a θ-módszer egy speciális esete ($\theta = 1/2$), az *implicit középponti szabály*:

$$k_1 = f(t_n + h/2, y_n + h/2 \cdot y_{n+1})$$
$$y_{n+1} = y_n + hk_1.$$

A második egy kétlépcsős

$$k_1 = f(t_n, y_n)$$
$$k_2 = f(t_n + 2h/3, y_n + h/3 \cdot (k_1 + k_2)))$$
$$y_{n+1} = y_n + h/4 \cdot (k_1 + 3k_2)$$

módszer Hammer és Hollingsworth-től [14]. Meglepő módon az első módszer a rendfeltételeket másodrendig, míg a második harmadrendig teljesíti, így a lépcsőszámok ellenére a konzisztenciarendjeik 2, illetve 3. Ez azzal magyarázható, hogy az $\mathbf{A}$ együtthatómátrix és $\mathbf{b}$ súlyvektor paramétereinek szabadsági foka IRK módszereknél jelentősen nagyobb, mint ERK módszerek esetén. Az eddigi módszerek és még egy, a trapézszabályon alapuló módszer Butcher-tablóját mutatja be a 2.7. táblázat.

1	1
	1

1/2	1/2
0	1

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

0	0	0
1	1/2	1/2
	1/2	1/2

2.7. táblázat. Balról jobbra az IE, implicit középponti szabály, Hammer–Hollingsworth és az implicit trapéz-módszer Butcher-tablója.

Kuntzmann és Butcher felfedezték, hogy minden s-re létezik olyan IRK típusú módszer, melynek konzisztenciarendje $2s$ [2], ezeket *maximális rendű* IRK módszereknek nevezzük. Egylépcsősre már láttunk példát, az implicit középponti szabályt. Magasabb lépcsőszámú, maximális rendű módszereket pedig a 2.8. táblázat tartalmaz. Megfigyelhető, hogy ezen módszerek c_i értékei egy nagyon speciális módon lettek megválasztva, éppen a Gauss- vagy

Gauss–Legendre-kvadratúra alappontjainak felelnek meg. Éppen ezért szokás az ilyen IRK módszereket *Gauss–Legendre* módszereknek is nevezni (röviden Gauss$p$, ahol p a módszer rendje).

$$\begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} \frac{1}{2}-\frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9}-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36}-\frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{1}{4}-\frac{\sqrt{15}}{10} \\ \frac{1}{2}+\frac{\sqrt{15}}{10} & \frac{5}{36}-\frac{\sqrt{15}}{30} & \frac{2}{9}+\frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

2.8. táblázat. Negyedrendű Gauss4 és hatodrendű Gauss6 módszerek Butcher-tablója.

Ilyen és ehhez hasonló sok másik IRK meghatározható egy, polinom interpoláción alapuló numerikus integráláshoz hasonló, úgynevezett *kollokációs* módszer segítségével. Az így kapott numerikus eljárást szintén kollokációs módszernek nevezzük. Nevezetes kollokációs módszer-család, a *Lobatto*: melyben a Gauss-féle c_i alappontok úgy kerülnek módosításra, hogy az első és utolsó alappont az intervallum végpontjaival essen egybe, vagyis $c_1 = 0$ és $c_s = 1$, ezek rendje $2s - 2$. Másik hasonlóan népszerű módszer-család a *Radau*, ahol csak az első (Radau IA) vagy az utolsó (Radau IIA) alappont kerül a lépés szélére, ezek rendje $2s-1$. Felmerülhet a kérdés, hogy miért is van szükség implicit módszerekre, mikor a nemlineáris egyenletek minden lépésben való megoldása óriási mértékben megnöveli a számítási munkát. Erre a kérdésre a következő alfejezetben választ adunk.

II.3.1. Stabilitás és merevség

Most megvizsgálunk a már említett "bizonyos" feladatok közül néhányat, minek során fény derül IRK típusú módszerek olyan kedvező tulajdonságaira, melyek ERK típusú módszereknél nem elvárhatóak. Először tekintsük a Dahlquist-féle lineáris tesztegyenletet:

$$u'(t) = \lambda u(t), \quad t \in [0, T], \quad u(0) = 1, \quad \lambda \leq 0^{8}. \tag{2.10}$$

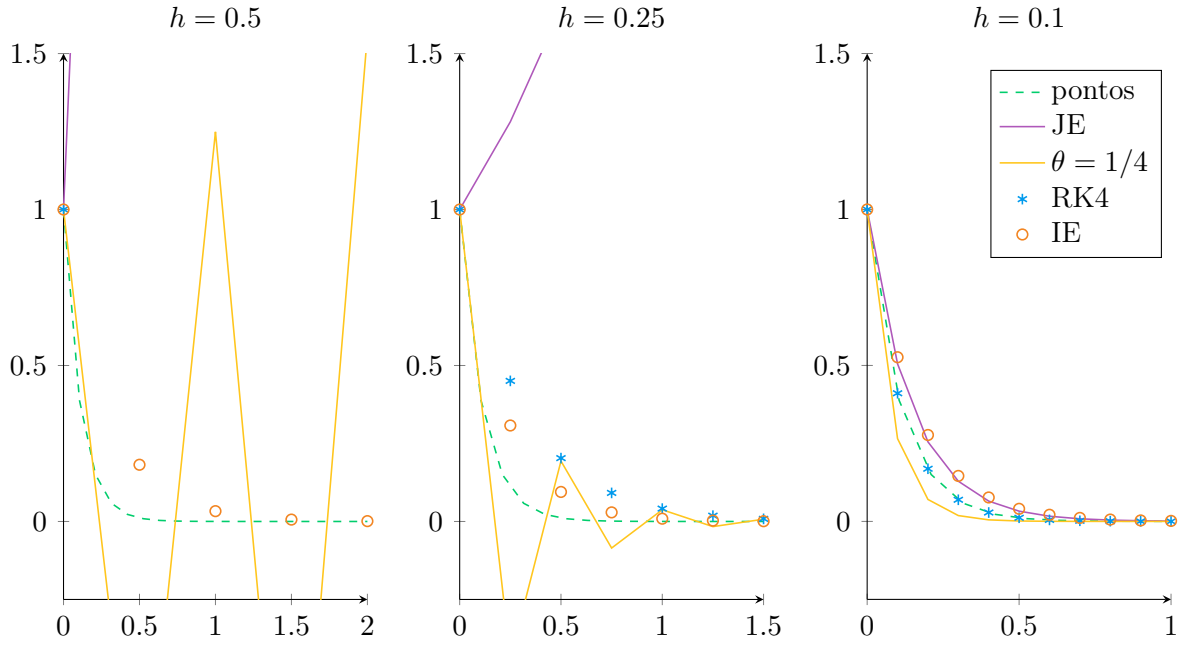
Majd oldjuk meg az EE és IE módszerekkel a $[0, 1]$ intervallumon, különböző λ és h értékek mellett is. Az $u(t) = e^{\lambda t}$ pontos megoldás ismeretében könnyen kiszámolható a hiba is, melyet a 2.9. táblázatban láthatunk. Figyeljük meg, hogy $\lambda = -9$ paraméterrel a hiba

[8] Gyakorlatban sokszor többdimenziós CF-ot kell megoldanunk, és az ekkor megjelenő sajátértékekről nem minden esetben feltehető, hogy valósak. Így általános esetben a $\mathrm{Re}(\lambda) \leq 0$ feltétellel dolgozunk.

	$\lambda = -9$		$\lambda = -99$		$\lambda = -999$	
h	EE	IE	EE	IE	EE	IE
0.1	$3.07\text{e}-01$	$1.20\text{e}-01$	$3.12\text{e}+09$	$9.17\text{e}-02$	$8.95\text{e}+19$	$9.93\text{e}-03$
0.01	$1.72\text{e}-02$	$1.60\text{e}-02$	$3.62\text{e}-01$	$1.31\text{e}-01$	$2.38\text{e}+95$	$9.09\text{e}-02$
0.001	$1.71\text{e}-03$	$1.60\text{e}-03$	$1.90\text{e}-02$	$1.75\text{e}-02$	$3.67\text{e}-01$	$1.32\text{e}-01$
0.0001	$1.66\text{e}-04$	$1.65\text{e}-04$	$1.78\text{e}-03$	$1.68\text{e}-03$	$1.92\text{e}-02$	$1.76\text{e}-02$
0.00001	$1.66\text{e}-05$	$1.66\text{e}-05$	$1.82\text{e}-04$	$1.18\text{e}-04$	$1.83\text{e}-03$	$1.83\text{e}-03$

2.9. táblázat. Hiba maximumnormában, különböző λ értékekekre.

pontosan a vártnak megfelelően alakul. Viszont a többi λ értékre az EE csak megfelelően kis h-ra ad használható eredményt, míg az IE továbbra is megfelelően működik. Ez a gyakorlati alkalmazások szempontjából gondot jelenthet, hiszen kis h mellett sok lépést kell végrehajtani, ami a számítások idejének megnövekedése mellett a gépi hibák felhalmozódásának lehetőségét eredményezi. Mint azt a 2.4. ábra mutatja, ez a jelenség nem az EE módszer sajátos tulajdonsága, minden explicit és számos implicit RK módszernél is jelen van. Viszont az IE a fenti táblázatnak megfelelően viselkedik minden adott paraméterre.



2.4. ábra. A $\lambda = -9$ paraméterű tesztfeladat megoldása különböző lépésközök mellett.

Ezen jelenség megértéséhez először vizsgáljuk meg közelebbről az EE és IE egy általános lépését a (2.10) tesztfeladaton:

$$\begin{aligned} \text{EE}: &\quad y_{n+1} = (1+h\lambda)y_n, \quad n = 0,1,\ldots,N-1 \quad y_0 = 1, \\ \text{IE}: &\quad y_{n+1} = \frac{1}{(1-h\lambda)}y_n, \quad n = 0,1,\ldots,N-1 \quad y_0 = 1. \end{aligned}$$

Ezt általánosan egy módszertől függő $R(h\lambda)$, úgynevezett *stabilitási függvénnyel*

$$y_{n+1} = R(h\lambda)y_n \tag{2.11}$$

formában írhatjuk le[9]. Bevezetve a $z = h\lambda$ jelölést, majd a (2.11) egyenletet iterálva a módszerek n-edik pontban vett megoldását előállíthatjuk $R(z)^n y_0$ alakban. Így könnyen látható, hogy a numerikus módszer megoldásának korlátossága megfogalmazható $R(z)$ függvényében is, ami azért fontos, mert a tesztfeladat pontos megoldása a $\mathrm{Re}(\lambda) \leq 0$ feltétel következtében mindig korlátos, így csak olyan numerikus megoldás közelítheti megfelelően, amely szintén hasonló tulajdonságú, azaz $R(z) \leq 1$. Ez valós, negatív λ esetén az EE, illetve IE módszerekre a következő feltételeket eredményezi:

$$\begin{aligned} |R_{\text{EE}}(z)| \leq 1 \quad &\Longleftrightarrow \quad h \leq \frac{2}{-\lambda}, \\ |R_{\text{IE}}(z)| \leq 1 \quad &\Longleftrightarrow \quad h > 0. \end{aligned}$$

Mivel h minden esetben pozitív, az IE mindig korlátos megoldást fog eredményezni, míg az EE csak megfelelően kis h-ra. Ez pedig összhangban áll a 2.9. táblázattal.

Mivel egy numerikus módszer konvergenciája csupán a $h \to 0$ melletti viselkedését jellemzi, egy rögzített rácshálón való működéséről nem ad információt. Emiatt van kiemelkedő szerepe azon módszereknek, melyek tetszőleges rácshálón is visszaadják a pontos megoldás fontosabb tulajdonságait, speciálisan a korlátosságot.

2.9. Definíció *Azon $z \in \mathbb{C}$ számok halmazát amelyre az*

$$|R(z)| \leq 1$$

feltétel teljesül, a numerikus módszer stabilitási tartományának *nevezzük. Azt mondjuk, hogy egy numerikus módszer* A-stabil, *ha a stabilitási tartománya tartalmazza a bal oldali* $\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Re}(z) \leq 0\} \subset \mathbb{C}$ *komplex félsíkot.*

[9] A két fenti esetben ez $R_{\text{EE}}(h\lambda) = 1 + h\lambda$ és $R_{\text{IE}}(h\lambda) = 1/(1-h\lambda)$-t jelenti.

Vagyis az A-stabil módszerek pont azok, melyek egy lineáris differenciálegyenletre minden $h > 0$ lépésközzel korlátos megoldást adnak, hogyha a pontos megoldás korlátos. A definíció Germund Dahlquisttól származik, aki először 1963-ban többlépéses módszerekre fogalmazta meg ezt a stabilitási tulajdonságot, azzal a kis eltéréssel, hogy $\mathrm{Re}(z) < 0$ esetén a numerikus megoldás 0-hoz való konvergenciáját követelte meg[10] [8]. A stabilitási függvények ismeretében könnyű ellenőrizni, hogy az IE A-stabil, míg az EE a valós tengelyen sem teljesíti a tulajdonságot. Hogy más RK típusú módszerek stabilitási tartományát is meg tudjuk vizsgálni, először azok stabilitási függvényét kell kiszámolnunk.

2.10. Tétel [16, 4.14] *Egy Runge–Kutta módszer $R(z)$ stabilitasi függvénye*[11]

$$R(z) = \frac{\det(\mathbf{I} - z\mathbf{A} + z\mathbf{e} \cdot \mathbf{b}^\top)}{\det(\mathbf{I} - z\mathbf{A})}.$$

Figyeljük meg, hogy amennyiben $\mathbf{A}$ egy szigorú alsó háromszögmátrix, azaz a módszer explicit, a nevező azonosan 1. Azaz $R(z)$ z egy polinomja, tehát a módszer stabilitási tartománya mindig korlátos. Így kapjuk a már korábban említett fontos eredményt:

2.11. Tétel *Az explicit Runge–Kutta típusú módszerek nem A-stabilak.*

Általános RK típusú módszerek esetén a 2.10. Tétel alapján $R(z)$ egy racionális törtfüggvény, ahol a nevezőben szereplő polinom implicit módszerekre legalább elsőfokú. A tesztfeladat pontos megoldását összehasonlítva egy p-edrendű numerikus megoldással az

$$e^z - R(z) = \mathcal{O}(h^{p+1}) = \mathcal{O}(z^{p+1})$$

összefüggést kapjuk. Tehát a stabilitási függvényről kiderül, hogy e^z egy p-edrendű Padé típusú közelítése. Így az s-lépcsős, s-edrendű ERK módszerek stabilitási függvénye az

$$R(z) = 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^s}{s!}$$

alakot ölti. Most tekintsük a θ-módszer

$$R(z) = \frac{1 + (1-\theta)z}{1 - \theta z}$$

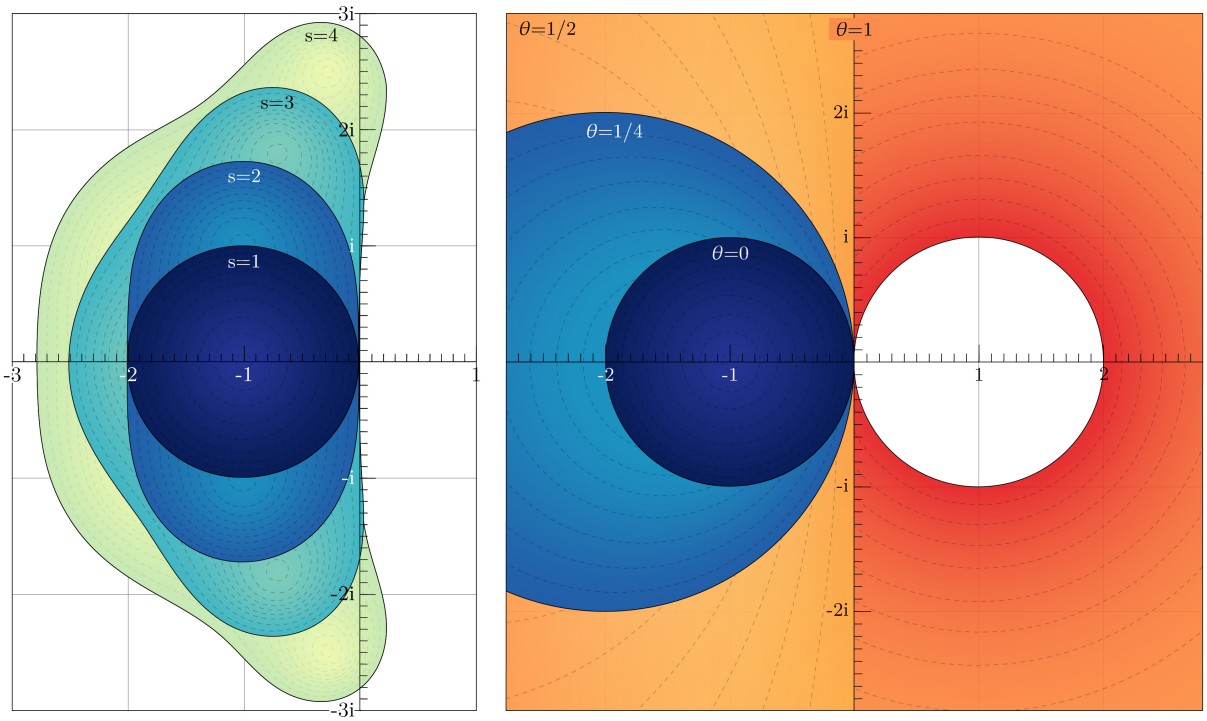
stabilitási függvényét, mely $\theta = 0$ és 1 választással R_{EE}, illetve R_{IE}, míg $\theta = 1/2$ az

$$R(z) = \frac{1 + 1/2 \cdot z}{1 - 1/2 \cdot z}$$

függvényt eredményezi, ami éppen megegyezik a trapéz-módszer stabilitási függvényével. Az így kiszámítható stabilitási tartományokat mutatja be a 2.5. ábra.

[10] Ez a stabilitási függvénnyel megfogalmazva az $R(z) < 1$ feltételt jelenti.

[11] Az $\mathbf{e} \cdot \mathbf{b}^\top$ kifejezés a két vektor diadikus szorzata, azaz $\mathbf{e} \cdot \mathbf{b}^\top \in \mathbb{R}^{s \times s}$.



2.5. ábra. Bal oldalon s-lépcsős s-edrendű ERK módszerek stabilitási tartománya, jobb oldalon a θ-módszer stabilitási tartománya különböző θ értékekre.

Mint látható, az ERK módszerek mindegyike és a θ-módszer $\theta < 1/2$ értékekre korlátos stabilitási tartománnyal rendelkezik. Viszont $\theta = 1/2$-re a θ-módszer (az implicit középponti szabály) stabilitási tartománya éppen egybeesik a bal oldali komplex félsíkkal, míg $\theta > 1/2$-re a bal oldali komplex félsíkon túl a jobb oldali félsík nagy részét lefedi. Tehát a trapéz-módszer, $\theta > 1/2$-re a θ-módszer és így az IE is A-stabil. Mint azt láthattuk, az A-stabil RK módszerek akkor előnyösek a tesztfeladaton explicit módszerekkel szemben, amikor $\mathrm{Re}(\lambda) \ll 0$. Ebben az esetben a pontos megoldás egy rövid, gyorsan változó – úgynevezett *tranziens* – szakasz után szinte állandó. Ez a jelentős eltérés a pontos megoldás különböző időskálái között nagyban jellemez sok, természettudományokban felbukkanó folyamatot. Ennek köszönhetően az A-stabilitás egy valóban fontos fogalommá vált.

Alkalmazásokban gyakran előfordulnak olyan kezdetiérték feladatok, melyek numerikus megoldása során stabilitási tulajdonságok jelentik a főbb akadályt. Így az, hogy milyen módszert választunk, egy kritikus fontosságú kérdéssé válhat. (Erre egy példa a fenti tesztfeladat megfelelő paraméterekkel.) Ugyanakkor annak eldöntése, hogy melyik CF kíván különös figyelmet korántsem egyszerű. A *merev* feladatok egy hatalmas ilyen, speciális bánásmódot igénylő családot alkotnak. Annak ellenére, hogy számtalan területen bukkannak

fel és elméletük mára hatalmasra nőtt, nagyon sokáig nem született kielégítő definíciójuk. Az első megfogalmazás Curtishez és Hirschfelderhez köthető [7], mely Hairer és Wanner [12, IV.1] merev feladatokat bevezető fejezetét is nyitja, mely magyarul az alábbi módon hangzik: *"merev feladatok azok, melyeken explicit módszerek nem működnek."* Azóta számos vetélkedő definíció született, ezt jól leírja Gustaf Söderlind 2015-ös *"... Sixty years in search of a definition"* publikációjának címe is, azaz már több mint 60 éve foglalkoztatja a matematikus közösséget ez a fogalom [31]. Ebben a publikációban a fogalom történelmének összefoglalása mellett, egy funkcionálanalitikus megközelítésű definíciót is kaphatunk. A további fogalmak tekintetében, inkább a merevségre jellemző tulajdonságok lesznek fontosak, így a dolgozatom a definíciót nem fogja tárgyalni.

Sok merev feladat lassan változó megoldásában jelen vannak tranziens szakaszok[12]. Tekintsük erre példaként az alábbi, kémiai reakciót leíró egyenletrendszert:

$$\begin{aligned} u_1'(t) &= -0.04u_1(t) + 10^4 u_2(t)u_3(t), && \text{(lassú)} \\ u_2'(t) &= 0.04u_1(t) - 10^4 u_2(t)u_3(t) - 3\cdot 10^7 u_2^2(t), && \text{(nagyon gyors)} \\ u_3'(t) &= 3\cdot 10^7 u_2^2(t), && \text{(gyors)} \end{aligned} \tag{2.12}$$

ahol a kezdeti feltétel $u_1(0) = 1,\ u_2(0) = 0,\ u_3(0) = 0$. Ennek megoldásához három különböző adaptív numerikus eljárást fogok alkalmazni, koordinátánként implementálva. Az első a MATLAB beépített ode45 differenciálegyenlet-megoldója[13]. A második, a már korábban ismertetett DOPRI egy egyszerű lépésköz-megválasztó eljárással, és végül az implicit középponti módszer, avagy Gauss2 a Richardson extrapoláción alapuló lépésköz választással. A numerikus eredmények a 2.6. ábrán láthatók. Vegyük észre, hogy mindhárom ERK típusú módszerrel tett próbálkozás több, mint 100 lépést igényelt és erősen "zajos" közelítő megoldást adott. Míg az A-stabil Gauss2 mindössze 13 lépést igényelt és a megoldása sima. Emellett érdemes azt is megfigyelni, hogy a toleranciaérték csökkenése nem eredményezte a lépésköz jelentős változását[14].

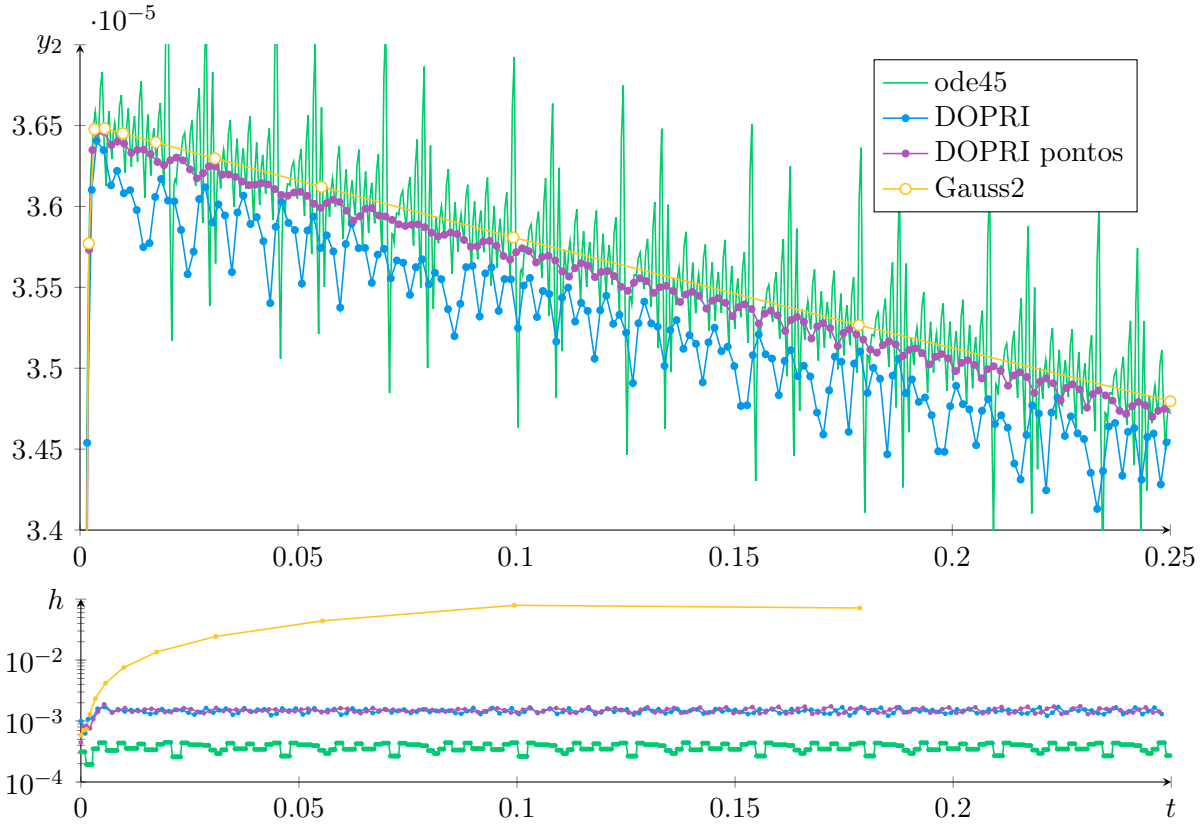
A fentiek alapján könnyen juthatnánk arra a következtetésre, hogy merev feladatokat A-stabil módszerekkel mindig zökkenőmentesen tudunk kezelni. A következő

$$u'(t) = -999(u(t) - \cos(t)), \quad u(0) = 0 \tag{2.13}$$

[12] Ugyanakkor ez a feltétel se nem szükséges, se nem elégséges.

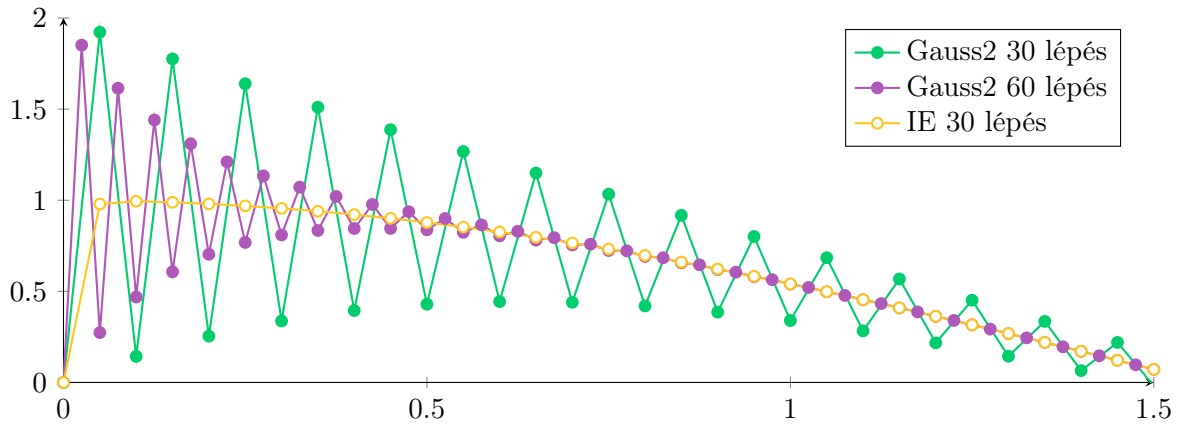
[13] Dormand–Prince módszeren alapul egy szofisztikált lépésköz-megválasztási technikát alkalmazva.

[14] Ez egy, a dolgozatom terjedelmén túlmutató számításmatematikai stabilitással hozható összefüggésbe.



2.6. ábra. A (2.12) CF numerikus megoldása a hozzá tartozó lépéshosszakkal ε, ATol, RTol $= 10^{-6}$, illetve "DOPRI pontos" esetében $\varepsilon = 2 \cdot 10^{-7}$ toleranciaértékek mellett.

egyszerű differenciálegyenlet erre ad ellenpéldát, melynek pontos megoldása egy rövid tranziens szakasz után $\approx \cos(t)$. Most pedig tekintsük meg az implicit középponti módszer és az IE módszer közelítő megoldását a 2.7. ábrán.



2.7. ábra. A (2.13) CF numerikus megoldása különböző A-stabil módszerekkel.

A 2.7. ábra szerint az implicit középponti szabály ugyan időben tart a pontos megoldáshoz, az oszcillálás mértéke nagyobb lépésköz mellett használhatatlanná teszi az eredményt, míg az IE egyáltalán nem mutat oszcillációt. Ez a jelenség azzal magyarázható, hogy a módszer $R_{\text{Gauss2}}(h\lambda)$ stabilitási függvénye $\text{Re}(\lambda) \ll 0$ esetben ugyan kisebb, mint 1, de nagyon közel van hozzá. Ez azt eredményezi, hogy a $\cos(t)$-től vett eltérés exponenciális módon, de lassan cseng le. Tehát azon merev feladatok, melyek perturbációi különösen gyorsan kerülnek csillapításra, olyan módszerrel kezelhetők jól, melyek R stabilitási függvénye nem rendelkezik a fenti problémával.

2.12. Definíció *Egy numerikus módszert* L-stabilnak *nevezünk, amennyiben* A*-stabil és*

$$\lim_{z\to\infty} R(z) = 0. \tag{2.14}$$

Könnyen látható, hogy az IE esetében a fenti határérték valóban 0, és mivel A-stabil, így L-stabil is. Hasonló módon ellenőrizhető, hogy a Gauss2 módszer nem L-stabil[15]. Vagyis sikerült megoldani a fenti, nemkívánatos jelenséget azon az áron, hogy egy szigorúbb stabilitási fogalmat vezettünk be. Ezzel is jól mutatva, hogy a merev feladatok megoldását többnyire stabilitás vezérli.

A fejezetet egy olyan jelenséggel zárom, mely – az eddigiekkel ellentétben – nem kíván tranziens szakaszt a pontos megoldásban. Ez a jelenség a *rendcsökkenés*, ami implicit módszerek merev feladatokra való alkalmazásában az elmélet szerinti és az adott feladaton valóban megfigyelt konvergenciarend eltérését jelenti. Tekintsük a következő

$$\begin{aligned} u_1'(t) &= -(\mu+2)u_1(t) + \mu u_2^2(t), \qquad & u_1(0) &= 1, \\ u_2'(t) &= u_1(t) - u_2(t) - u_2^2(t), & u_2(0) &= 1 \end{aligned} \tag{2.15}$$

valós μ-vel paraméterezett KDE rendszert a $[0,1]$ intervallumon, melynek pontos megoldása

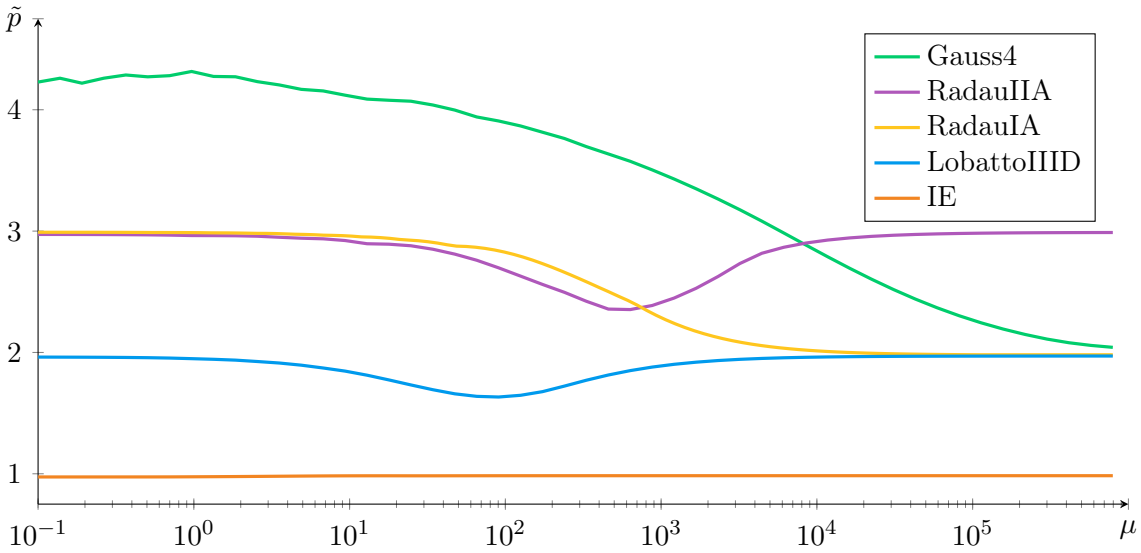
$$u_1(t) = e^{-2t} \quad \text{és} \quad u_2(t) = e^{-t}.$$

Ebben a példában a megoldás nem függ a μ paramétertől. Viszont a CF *merevsége* igen, pontosabban, nagyobb paraméter *merevebb* feladatot fog jelenteni [27]. Most oldjuk meg a (2.15) feladatot különböző IRK típusú módszerekkel. Jelölje a megfigyelt rendet $\tilde{p}$ és becsüljük a II.2.2. fejezetben bemutatott extrapolációs technika átalakításával kapott

$$\tilde{p} = \lg(e(0.1)/e(0.01)) \tag{2.16}$$

[15] Olyan módszerekről, melyek stabilitási tartománya éppen egybeesik a bal oldali komplex félsíkkal, belátható, hogy a (2.14) határérték mindig 1, így ezek nem L-stabiliak.

eljárással, ahol $e(h)$ a lépések során vett legnagyobb hibát jelöli h lépésköz mellett[16]. Most pedig vessük össze az így kiszámítható rend becslését különböző μ paraméterekkel.



2.8. ábra. A (2.16)-os $\tilde{p}$ rend becslések a μ paraméterű (2.15) CF függvényében.

Figyeljük meg, hogy amennyiben μ nagyságrendileg közel áll a lépésközhöz (nem merev eset), a megfigyelt rend jó becslése az elmélet szerinti rendnek. Viszont nagyságrendileg magasabb μ értékekre (merev eset) a konvergenciarend több módszernél is csökken és azt nem mindegyik nyeri vissza. Ezt a jelenséget először Prothero és Robinson magyarázta meg egy (2.10) általánosításaként kapott

$$u'(t) = \lambda(u(t) - \phi(t)) + \phi'(t), \quad t \in [0, T]$$

tesztegyenlet segítségével [24]. Ekkor, ezen a λ-val paraméterezett egyenletcsaládon két eltérő konvergenciarendről is beszélhetünk. Az első a klasszikus rend, mely a módszer hibájának viselkedését írja le $h \to 0,\ h\lambda \to 0$ esetben. A másik, egy merev feladatok jellemzése során felmerülő $h \to 0,\ h\lambda \to \infty$, úgynevezett *merev-határérték* mentén megjelenő konvergenciarend [17]. Az utóbbi eset kezelésére új fogalmakat kell bevezetni (lásd [17, 27]). Dolgozatomban nem térek ki ezek tárgyalására, csupán megemlítem, hogy lehetséges olyan feltételeket szerkeszteni, melyek garantálni fogják a rend visszanyerését. Továbbá léteznek úgynevezett *merev-pontos* módszerek is (implicit trapéz, vagy az ábrán IE), melyek teljes mértékben ellenállnak a rendcsökkenésnek.

[16] Itt a *hiba* a kétkomponensű eltérés euklideszi normája.

III. Nemstandard véges differencia módszerek

A fejezet Ronald E. Mickens *"Nonstandard Finite Difference Schemes"* és *"Advances in the Applications of Nonstandard Finite Difference Schemes"* című könyveinek releváns részeit dolgozza fel[1]. A cél, hogy betekintést nyerjünk közönséges differenciálegyenletek egy, a standard elmélettől eltérő megoldási módjába. Az NSFD *metodika* egy általános szemléletmód arra vonatkozóan, hogy hogyan is kell véges differencia módszereket szerkeszteni. A szemlélet egyik kulcsfontosságú része az, hogy minden egyes differenciálegyenletet más módon kell megközelítenünk, annak függvényében, hogy maguk az egyenletek mit modelleznek, és hogy milyen célt szolgál ezek megoldása.

Mivel a standard elmélet nem egy adott DE megoldására nyújt megoldó módszert, fenn áll a kérdés: ha egyáltalán létezik, akkor melyik módszer a legjobb egy adott probléma kezelésére? Ezen egyértelműség hiányát próbálja kiküszöbölni az NSFD metodika egy extra megkötéseket nyújtó elv, a *dinamikus konzisztencia* segítségével.

A módszerek szerkesztését vezérlő elv bevezetése érdekében tekintsük az alábbi

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(u, t, \lambda), \quad u(0) = u_0 \tag{3.1}$$

közönséges differenciálegyenletet, ahol $\lambda \in \mathbb{R}^m$ az egyenletet meghatározó paraméter. Továbbá legyen (3.1) egy véges differencia reprezentációja

$$y_{n+1} = F(y_n, y_{n+1}, t_n, h, \lambda), \quad t_n = nh, \tag{3.2}$$

ahol h a lépésköz és y_n éppen $u(t_n)$ közelítése.

3.1. Definíció *Legyen a* (3.1) *alakú DE és/vagy ennek megoldásai P tulajdonságúak. Azt mondjuk, hogy* (3.2) dinamikusan konzisztens P tekintetében (3.1)*-el, amennyiben* (3.2) *és/vagy ennek megoldásai szintén rendelkeznek a P tulajdonsággal.*

[1] Ahogy azt az irodalomban megfigyelni véltem, az NSFD világban sokkal gyakrabban használt a "séma", mint a "módszer" kifejezés, mikor véges numerikus eljárásokról esik szó. Így ezt dolgozatom is tükrözni fogja azzal, hogy nemstandard elemeket tartalmazó eljárásokra sémaként fogok hivatkozni.

Vagyis minél több P tulajdonság tekintetében dinamikusan konzisztens (röviden DK) egy véges differencia egyenlet, megoldása annál "jobb" közelítése a DE megoldásának. Ez akkor nagyon hasznos, amikor elvárás, hogy a numerikus közelítés a pontos megoldás bizonyos kvalitatív tulajdonságait megőrizze. Emlékezzünk vissza, hogy a II.3.1. fejezetben is egy kvalitatív tulajdonság vizsgálata vezetett az A-stabilitáshoz. Gyakran előforduló, fontos P tulajdonságok lehetnek [6, 11, 23]:

- A megoldás korlátossága.
- A megoldás monotonitása (speciálisan oszcillációk kizárása).
- A megoldás nemnegativitása.
- Az egyensúlyi pontok száma és stabilitásuk.
- Megmaradási törvények (például energiamegmaradás).
- Periodikus és más speciális megoldások.

Bevezető példaként tekintsük az

$$u'(t) = \lambda u(t), \quad u(0) = u_0, \quad \lambda < 0 \tag{3.3}$$

tesztfeladatot. Figyeljük meg, hogy a pontos megoldás ismerete nélkül is rögzíthetünk bizonyos tulajdonságokat, amit egy "jó" közelítéstől elvárnánk:

- P_1: $u(t) = 0$ egy egyensúlyi pont;
- P_2: $u(t)$ abszolút értéke szigorúan monoton tart a nullához;
- P_3: feltéve, hogy $u_0 \neq 0$ minden $t > 0$-ra $u_0 \cdot u(t) > 0$.

Az

$$u(t) = \frac{u^2(t)}{u(t)} \approx \frac{y_n y_{n+1}}{(y_n + y_{n+1})/2}$$

közelítést használva, majd az EE módszert alkamazva adódik az

$$\frac{y_{n+1} - y_n}{h} = \lambda \frac{2y_n y_{n+1}}{y_n + y_{n+1}} \tag{3.4}$$

séma, melynek a nemnegatív megoldása[2]

$$y_{n+1} = h\lambda y_n + y_n\sqrt{1 + h^2\lambda^2} = \left(\sqrt{1 + h^2\lambda^2} + h\lambda\right) y_n. \tag{3.5}$$

[2] Átrendezve y_{n+1}-ben másodfokú, így explicit megadható.

Számolással belátható, hogy minden $h > 0, \lambda < 0$ értékre

$$0 < \sqrt{1 + h^2\lambda^2} + h\lambda < 1,$$

azaz a (3.5) által megadott ($y_0 = u_0$) explicit numerikus módszer P_1, P_2 és P_3 tulajdonság tekintetében is DK a (3.3) tesztfeladattal. Vagyis a módszer által adott közelítő megoldás a (3.1) feladatra minden lépésköz mellett korlátos. Ezen eredményt érdemes összevetni a 2.11. Tétel állításával, miszerint explicit RK módszerek nem lehetnek A-stabilak, ennek következménye, hogy csak implicit RK módszerek tudják garantálni a tesztfeladatra adott közelítő megoldás korlátosságát tetszőleges lépésköz mellett. Tehát egy olyan módszert szerkesztettünk, mely explicitsége ellenére implicit RK módszerek tulajdonságával rendelkezik. Ezt az előnyös tulajdonságot az NSFD metodika nyelvén úgy fogalmaznánk meg, hogy a (3.3) DE (3.4) diszkrét modellje *numerikus instabilitás mentes*, vagyis a diszkrét modellnek nincs olyan megoldása, ami nem feleltethető meg a DE egy megoldásával sem. Ellenkező esetben azt mondjuk, hogy egy DE diszkrét modellje *numerikus instabilitással* rendelkezik, hogyha létezik olyan megoldása, mely a DE egy lehetséges megoldásának sem feleltethető meg kvalitatív módon[3].

A DK elv bevezetésének az egyik fő célja többek között az, hogy numerikus instabilitásoktól mentes, vagy ezeket gyorsan csillapító diszkretizáció szerkesztését tegye lehetővé. Ehhez pedig szükséges a feladat alapos ismerete, hogy a numerikus instabilitásokat okozó kvalitatív tulajdonságban való eltéréseket felismerjük. A II.3.1. fejezetben oszcilláló, vagy nem korlátos megoldások formájában már láttunk példát numerikus instabilitásokra. Ezek megjelenésének alapvető oka az, hogy a véges differencia modellek paramétertere mindig bővebb lesz, mint általa modellezett DE. Ennek jó szemléltetője az általános (3.1)-(3.2) egyenletpár. Míg a differenciálegyenletnek m paramétere van, a differenciaegyenlet a lépésköznek köszönhetően $m + 1$ szabad paraméterrel rendelkezik.

III.1. Pontos véges differrencia sémák

Fontos megjegyezni, hogy az alfejezet annyiban rendhagyó, hogy bizonyos KDE-k analitikus megoldását fel fogja használni az őket megoldó véges differencia sémák szerkesztése során. Ennek célja az, hogy az így nyert sémák struktúrájából olyan tanulságokra tehessünk szert, melyeket a későbbiekben valódi alkalmazások során fel tudunk használni.

[3] A fogalom jobban kifejthető [19], viszont a jelen cél szempontjából a fenti "definíció" kielégítő.

3.2. Definíció *A* (3.1) *és* (3.2) *alakú egyenletekről azt mondjuk, hogy ugyanazzal az* általános megoldással *rendelkeznek, amennyiben minden* $h > 0$ *lépésközre*

$$\forall t_n \in \omega_h : \; u(t_n) = y_n$$

teljesül, ahol ω_h *a* h *lépésközű ekvidisztáns rácsháló a kitűzött időintervallumon.*

3.3. Definíció *Egy differencia sémát* pontosnak *nevezünk, amennyiben ugyanaz az általános megoldása, mint az általa modellezett differenciálegyenletnek.*

A pontosság definíciójából következik, hogy ilyen sémák numerikus instabilitás mentesek, mivel minden megoldásuk egyértelműen megfeleltethető a hozzátartozó DE megoldásának. Az analitikus megoldások ismeretében Mickens technikáját [19, 3.2] fogjuk alkalmazni. Ehhez tekintsük a

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = \boldsymbol{f}(\boldsymbol{u}, t, \lambda), \quad \boldsymbol{u}(t_0) = \boldsymbol{u}_0$$

m dimenziós CF-t. Ekkor a hozzátartozó pontos differencia sémát az

$$\begin{cases} \boldsymbol{u}(t) \longrightarrow \boldsymbol{y}_{n+1} \\ \boldsymbol{u}(t_0) \longrightarrow \boldsymbol{y}_n \\ t \longrightarrow t_{n+1} \\ t_0 \longrightarrow t_n \end{cases} \tag{3.6}$$

helyettesítésekkel kaphatjuk meg.

3.4. Példa *Elsőnek tekintsük a már jól ismert* (3.1) *tesztfeladatot, melynek megoldása*

$$u(t) = u_0 e^{-\lambda(t-t_0)}.$$

A (3.6) helyettesítést követve kapjuk az

$$y_{n+1} = y_n e^{-\lambda h}.$$

differencia sémát, melyet átrendezve az

$$y_{n+1} - y_n = (e^{-\lambda h} - 1) y_n = -\lambda \left(\frac{1 - e^{-\lambda h}}{\lambda} \right) y_n,$$

egyenletet kapjuk, amit végül

$$\frac{y_{n+1} - y_n}{\left(\frac{1-e^{-\lambda h}}{\lambda} \right)} = -\lambda y_n$$

alakba hozhatunk. Így a pontos differencia séma a bal oldali nevező kivételével nagyban hasonlít az (2.3) standard diszkretizációra.

3.5. Példa *Második példának vegyük az*

$$u'(t) = \lambda_1 u(t) - \lambda_2 (u(t))^2, \quad u(t_0) = u_0$$

általános logisztikus differenciálegyenletet, ahol a CF analitikus megoldása

$$u(t) = \frac{\lambda_1 u_0}{(\lambda_1 - u_0\lambda_2)e^{-\lambda_1(t-t_0)} + \lambda_2 u_0}. \tag{3.7}$$

A (3.6) behelyettesítést a (3.7) egyenletre alkalmazva az

$$y_{n+1} = \frac{\lambda_1 y_n}{(\lambda_1 - \lambda_2 y_n)e^{-\lambda_1 h} + \lambda_2 y_n} \tag{3.8}$$

differencia sémát kapjuk. Az előzőhöz hasonló átrendezések után (3.8) az alábbi

$$\frac{y_{k+1} - y_k}{(e^{\lambda_1 h} - 1)} = \lambda_1 y_k - \lambda_2 y_{k-1} y_k$$

alakra hozható, ami szintén a haladó differencia séma diszkretizációra emlékeztető.

3.6. Példa *Végül tekintsük az*

$$u''(t) + \omega^2 u(t) = 0, \quad u(t_0) = u_0, \quad u'(t_0) = u_0'$$

másodrendű DE-t, a harmonikus oszcillátort, melyet felírhatunk

$$\begin{aligned} u'(t) &= v(t), & u(t_0) &= u_0 \\ v'(t) &= -\omega^2 u(t), & v(t_0) &= v_0 \end{aligned} \tag{3.9}$$

alakban, így egy lineáris DE-rendszer, melynek megoldása pontosan megadható.

A (3.9) rendszer megoldásából helyettesítés után az

$$\begin{aligned} y_{n+1} &= \cos(\omega h) y_n + \sin(\omega h) z_n \\ z_{n+1} &= \sin(\omega h) y_n + \cos(\omega h) z_n \end{aligned}$$

differencia egyenletrendszert kapjuk, ahol z_n kiejthető, így kapjuk az

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\frac{4}{\omega^2}\sin^2\left(\frac{\omega h}{2}\right)} + \omega^2 y_n = 0$$

másodrendű differenciaegyenletet. Figyeljük meg, hogy az így kapott formula a másodrendű centrális differencia sémával szerkeszthető diszkretizációra hasonlít a nevezőbeli eltérést figyelmen kívül hagyva.

III.2. Az NSFD metodika első eszközei

Először is gyűjtsük össze az előző fejezet pontos differrencia sémáiban felfedezhető szabályszerűségeket.

- A pontos differencia sémában szereplő derivált közelítés mindig hasonlít egy, a DE rendjével megegyező véges differencia közelítésre.
- Az említett véges differencia közelítésekben a nevezőben szereplő függvény összetettebb, mint a standard közelítés nevezője.
- A 3.5. Példában szereplő nemlineáris u^2 tag a (3.7) pontos differencia sémában *nem lokálisan*, azaz nem egy rácspontbeli érték függvényeként szerepelt, vagyis

$$u^2(t) \longrightarrow y_{n+1}y_n.$$

Vegyük észre, hogy a határérték

$$\lim_{\substack{h\to 0\\ n\to\infty\\ nh=t \text{ fix}}} y_{n+1}y_n = \lim_{\substack{h\to 0\\ n\to\infty\\ nh=t \text{ fix}}} y_n^2 = u^2(t),$$

de véges értékekre a nemlineáris tag két reprezentációja eltérő értéket jelenthet. Továbbá azt is megfigyelhetjük, hogy a második pontban szereplő NSFD derivált közelítés egy (3.1) alakú DE esetében megfogalmazható

$$\frac{\mathrm{d}u}{\mathrm{d}t} \longrightarrow \frac{y_{n+1}-y_n}{\phi(h,\lambda)}$$

formában is. Így láthatóvá válik, hogy a standard derivált közelítésének egy általánosítását kaptuk[4]. Belátható, hogy az általánosított derivált közelítés rögzített λ értékre $h \to 0$ határértékben meg is egyezik a valódi derivált értékkel, amennyiben a ϕ *nevezőfüggvény*

$$\phi(h,\lambda) = h + \mathcal{O}(h^2), \quad \text{második derivált esetén} \quad h^2 + \mathcal{O}(h^4) \tag{3.10}$$

alakú. A fenti példákban éppen ilyen tulajdonságú nevezőfüggvényeket kaptunk, további lehetséges választások első derivált esetén:

$$\phi(h,\lambda) \in \left\{ h,\ \sin(h),\ e^h-1,\ 1-e^{-h},\ \frac{1-e^{-\lambda h}}{\lambda},\ \dots \right\}.$$

A következő részben megvizsgáljuk, hogy hogyan alkalmazhatóak a megfigyeléseink.

[4] A $\phi(h,\lambda) = h$ választás éppen a véges haladó differencia közelítést eredményezi.

3.7. Példa *Egy kiemelkedő szerepű egyenlet a nemlineáris oszcillációs jelenségek vizsgálatában a Duffing egyenlet [28]. Ennek az*

$$u''(t) + \omega^2 u(t) + \lambda u^3(t) = 0 \tag{3.11}$$

speciális esetét fogjuk megvizsgálni [20].

Az egyenlet első integrálja

$$\frac{(u''(t))^2}{2} + \frac{\omega^2 u^2(t)}{2} + \frac{\lambda u^4(t)}{4} = E,$$

ahol E az energia, ami időben konstans, amennyiben λ nemnegatív. Mikor az egyenletről ismert, hogy megtartja az energiát, érdemes olyan módszerrel megoldani, ami szintén hasonló tulajdonsággal rendelkezik. Először is tekintsük (3.11)

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} + \omega^2 y_n + \lambda y_n^3 = 0 \tag{3.12}$$

egyik lehetséges standard differencia modelljét. Mindkét oldalt szorozva az

$$(y_{n+1} - y_n) + (y_n - y_{n-1}) = y_{n+1} - y_{n-1}$$

kifejezéssel, majd átrendezve megkapjuk az

$$\left(\frac{y_{n+1} - y_n}{h}\right)^2 + \omega^2 y_{n+1} y_n + \lambda y_{n+1} y_n^3 = \left(\frac{y_n - y_{n-1}}{h}\right)^2 + \omega^2 y_n y_{n-1} + \lambda y_{n-1} y_n^3 \tag{3.13}$$

egyenletet. Hogyha (3.13) megőrzi az energiát, a bal oldali $n \longrightarrow (n+1)$ csere a két oldal megegyezését kell, hogy eredményezze. Viszont az utolsó tagra ez nem teljesül, így a (3.12) differenciaegyenlet nem őrzi meg az energiát. Most tehát alkalmazzuk az NSFD modellezési megfigyeléseinket. A diszkrét derivált rendje megegyezik a DE rendjével, így módosítsuk a nevezőfüggvényt és a nemlineáris tagot az alábbi

$$h^2 \longrightarrow \phi(h, \omega, \lambda), \quad y_n^3 \longrightarrow y_n^2 \left(\frac{y_{n+1} + y_{n-1}}{2}\right)$$

módon úgy, hogy ϕ teljesítse a (3.10) tulajdonságot. Így a fenti számolást elvégezve az

$$\frac{(y_{n+1} - y_n)^2}{2\phi} + \frac{\omega^2 y_{n+1} y_n}{2} + \frac{\lambda y_{n+1}^2 y_n^2}{4} = \frac{(y_n - y_{n-1})^2}{2\phi} + \frac{\omega^2 y_n y_{n-1}}{2} + \frac{\lambda y_{n-1}^2 y_n^2}{4}$$

egyenletet kapjuk, mely jobb oldala az $n \longrightarrow (n+1)$ csere következtében a bal oldallal lesz egyenlő a ϕ függvény pontos alakjától függetlenül. Ez azt jelenti, hogy n-től függetlenül mindig teljesülni fog az eredeti egyenlet energiamegőrző tulajdonsága.

III.2.1. Autonóm egyenletek sémái

Ebben a fejezetben bemutatásra kerül egy speciális nevezőfüggvényeken alapuló eljárás, aminek segítségével olyan NSFD sémát leszünk képesek szerkeszteni, mely pontosan visszaadja egy skalár DE egyensúlyi pontjainak lineáris stabilitási tulajdonságait. Ehhez tekintsük az

$$u'(t) = f(u(t)) \tag{3.14}$$

autonóm DE azon esetét, amikor f olyan, hogy az

$$f(u) = 0$$

egyenletnek csupán egyszeres gyökei vannak. Jelölje

$$x_1, x_2, \ldots, x_M$$

a (3.14) egyenlet egyensúlyi pontjait. Továbbá legyen

$$R_i := f'(x_i), \quad i = 1, 2, \ldots, M \quad \text{és} \quad R^* := \max\{|R_i| : i = 1, 2, \ldots, M\}.$$

Ekkor az $u(t) = x_i$ egyensúlyi pont $R_i > 0$ esetben instabil, $R_i < 0$ esetben pedig stabil. Most tekintsük az

$$\frac{y_{n+1} - y_n}{\left[\frac{\phi(hR^*)}{R^*}\right]} = f(y_n) \tag{3.15}$$

NSFD sémát, ahol ϕ az alábbi

$$\begin{aligned} z \to 0: &\ \ \phi(z) = h + \mathcal{O}(h^2), \\ z > 0: &\ \ 0 < \phi(z) < 1 \end{aligned} \tag{3.16}$$

tulajdonságokkal rendelkezik.

3.8. Tétel *A (3.15) véges differencia séma egyensúlyi pontjai pontosan azzal a lineáris stabilitási tulajdonsággal rendelkeznek, mint a hozzátartozó*

$$u'(t) = f(u(t))$$

DE egyensúlyi pontjai minden $h > 0$-ra.

Bizonyítás Az i-edik egyensúlyi pont körüli perturbációt felírhatjuk

$$y_n = x_i + \varepsilon_k$$

formában. Majd x_i körüli linearizálás után az

$$\frac{\varepsilon_{n+1}-\varepsilon_n}{\left[\frac{\phi(hR^*)}{R^*}\right]} = R_i\varepsilon_n$$

hiba egyenletet kapjuk ε_n-re, melyet átrendezve az

$$\varepsilon_{n+1} = \left[1+\left(\frac{R_i}{R^*}\right)\phi(hR^*)\right]\varepsilon_n = \varepsilon_0\left[1+\left(\frac{R_i}{R^*}\right)\phi(hR^*)\right]^n$$

képletet kapjuk az egyensúlyi ponttól vett eltérésre. Most ϕ (3.16) tulajdonságait kihasználva leellenőrizhetjük, hogy amennyiben

- $R_i > 0: \quad 1+\left(\frac{R_i}{R^*}\right)\phi(hR^*) > 1, \quad h>0, \quad$ azaz $y_n \equiv x_i$ instabil;
- $R_i < 0: \quad 0 < 1-\left(\frac{|R_i|}{R^*}\right)\phi(hR^*) < 1, \quad h>0, \quad$ azaz $y_n \equiv x_i$ stabil.

Vagyis a (3.15) véges differencia séma egyensúlyi pontjainak lineáris stabilitása valóban megegyezik a tételben szereplő DE egyensúlyi pontjainak lineáris stabilitásával. □

3.9. Példa *Vegyük a talán legegyszerűbb három egyensúlyi ponttal rendelkező 3.8. Tétel feltételeinek megfelelő*

$$u'(t) = u(t)(1-u^2(t))$$

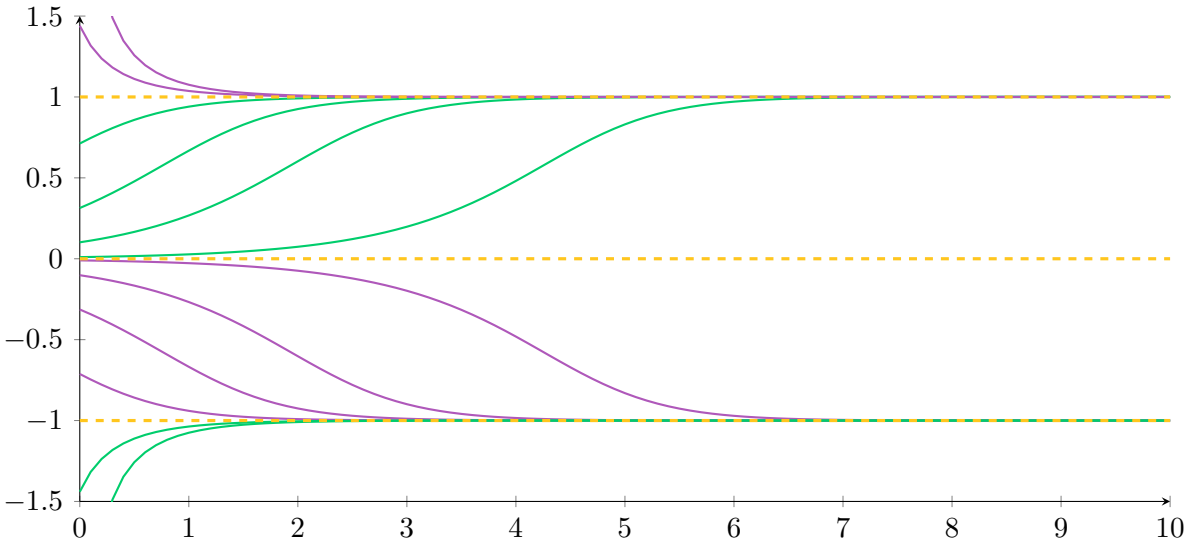
autonóm differenciálegyenletet. Ebben az esetben

$$f(u) = u(1-u^2), \quad x_1 = 0, \quad x_2 = 1, \quad x_3 = -1,$$
$$R_1 = 1, \quad R_2 = R_3 = -2, \quad R^* = 2.$$

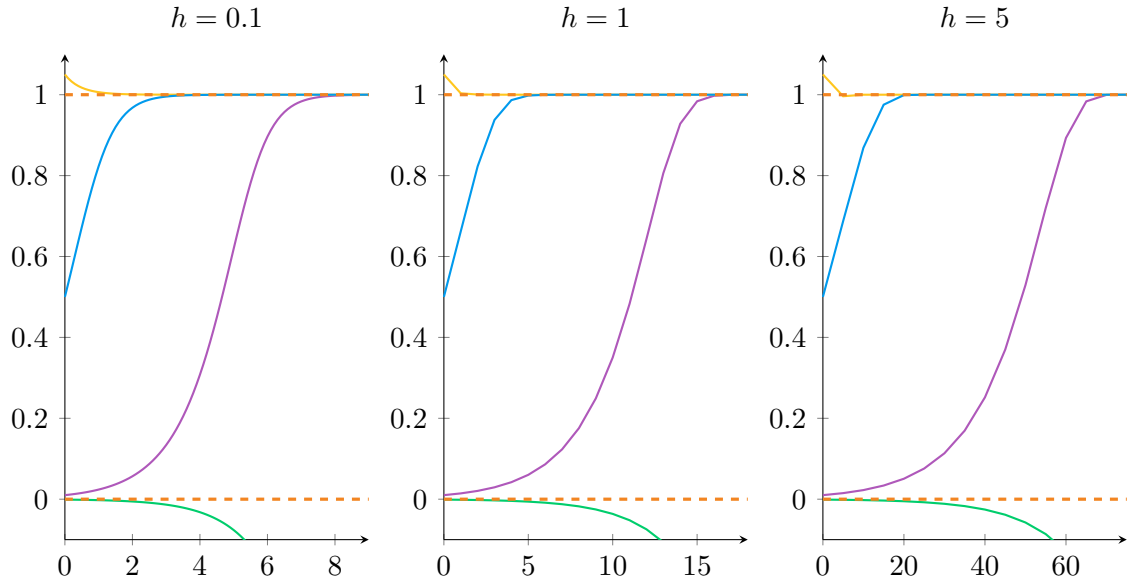
A példa megoldására alkalmazzuk a (3.15) sémát $\phi(h) = 1-e^{-h}$ választással. Ekkor

$$\frac{y_{n+1}-y_n}{\left(\frac{1-e^{-2h}}{2}\right)} = y_n(1-y_n^2) \tag{3.17}$$

a kapott NSFD séma. A numerikus eredményeket a 3.2. ábrán láthatjuk, melyeket összevethetünk a 3.1. ábán látható pontos megoldásokkal. Érdemes megjegyezni, hogy csupán a nevezőfüggvény tér el egy standard diszkretizációtól, mégis tetszőleges lépésköz mellett megtartja az egyensúlyi pontok stabilitási tulajdonságát.



3.1. ábra. A (3.14) egyenlet tipikus megoldásai különböző kezdeti értékekkel.



3.2. ábra. A (3.17) NSFD-ből származó numerikus megoldások különböző lépésközök mellett.

Hogyha a

$$h \longrightarrow \frac{1 - e^{-R^* h}}{R^*}$$

nevezőfüggvény normalizálás mellett nemlokális módon reprezentáljuk a nemlineáris tagot,

további javulást tudunk elérni. Egy maximális szimmetriával rendelkező lehetséges eljárás:

$$u(1-u^2) = u(1+u)(1-u)$$
$$\downarrow$$
$$\frac{y_{n+1}(1+y_n)(1-y_n)}{2} + \frac{y_n(1+y_{n+1})(1-y_n)}{2}$$
$$+\frac{y_n(1+y_{n+1})(1-y_{n+1})}{2} - \frac{y_n(1+y_n)(1-y_n)}{2}.$$

Ezzel a módosítással a (3.17) differencia séma megoldása nemcsak lokálisan adja vissza az egyensúlyi pontok stabilitását, hanem pontosan az eredeti egyenletnek megfelelő pontokhoz fog tartani tetszőleges lépésköz mellett [18].

III.3. NSFD modellezési eljárás

Mickens azon diszkrét modelleket *legjobb véges differencia sémának* nevezi, melyek az alábbi szabályok szerint kerültek megalkotásra:

1. A diszkrét deriváltak rendje egyezzen meg a DE-ben szereplő derivált rendjével.
2. A diszkrét deriváltban szereplő nevezőfüggvény általánosságban összetettebb függvénye legyen a lépésköznek, mint standard modellek esetén.
3. A nemlineáris tagok általában legyenek nemlokális módon reprezentálva a rácshálón.
4. A DE speciális megoldásainak egyben a véges differencia séma speciális (diszkrét) megoldásának is kell lenniük.
5. A véges differencia sémának ne legyen olyan megoldása, mely nem felel meg a DE egyetlen lehetséges megoldásának sem.

Az öt szabály alkalmazása nem minden esetben eredményez egy egyértelmű, vagy pontos differencia sémát, viszont az általuk nyert legjobb véges differencia sémák számos kívánatos tulajdonsággal rendelkezhetnek.

A fejezet zárásaként érdemesnek tartom megemlíteni, hogy a bemutatott általános eljárások, ötletek és szabályok kiterjeszthetők parciális differenciálegyenletek kezelésére is. Érdeklődő Olvasók figyelmébe ajánlom az említett irodalom [19] 5-11. fejezetét, melyben a dolgozatomban bemutatott alapok továbbfejlesztéseképpen számos összegyűjtött példát és alkalmazást találhatunk.

IV. Egy általános NSFD séma-család merev feladatokra

"It is obvious that explicit methods for stiff problems of ordinary differential equations have to be nonlinear in some sense."
(Olavi Nevanlinna és Aarne H. Sipila [21])

Dolgozatom utolsó fejezetében az egyik legújabb eredményeket tartalmazó, merev feladatokat nemstandard módon kezelő cikket [15] fogom feldolgozni.

Mint azt a II.3.1. fejezetben is tapasztalhattuk, merev feladatok megbízható megoldását A- és L-stabilitáshoz hasonló, kivételes stabilitási tulajdonsággal rendelkező módszerekkel tudjuk biztosítani. A 2.11. Tételben már láttuk, hogy az explicit RK típusú módszerek nem lehetnek A-stabilak, így L-stabilak sem. Sajnos explicit esetben ez az eredmény egy sokkal általánosabb "lineáris" módszer osztályra is kiterjeszthető, mely többek között tartalmazza a *lineáris többlépéses* és *prediktor-korrektor (avagy "jósló-javító")* módszereket [21]. Implicit módszerek ugyan lehetnek A- és L-stabilak, ugyanakkor a nemlineáris egyenletrendszer megoldása minden egyes lépés során komoly hátrányt jelent ezen módszerek alkalmazásakor. Így több alternatív ötlet is született explicit megoldásokra, melyekben a pontos megoldás valamilyen nemlineáris függvénnyel való közelítése szerepel.

IV.1. Másodrend és A-stabilitás

Niekerk 1987-es inverz-polinomszerű ötlete [33] alapján Ramos 2007-ben az alábbi

$$u(t_n + h) \approx u(t_n) + \frac{hu'(t_n)}{\psi(h) + u(t_n)}$$

alakú explicit közelítést javasolta a (2.1) CF pontos megoldására, ahol $\psi(h)$ egy megfelelően sima, még nem ismert függvény [25]. Taylor-sorbafejtés segítségével ψ meghatározható oly módon, hogy a kapott

$$y_{n+1} = y_n + \frac{2hf_n^2}{2f_n - hf_n'} \tag{4.1}$$

explicit séma másodrendben legyen konvergens, ahol

$$f_n := f(t_n, y_n), \quad f'_n := \partial_1 f(t_n, y_n) + \partial_2 f(t_n, y_n) \cdot f_n.$$

A (2.10) Dahlquist-féle tesztegyenlet segítségével az

$$R(z) = \frac{2+z}{2-z}$$

stabilitási függvény is meghatározható, melyet megvizsgálva kiderül, hogy a numerikus séma A-stabil, de nem L-stabil. Ugyan nem az NSFD metodika szabályai szerint építettük fel a sémát, de nemlineáris jellegéből látható, hogy mennyire nemstandard lépéseket kaptunk. Érdemes megemlíteni, hogy a módszer explicit, de alkalmazásához (mint implicit módszerek esetén a Newton-iterációhoz) szükséges az f jobb oldali függvény parciális deriváltjainak ismerete.

IV.2. Az L-stabil séma-család

A kiindulópont Ramos (4.1) explicit sémája lesz. Ezt a módszert fogjuk általánosítani két extra függvény hozzáadásával, melyek a szükséges szabadsági fokot fogják nyújtani. A bal oldali deriváltnak vegyük a standard haladó

$$u'(t) \approx \frac{y_{n+1} - y_n}{h}$$

diszkretizációját. Majd a jobb oldalt írjuk fel az alábbi

$$\begin{aligned} f(t_n, u(t_n)) &= f(t_n, u(t_n)) + [-u(t_n) + u(t_n)]\, A(t_n, u(t_n)) \\ &= f(t_n, u(t_n)) + \big[-u(t_n)A(t_n, u(t_n)) + u(t_n)A(t_n, u(t_n))\big] \end{aligned}$$

formában. Ezután az NSFD metodika szerint modellezük a 0-függvényt nemlokálisan:

$$0 = -y_n + y_n \longrightarrow -y_n + y_{n+1}.$$

Ezt a 0 közelítést, két $\alpha, \beta \in \mathbb{R}$ paramétert és egy

$$B(t, y) : \ \mathbb{R}^2 \to \mathbb{R}, \quad 0 \approx \beta h B(t_n, u(t_n)) \tag{4.2}$$

függvényt bevezetve felírhatjuk a jobb oldal

$$f(t_n, y_n) \approx f(t_n, y_n) - \alpha y_n A(t_n, y_n) + \alpha y_{n+1} A(t_n, y_n) + \beta h B(t_n, y_n)$$

nemstandard közelítését, amivel dolgozni fogunk. Ekkor a véges differencia sémát az

$$\frac{y_{n+1}-y_n}{h} = f(t_n,y_n) - \alpha y_n A(t_n,y_n) + \alpha y_{n+1} A(t_n,y_n) + \beta h B(t_n,y_n) \tag{4.3}$$

alakban kereshetjük. Kihasználva a (4.2) tulajdonságot, átrendezés segítségével az alábbi

$$y_{n+1} = \frac{y_n + hf_n - h\alpha y_n A_n}{1 - h\alpha A_n - h^2\beta B_n} = y_n + \frac{hf_n + h^2\beta y_n B_n}{1 - h\alpha A_n - h^2\beta B_n} \tag{4.4}$$

explicit módon adhatjuk meg a (4.3) sémát, ahol

$$f_n := f(t_n,y_n), \quad A_n := A(t_n,y_n), \quad B_n := B(t_n,y_n).$$

Vegyük észre, hogy megfelelően kis h-ra

$$|1 - h\alpha A_n - h^2\beta B_n| > 0,$$

azaz (4.4) értelmes.

4.1. Tétel *A (4.4) egylépéses numerikus eljárás másodrendben konzisztens, azaz a lokális approximációs hibája $\mathcal{O}(h^3)$, pontosan akkor, ha*

$$2\beta y_n B_n + 2\alpha A_n f_n = f'_n, \quad \textit{ahol } f'_n := \partial_1 f(t_n,y_n) + \partial_2 f(t_n,y_n)\cdot f_n. \tag{4.5}$$

4.2. Bizonyítás Először is írjuk fel a pontos megoldás

$$u(t_n+h) = u(t_n) + hf(t_n,u(t_n)) + \frac{h^2}{2} f'(t_n,u(t_n)) + \mathcal{O}(h^3) \tag{4.6}$$

Taylor-sorát. Majd a könnyebb átláthatóság kedvéért vezessük be az alábbi

$$f_D(t,y,h) = y + \frac{hf(t,y) + h^2\beta y B(t,y)}{1 - h\alpha A(t,y) - h^2\beta B(t,y)}$$

egyszerűsítő jelölést (4.4) jobb oldalára. Ekkor

$$f_D(t,y,0) = y, \quad \partial_3 f_D(t,y,0) = f(t,y),$$
$$\partial_3^2 f_D(t,y,0) = 2\alpha A(t,y) f(t,y) + 2\beta y B(t,y),$$

így a numerikus megoldás Taylor-sora:

$$y_{n+1} = y_n + hf(t_n,y_n) + \frac{h^2}{2}\Big[2\alpha A(t_n,y_n) f(t_n,y_n) + 2\beta y_n B(t_n,y_n)\Big] + \mathcal{O}(h^3). \tag{4.7}$$

Tehát (4.6) és (4.7) alapján $y_{n+1} - u(t_{n+1}) = \mathcal{O}(h^3)$ pontosan akkor, ha (4.5) teljesül. □

Számolással az is belátható, hogy a (2.4) egylépéses módszerhez tartozó Φ függvény lokálisan Lipschitzes a megfelelő változóban, vagyis a 2.7. Tétel szerint konvergens is másodrendben.

Vegyük észre, hogy amennyiben $\beta = 0$, a (4.5) feltétel csupán

$$A_n = \frac{f'_n}{2\alpha f_n}, \ \alpha f_n \neq 0,$$

ami éppen a (4.1) sémát eredményezi, így nem lehet L-stabil, tehát foglalkozzunk a $\beta \neq 0$ esettel. Ha ezentúl feltesszük azt is, hogy $\alpha \neq 0$, (4.5) következtében

$$B_n = \frac{f'_n - 2\alpha A_n f_n}{2\beta y_n}, \ y_n \neq 0.$$

Most az $A(t, y)$ függvényt kell ügyesen megválasztani. Tekintsük az

$$A(t, y) = \partial_2 f(t, y)$$

választást. Ekkor a (4.4) sémát ismeretlen függvények nélkül, teljesen explicit alakban

$$y_{n+1} = \frac{2y_n^2 + 2hy_n f_n - 2h\alpha y_n^2 f_{y,n}}{2y_n - 2h\alpha y_n f_{y,n} - h^2 f'_n + 2h^2 \alpha f_{y,n} f_n}, \quad f_{y,n} := A(t_n, y_n) = \partial_2 f(t_n, y_n) \tag{4.8}$$

módon is felírhatjuk.

4.3. Tétel *A (4.8) által meghatározott nemlineáris, másodrendű véges differencia séma*

- $\alpha \geq 1/2$ *értékekre A-stabil;*
- $\alpha > 1/2$ *értékekre L-stabil.*

4.4. Bizonyítás Alkalmazzuk a sémát a (2.10) tesztegyenletre, ekkor

$$f_n = \lambda y_n, \quad f'_n = 0 + \lambda^2 y_n, \quad f_{y,n} = \lambda, \quad z = h\lambda,$$

majd ezeket felhasználva

$$y_{n+1} = \frac{2y_n^2 + 2h\lambda y_n^2 - 2h\alpha y_n^2 \lambda}{2y_n - 2h\alpha y_n \lambda - h^2\lambda^2 y_n + 2h^2\lambda^2 \alpha y_n} = y_n \frac{2 + 2z - 2\alpha z}{2 - 2\alpha z + 2\alpha z^2 - z^2}.$$

Vagyis az egylépéses eljárás stabilitási függvénye

$$R(z) = \frac{2 + (2 - 2\alpha)z}{2 - 2\alpha z + (2\alpha - 1)z^2}.$$

Egy-két lépésnyi számolással belátható, hogy $|R(z)| \leq 1$ pontosan akkor, ha

$$\begin{aligned}(16\alpha - 8)(\mathrm{Re}(\lambda))^2 + (2\alpha - 1)^2\left[(\mathrm{Re}(\lambda))^4 + (\mathrm{Im}(\lambda))^4\right] - 4\alpha(2\alpha - 1)(\mathrm{Re}(\lambda))^3 \\ +8\,\mathrm{Re}(\lambda) + (4\alpha - 8\alpha^2)\,\mathrm{Re}(\lambda)(\mathrm{Im}(\lambda))^2 + 2(2\alpha - 1)^2(\mathrm{Re}(\lambda))^2(\mathrm{Re}(\lambda))^2 \geq 0.\end{aligned} \tag{4.9}$$

Mivel $\alpha \geq 1/2$, ezért a következő

$$16\alpha - 8 \geq 0, \quad 4\alpha(2\alpha - 1) \geq 0 \quad \text{és} \quad (4\alpha - 8\alpha^2) \leq 0$$

egyenlőtlenségek állnak fent, ami azt jelenti, hogy $\mathrm{Re}(\lambda) \leq 0$ esetben (4.9) teljesül, azaz $|R(z)| \leq 1$, vagyis a séma valóban A-stabil $\alpha \geq 1/2$ paraméterrel. Ezenkívül a stabilitási függvényről könnyen látható, hogy

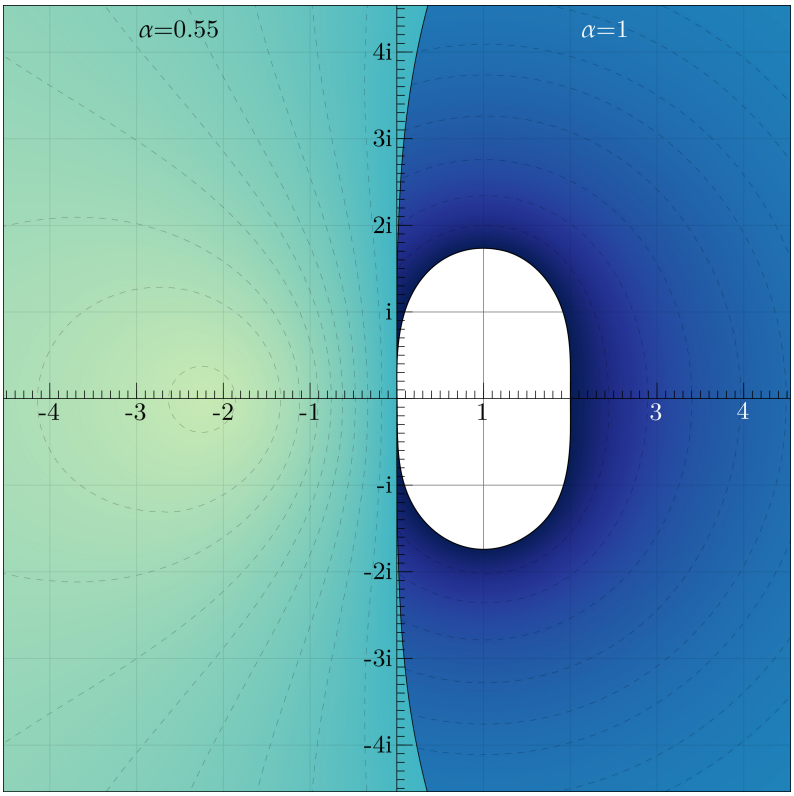
$$\lim_{z \to -\infty} R(z) = \begin{cases} -1, & \alpha = 1/2, \\ 0, & \alpha \neq 1/2. \end{cases}$$

Vagyis a (4.8) numerikus eljárás $\alpha > 1/2$ paraméterrel L-stabil. □

Tehát egy olyan másodrendű α valós számmal paraméterezett egylépéses véges differencia séma-családot sikerült szerkesztenünk, amely amellett, hogy explicit, még L-stabilitási tulajdonsággal is rendelkezik. Ezen remek tulajdonságok jelentőségét a következő alfejezetben fogom bemutatni, de ezelőtt még vegyük számba milyen *áron* értük el ezt eredményt (avagy mi a *"trade-off"*):

- A séma bal oldalán szereplő számláló minden tagjában szerepel az y_n érték, vagyis 0-ból indítva mindig végig 0 eredményt kapunk függetlenül a feladattól.
- Minden egyes lépéshez szükséges f parciális deriváltjainak ismerete. Mivel ez nincs minden esetben megadva, numerikusan kell közelítünk, ami pedig a számítási költség növelését jelenti.

A második pont nehézsége minden valamilyen Newton-iterációt alkalmazó módszernél jelen van, még Ramos (4.1) explicit sémájában is. Az első pontbeli problémát az értékek kis perturbálásával próbálhatjuk meg kezelni. Emellett érdemes azt is megjegyezni, hogy az α paraméter függvényében nagyban változik a séma stabilitási tartománya (lásd 4.1 ábra). Így az, hogy egy adott feladatot milyen α választással oldunk meg, nem egy triviális kérdés.



4.1. ábra. A (4.8) differencia séma stabilitási tartománya különböző α paraméterekre.

IV.3. Numerikus eredmények

A most következő példákban az új séma-család teljesítményét fogom összehasonlítani a kiindulópontnak tekintett (4.1) NSFD sémával és implicit módszerekkel különböző merev feladatokon. Az egyszerűbb hivatkozás kedvéért vezessük be a másodrendű L-stabil explicit nemlineáris módszert leíró (4.8) sémára az LENM2 jelölést, míg Ramos (4.1) másodrendű A-stabil explicit nemlineáris módszert adó sémájára az AENM2 jelölést.

4.5. Példa *Tekintsük az alábbi*

$$u'(t) = u^2(t) - e^{-2000t} - 1002e^{-1000t} - 1, \quad u(0) = 2$$

merev feladatot, melynek pontos megoldása $u(t) = e^{-1000t} + 1$*, azaz egy gyors tranziens szakasz található a megoldás első szakaszában [15].*

A CF megoldására alkalmazzuk most a két másodrendű NSFD sémát különböző lépésközök mellett. A 4.1. táblázat a végpontban vett közelítő megoldás és pontos megoldás eltérése ($e_{\text{vége}}$) mellett az intervallumon vett legnagyobb hibát ($e_{\max}$) is tartalmazza.

	LENM2 ($\alpha = 0.55$)		AENM2	
h	$e_{\max}$	$e_{\text{vége}}$	$e_{\max}$	$e_{\text{vége}}$
10^{-1}	0.96078	0.96078	0.96078	0.96078
10^{-2}	0.74705	0.74705	0.74747	0.74747
10^{-3}	3.4546e − 2	9.687e − 3	6.6065e − 2	6.6065e − 2
10^{-4}	2.3756e − 4	1.5504e − 4	9.6796e − 4	9.6796e − 4
10^{-5}	2.2889e − 6	1.6204e − 6	1.0117e − 5	1.0117e − 5
10^{-6}	2.2804e − 8	1.6276e − 8	1.0163e − 7	1.0163e − 7

4.1. táblázat. Hibák a 4.5. Példára a $[0, 0.1]$ intervallumon.

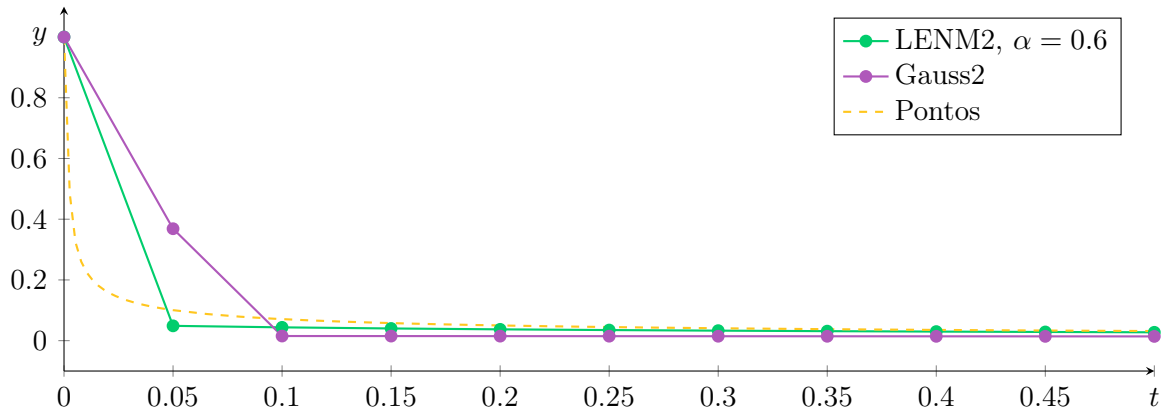
Ezekből az adatokból látható, hogy a két séma megfelelően kis h lépésköz mellett valóban másodrendben konvergál, viszont a LENM2 séma a nemlokális közelítésből származó általánosítás következtében jobban ellenáll a merevségnek, mint az AENM2 séma ezen a feladaton.

4.6. Példa *Végső példának vegyük az*

$$u'(t) = -999 \cdot u^3(t), \quad u(0) = 1, \quad t \in [0, 0.5]$$

feladatot, melynek pontos megoldása $u(t) = \dfrac{1}{\sqrt{1 - 2 \cdot 999t}}$.

Ezen feladat megoldásra alkalmazzuk a LENM2 sémát $\alpha = 0.6$ paraméterrel és egy A-stabil másodrendű RK módszert, a Gauss2 numerikus eljárást. Megfigyelve a 4.2. ábrát és 4.2. táblázatot láthatjuk, hogy ezen feladat megoldásában mennyire hatásosnak bizonyul az explicit séma stabilitási tulajdonsága. A LENM2 séma még akkor is jobb közelítést eredményezett, amikor az egyenlet időskálájához képest nagy lépésközt használtunk, így ebben az esetben sikerült áthidalnunk az explicit módszerek egyik jelentős nehézségét.



4.2. ábra. Numerikus megoldások a 4.6. Példára $h = 0.05$ lépésközzel.

	LENM2 ($\alpha = 0.6$)		Gauss2	
h	$e_{\max}$	$e_{\text{vége}}$	$e_{\max}$	$e_{\text{vége}}$
$5 \cdot 10^{-1}$	0.026334	0.026334	0.73083	0.73083
$5 \cdot 10^{-2}$	0.050757	4.0849e − 3	0.49298	3.497e − 2
$5 \cdot 10^{-3}$	0.015771	1.6778e − 5	0.18081	8.7419e − 4
$5 \cdot 10^{-4}$	1.7515e − 3	3.4669e − 7	1.167e − 2	2.0286e − 6
$5 \cdot 10^{-5}$	2.3075e − 5	3.9314e − 9	1.1597e − 4	1.9711e − 8

4.2. táblázat. Hibák a 4.6. Példára a $[0, 0.5]$ intervallumon.

Irodalomjegyzék

[1] Francis Bashforth és John Couch Adams. *An attempt to test the theories of capillary action by comparing the theoretical and measured forms of drops of fluid*. University Press, 1883.

[2] John C. Butcher. Implicit Runge-Kutta processes. *Mathematics of computation*, 18(85):50–64, 1964.

[3] John C. Butcher. On Runge-Kutta processes of high order. *Journal of the Australian Mathematical Society*, 4(2):179–194, 1964.

[4] John C. Butcher. Numerical methods for ordinary differential equations in the 20th century. *Journal of Computational and Applied Mathematics*, 125(1-2):1–29, 2000.

[5] A. L. Cauchy. Resumé des Leçons données à l'Ecole Royale Polytechnique. Suite du Calcul Infinitésimal. https://archive.org/embed/EquationsDifferentiellesOrdinaires, 1924. publikálva: Équations Differentielles Ordinaires (1981).

[6] John Crank. *The mathematics of diffusion*. Oxford university press, 1979.

[7] Charles Francis Curtiss és Joseph O. Hirschfelder. Integration of stiff equations. *Proceedings of the national academy of sciences*, 38(3):235–243, 1952.

[8] Germund G Dahlquist. A special stability problem for linear multistep methods. *BIT Numerical Mathematics*, 3(1):27–43, 1963.

[9] John R. Dormand és Peter J. Prince. A family of embedded Runge-Kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.

[10] Leonhard Euler. *Institutionum Calculi Integralis*, volume XI. 1768. Volumen Primum, Opera Omnia.

[11] Ernst Hairer, Christian Lubich és Gerhard Wanner. Geometric numerical integration. *Geometric Numerical Integration*, 2011.

[12] Ernst Hairer és Gerhard Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems.* Springer Berlin, Heidelberg, 1996.

[13] Ernst Hairer, Gerhard Wanner és Syvert P. Nørsett. *Solving Ordinary Differential Equations I: Nonstiff Problems.* Springer Berlin, Heidelberg, 1993.

[14] Preston C. Hammer és Jack W. Hollingsworth. Trapezoidal methods of approximating solutions of differential equations. *Mathematical Tables and Other Aids to Computation*, pages 92–96, 1955.

[15] Manh Tuan Hoang és Matthias Ehrhardt. A general class of second-order L-stable explicit numerical methods for stiff problems. *Applied Mathematics Letters*, 149:108897, 2024.

[16] Faragó István. *Numerikus modellezés és közönseges differenciálegyenletek numerikus megoldási módszerei.* 2013.

[17] David I. Ketcheson, Benjamin Seibold, David Shirokoff és Dong Zhou. DIRK schemes with high weak stage order. *Spectral and High Order Methods for Partial Differential Equations*, page 453, 2020.

[18] R. E. Mickens. *Difference Equations: Theory and Applications.* New York: Van Nostrand Reinhold, 1990.

[19] R. E. Mickens. *Nonstandard finite difference schemes: methodology and applications.* World Scientific, 2020.

[20] R. E. Mickens, O. Oyedeji és C. R. McIntyre. A difference-equation model of the duffing equation. *Journal of Sound Vibration*, 130(3):509–512, 1989.

[21] Olavi Nevanlinna és Aarne H. Sipilä. A Nonexistence Theorem for Explicit A-Stable Methods. *Mathematics of Computation*, 28(128):1053–1055, 1974.

[22] J. Oliver. A curiosity of low-order explicit Runge-Kutta methods. *Mathematics of Computation*, 29(132):1032–1036, 1975.

[23] Elaine S. Oran és Jay P. Boris. Numerical simulation of reactive flow. *NASA STI/-Recon Technical Report A*, 88:44860, 1987.

[24] A. Prothero és A. Robinson. On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. *Mathematics of Computation*, 28(125):145–162, 1974.

[25] Higinio Ramos. A non-standard explicit integration scheme for initial-value problems. *Applied Mathematics and Computation*, 189(1):710–718, 2007.

[26] Carl Runge. Über die numerische Auflösung von Differentialgleichungen. *Mathematische Annalen*, 46(2):167–178, 1895.

[27] Leonid Markovich Skvortsov. How to avoid accuracy and order reduction in Runge–Kutta methods as applied to stiff problems. *Computational Mathematics and Mathematical Physics*, 57:1124–1139, 2017.

[28] J. J. Stokes. Nonlinear vibrations. *Intersciences, New York*, 1950.

[29] Gustaf Söderlind. Automatic control and adaptive time-stepping. *Numerical Algorithms*, 31:281–310, 2002.

[30] Gustaf Söderlind. Digital filters in adaptive time-stepping. *ACM Transactions on Mathematical Software (TOMS)*, 29(1):1–26, 2003.

[31] Gustaf Söderlind, Laurent Jay és Manuel Calvo. Stiffness 1952–2012: Sixty years in search of a definition. *BIT Numerical Mathematics*, 55(2):531–558, 2015.

[32] Gustaf Söderlind és Lina Wang. Adaptive time-stepping and computational stability. *Journal of Computational and Applied Mathematics*, 185(2):225–243, 2006.

[33] F. D. Van Niekerk. Non-linear one-step methods for initial value problems. *Computers & Mathematics with Applications*, 13(4):367–371, 1987.

Appendix

A. Brusselator

A Brusselator egy autokatalitikus reakció elméleti modelljét leíró differenciálegyenlet, mely $a, b > 0$ rögzített paraméterekkel az

$$X'(t) = 1 - (b+1)X(t) + aX^2(t)Y(t)$$
$$Y'(t) = bX(t) - aX^2(t)Y(t)$$

általános alakban írható le. A dolgozatban csak az $a = 1,\ b = 3$ esetet tekintettük.

B. MATLAB kódok

A dolgozatom megírásához készített MATLAB® kódok mind megtalálhatók a https://github.com/mate-oro/BSc_Thesis_code.git GitHub oldalon, melyek segítségével a bemutatott numerikus eredmények reprodukálhatók.