

Implicit-explicit módszerek időfüggő differenciálegyenletek megoldására

Szakdolgozat

Készítette: Horlik Zalán Zoltán
ELTE TTK, Matematika alapszak

Témavezető: Dr. Fekete Imre
Alkalmazott Analízis és Számításmatematikai Tanszék



Budapest, 2025

Tartalomjegyzék

1. Bevezetés	3
2. Numerikus módszerek kezdetiérték-feladatok megoldására	4
2.1. Alapfogalmak	4
2.2. Alapvető egy lépéses módszerek	5
2.2.1. Az explicit Euler-módszer	5
2.2.2. Az implicit Euler-módszer	6
2.2.3. Numerikus módszerek hibái és konvergenciája	7
2.3. Runge-Kutta módszerek	7
2.3.1. Explicit Runge-Kutta módszerek	8
2.3.2. Runge-Kutta módszerek rendfeltételei	10
2.3.3. Implicit Runge-Kutta módszerek	11
2.4. Többlépéses módszerek	12
2.4.1. Adams-módszerek	13
2.4.2. BDF-módszerek	14
3. Numerikus módszerek stabilitása	15
3.1. A Dahlquesit-féle tesztfeladat	15
3.2. A-stabilitás és merevség	16
4. IMEX-módszerek	21
4.1. Többlépéses IMEX-módszerek	22
4.1.1. Stabilitásvizsgálat	23
4.2. IMEX Runge-Kutta módszerek	25
4.2.1. Stabilitás kérdése	28
4.2.2. Numerikus kísérlet	29
5. Összegzés	31

Köszönetnyilvánítás

Ezúton szeretném megköszönni az egész családomnak az évek során nyújtott kitartó támogatásukat. Hálás vagyok, hogy minden vizsga előtt szorítottak értem, biztattak a nehéz időszakokban, és mellettem álltak a mélypontokon is. Nélkülük ez az egyetemi út sokkal nehezebb lett volna.

Illetve, szeretném kifejezni hálámat Dr. Fekete Imrének, akinek fáradságos munkája nélkül ez a dolgozat nem tudott volna megvalósulni.

1. fejezet

Bevezetés

Közönséges és parciális differenciálegyenlet rendszerek időbeli vizsgálatára többféle módszer dolgoztak ki az elmúlt évtizedek során. Egyik fontos esete, ha a rendszer merev, azaz gyors és lassú komponensek is jelen vannak benne. A klasszikus módszerek –mint az explicit és implicit Runge–Kutta, vagy lineáris többlépéses módszerek– korlátai világosan megmutatkoznak ilyen esetekben: az explicit módszerek stabilitása rendkívül kis időlépést kíván, míg az implicit sémák stabilak ugyan, de számításmatematikai értelemben költségesek.

E kihívásokra válaszul alakultak ki olyan vegyes (hibrid) módszerek, amelyek ötvözik az explicit és implicit módszerek előnyeit. Ezek közé tartoznak az IMEX (Implicit-Explicit) sémák, melyek lényege, hogy a differenciálegyenlet jobboldalát két részre bontjuk: merev és nem merev részekre. A merev, rendszerint lineáris komponenst implicit módon, míg a nem merev, általában nemlineáris részt explicit módon kezeljük.

A dolgozat célja az IMEX módszerek alapötletének, elméleti hátterének és numerikus tulajdonságainak áttekintése, valamint azok alkalmazásának bemutatása konkrét feladaton keersztül.

A 2. és 3. fejezetek elméleti háttere főként Faragó István [5] jegyzetére épül. A 2. fejezetben megismerkedünk az a Cauchy-feladatok alapfogalmaival, valamint azok numerikus megoldási módszereivel. Részletesen tárgyaljuk a két legfontosabb módszer családot: a Runge–Kutta módszereket és a lineáris több lépéses módszereket ismertetve azok általános formuláját, pontossági feltételeit (rendfeltételek), valamint leginkább ismert verzióit.

A 3. fejezet középpontjában a numerikus stabilitás áll. Itt definiáljuk a stabilitás fogalmát, megvizsgáljuk a Dahlquist-féle tesztfeladatra a numerikus módszerekhez tartozó stabilitási függvényeket, valamint bemutatjuk a merevség jelenségét, amely kulcsfontosságú az IMEX módszerek motivációjának megértéséhez.

A 4. fejezetben az IMEX módszerek részletes tárgyalása következik. Először azt tekintjük át, hogy milyen típusú feladatok esetén előnyös az alkalmazásuk, és milyen szerkezeti tulajdonságokat használnak ki. Ezt követően Ascher, Ruuth és Spiteri publikációi ([1], [2], [3],[10]) alapján részletesen ismertetjük az IMEX lineáris többlépéses, valamint az IMEX Runge–Kutta módszereket. A fejezetben egy egyszerű advekción problémán keresztül vizsgáljuk a módszerek stabilitási viselkedését is. Legvégül pedig egy Fischer–Kolmogorov családba tartozó egyenleten demonstráljuk egy adott IMEX Runge–Kutta módszer előnyét a konvencionális Runge–Kutta módszerekkel szemben.

Az 5. fejeztben a szakdolgozat eredményeit és meglátásait összegezzük.

2. fejezet

Numerikus módszerek kezdetiérték-feladatok megoldására

A bevezetésben már említettük, hogy az IMEX módszereket parciális differenciálegyenletek numerikus megoldására alkalmazzák. Ezen egyenletek megfelelő térbeli diszkrétizáció után, például véges differenciákkal, időfüggő közönséges differenciálegyenletek rendszerének kezdetiérték-feladatára vezethetők vissza. Ezért még az IMEX módszerek vizsgálata előtt érdemes ezen rendszerek alapfogalmainak megismerésével kezdenünk.

2.1. Alapfogalmak

2.1.1. Definíció. Legyen $T \subset \mathbb{R} \times \mathbb{R}^d$ tartomány és $f : T \rightarrow \mathbb{R}^d$ folytonos függvény a $(t_0, u_0) \in T$ pontban. Ekkor a

$$\begin{cases} u'(t) = f(t, u(t)) \\ u(t_0) = u_0 \end{cases}$$

feladatot kezdetiérték-feladatnak, vagy Cauchy-feladatnak nevezzük.

Ennek megoldása az össze olyan $u : I \rightarrow \mathbb{R}^d$ meghatározása, amely valamely $I \subset \mathbb{R}$ intervallumon behelyettesíthető a feladatba, illetve kielégíti azt.

2.1.2. Definíció. Az olyan $u : I \rightarrow \mathbb{R}^d$ folytonosan differenciálható függvényt, amelyre

- $\{(t, u(t)) : t \in I\} \subset G$ és $t_0 \in I$,
- $u'(t) = f(t, u(t)) \quad \forall t \in I$,
- $u(t_0) = u_0$

a Cauchy-feladat megoldásának nevezzük.

Gyakorlati feladatok esetében többnyire fizikai, kémiai, biológiai, illetve közgazdasági problémákat leíró Cauchy-feladatok egyértelmű megoldását szeretnénk megtalálni. Ennek feltételei találhatóak meg az alábbi klasszikus egzisztencia és unicitás tételben, melynek bizonyítása például [12] monográfia hatodik fejezetében olvasható.

2.1.1. Tétel (Picard-Lindelöf). Legyen $D \subset T$ téglá és $(t_0, u_0) \in D$, valamint f a második vektor változójában Lipschitz folytonos D -n, azaz tetszőleges $(t, u_1), (t, u_2) \in D$ pontra létezik $L > 0$ állandó, hogy

$$\|f(t, u_1) - f(t, u_2)\| \leq \|u_1 - u_2\|.$$

Ekkor létezik olyan $K(t_0)$ környezet, hogy létezik egyetlen $u : \overline{K(t_0)} \rightarrow \mathbb{R}^d$ megoldása a kezdetiérték-feladatnak.

A Cauchy-feladatok esetében a t változó tipikusan az idő változását jelöli és így természetes, hogy a $t > t_0$ -ra az $u(t)$ értéket szeretnénk megkapni. Ezért a kezdeti t_0 időpillanatot vehetjük tetszés szerint 0-nak, ezáltal az I intervallumot pedig $[0, T]$ -nak. Ekkor a feladat a következő alakot ölti:

$$\begin{cases} u'(t) = f(t, u(t)), & t \in [0, T], \\ u(0) = u_0. \end{cases}$$

2.2. Alapvető egylépéses módszerek

A kezdetiérték-feladatok megoldását sok esetben megkaphatjuk direkt számításokkal is, azonban vannak esetek, amikor a megoldások nem explicitek, így általánosságban *numerikus megoldást* szoktunk keresni valamilyen közelítő eljárás segítségével. A következőkben megismerekedünk ezen módszerekkel és tulajdonságaikkal.

Az egyik nagy családja ezen numerikus módszereknek az ún. egylépéses módszerek. Az elnevezés onnan ered, hogy a közelítés kiszámításához közvetlen az előző időegységben kapott értéket használják fel.

A továbbiakban a szemléletesség és az egyszerűbb jelölésrendszer jegyében, az egydimenziós Cauchy-feladatot fogom vizsgálni.

$$\begin{cases} u'(t) = f(t, u(t)), & t \in [0, T], \\ u(0) = u_0. \end{cases} \quad (2.1)$$

2.2.1. Az explicit Euler-módszer

Vegyük a $[0, T]$ intervallumon a következő egyenlő lépésközű rácsot:

$$\omega_\tau = \{t_n = n\tau, n = 0, 1, \dots, N, \tau = T/N\}.$$

A dolgozat további részében ezen rácsháló pontjaiban szeretnénk kiértékelni a (2.1) feladatot kielégítő $u(t)$ megoldásfüggvény numerikus közelítését.

A numerikus formula készítéséhez integráljuk az $u'(t)$ függvényt a $[t_n, t_{n+1}]$ intervallumon:

$$\int_{t_n}^{t_{n+1}} u'(t) dt = \int_{t_n}^{t_{n+1}} f(t, u(t)) dt.$$

Az egyenlet baloldalát a Newton–Leibniz formula segítségével könnyen megkapjuk. A jobb oldalon lévő integrálnál azonban nem feltétlenül tudunk pontos értéket számolni, így ott valamilyen integrál közelítéssel kell élnünk. Ennek az egyik legegyszerűbb módja

a baloldali téglalap szabály, azaz a $\tau f(t_n, u(t_n))$ értékkel való approximáció. Ekkor a következő képletet kapjuk átrendezés után:

$$u(t_{n+1}) \approx u(t_n) + \tau f(t_n, u(t_n)).$$

Alapvető célunk, hogy minden ω_τ -beli t_n rácspontra igaz legyen, hogy a rácspontbeli $y(t_n)$ (tömören jelölve y_n) közelítő érték közel legyen a pontos megoldás rácspontbeli $u(t_n)$ értékéhez. A természetes $y_0 = u_0$ kezdetiérték választással adódik az $y_1 = y_0 + \tau f(t_0, y_0)$ numerikus közelítés. Az eljárást folytatva az

$$y_{n+1} = y_n + \tau f(t_n, y_n), \quad n = 0, 1, \dots, N-1 \quad (2.2)$$

ún. *explicit Euler-módszert* kapjuk.

A (2.2) explicit Euler-módszer a nevét onnan kapta, hogy az y_{n+1} kiszámításához a már ismert t_n és y_n értékekre van szükség. A módszer számítási költsége alacsony, hiszen a következő lépéshez lényegében csak egy az f függvényértéket kell kiszámolnia. Azonban vannak módszerek, amelyek függnek y_{n+1} -től és ekkor már nemlineáris f esetén egy nemlineáris algebrai egyenletet (általánosságban rendszert) kell megoldanunk a következő lépéshez. A következőkben ennek a legismertebb és legegyszerűbb esetével foglalkozunk.

2.2.2. Az implicit Euler-módszer

A (2.2) explicit Euler-módszer származtatásához az $\int_{t_n}^{t_{n+1}} f(t, u(t)) dt$ integrált a baloldali téglalap szabállyal közelítettük. Most azonban nézzük meg, hogy mi történik, ha a jobboldali téglalap szabállyal közelítünk:

$$u(t_{n+1}) \approx u(t_n) + \tau f(t_{n+1}, u(t_{n+1})).$$

Az explicit Euler-módszernél látott gondolatmenetet használva adódik az

$$y_{n+1} = y_n + \tau f(t_{n+1}, y_{n+1}), \quad n = 0, 1, \dots, N-1 \quad (2.3)$$

ún. *implicit Euler-módszer*.

Azért nevezzük a (2.3) módszert implicitnek, mert általánosan nem fejezhető ki belőle a y_{n+1} érték, így minden egyes iterációnál az

$$x - y_n - \tau f(t_{n+1}, x) = 0$$

nemlineáris egyenletet kell megoldanunk. Ebből kifolyólag a számítási költsége is jelentősen nagyobb az explicit módszerekhez képest.

Az explicit és implicit Euler-módszerekből merítve tekinthetünk egy általános

$$y_{n+1} = y_n + \tau \Phi(\tau, t_n, y_n, y_{n+1}) \quad (2.4)$$

egylépéses numerikus módszert, ahol Φ a módszert meghatározó függvény.

2.2.1. Definíció. Azokat a módszereket, amelyekre $\Phi \equiv \Phi(\tau, t_n, y_n)$ –azaz a Φ függvény nem függ y_{n+1} -től– explicit módszereknek nevezzük. Amennyiben $\Phi \equiv \Phi(\tau, t_n, y_n, y_{n+1})$, a módszert implicitnek nevezzük.

2.2.3. Numerikus módszerek hibái és konvergenciája

2.2.2. Definíció. Legyen $t_* \in [0, T]$. Az $e_{t_*}(\tau) = y_n - u(t_*)$ függvényt a Φ numerikus módszer t_n pontbeli globális diszkretizációs hibafüggvényének nevezzük, ahol $t_* = n\tau$. Azt mondjuk, hogy a Φ numerikus módszer konvergens a t_n pontban, ha

$$\lim_{\tau \rightarrow 0} e_{t_*}(\tau) = 0.$$

Ha a Φ numerikus módszer konvergens a $[0, T]$ minden pontjában, akkor egyszerűen konvergens módszernek nevezzük. A (2.2.2) konvergenciájának rendjét a Φ numerikus módszer konvergenciarendjének nevezzük.

2.2.3. Definíció. Az $l_n(\tau) = \tilde{u}_n - u(t_n)$ függvényt, ahol $\tilde{u}_{n+1} = u(t_n) + \tau\Phi(\tau, t_n, u(t_n), \tilde{u}_{n+1})$ és $t_n \in \omega_\tau$ a Φ -numerikus módszer t_n pontbeli lokális diszkretizációs hibafüggvényének nevezzük.

2.2.4. Definíció. A

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

függvényt a (2.4) alakú Φ numerikus módszer $t_n \in \omega_\tau$ pontbeli lokális approximációs hibafüggvényének (vagy képlethibájának) nevezzük. Azt mondjuk, hogy a Φ numerikus módszer p -edrendben konzisztens, ha

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

valamely $p > 0$ állandóval tetszőleges $t_n \in \omega_\tau$ rácspontra.

A (2.2) explicit Euler-módszer esetén a képlethiba $g_n(\tau)$ meghatározásához az

$$u(t_{n+1}) = u(t_n) + \tau u'(t_n) + \frac{\tau^2}{2} u''(t_n) + \mathcal{O}(\tau^3),$$

Taylor-sorfejtés és $u'(t_n) = f(t_n, u(t_n))$ összefüggés felhasználásával adódik, hogy

$$g_n(\tau) = -y(t_{n+1}) + y(t_n) + \tau f(t_n, y(t_n)) = -\frac{\tau^2}{2} y''(t_n) + \mathcal{O}(\tau^3).$$

Ez azt jelenti, hogy a módszer elsőrendben konzisztens. A (2.3) implicit Euler-módszer képlethibáját hasonló módon határozhatjuk meg és vezethetjük le, hogy a módszer elsőrendben konzisztens.

2.3. Runge-Kutta módszerek

Az explicit és az implicit Euler-módszer is csak elsőrendben konzisztens. Ezt a rendet úgy többek között úgy tudjuk növelni, ha a t_n és t_{n+1} rácspont között több helyen is kiértékeljük a függvényt és ezen értékeket alkalmas módon súlyozzuk. Az ezen elven alapuló módszercsaládok vezetnek az ún. *Runge-Kutta módszerekhez*. A módszercsalád a nevét Carl Runge-ról és Martin Willhelm Kutta-ról kapta. Runge 1985-ös cikkében mutatta be az alapötletet az első modern numerikus differenciálegyenlet megoldó módszerre vonatkozóan [9]. Később Kutta továbbfejlesztette a módszert, amely ma az általánosan ismert Runge-Kutta módszerek alapját képezi [8].

Ahhoz, hogy ezeket a magasabbrendű módszereket származtassuk, az alábbi

$$\int_{t_n}^{t_{n+1}} f(t, u(t)) dt$$

integrált kell minél jobban közelítenünk valamilyen kvadratura alapú módszerrel. Ha ismerünk vagy, legalább ki tudunk számolni csomópontbeli ún. *lépcső* értékeket a $[t_n, t_{n+1}]$ intervallumon, akkor ezeknek a lineáris kombinációja jó becslés lehet az integrálra. Természetesen ezen módszer családnál is megkülönböztethetünk explicit, illetve implicit módszereket. Ezeket a következőkben külön fogjuk tárgyalni.

2.3.1. Explicit Runge-Kutta módszerek

2.3.1. Definíció. Legyen $s \geq 1$ egy adott egész szám. A rögzített a_{ij} , c_i , $b_i \in \mathbb{R}$ értékek mellett

$$Y_i = y_n + \tau \sum_{j=1}^s a_{ij} f(t_n + c_j \tau, Y_j), \quad i = 1, 2, \dots, s, \quad (2.5)$$

$$y_{n+1} = y_n + \tau \sum_{i=1}^s b_i f(t_n + c_i \tau, Y_i), \quad i = 1, 2, \dots, s \quad (2.6)$$

numerikus módszert s -lépcsős Runge-Kutta típusú módszernek nevezzük.

2.3.1. Megjegyzés.

- A Runge-Kutta típusú módszereket $s^2 + 2s$ darab paraméter határozza meg. Egy adott módszert egy mátrix és két vektor segítségével szokták megadni, ahol $A = (a_{ij})_{i,j=1}^s \in \mathbb{R}^{s \times s}$, $b^T = (b_i)_{i=1}^s \in \mathbb{R}^s$, valamint $c = (c_i)_{i=1}^s \in \mathbb{R}^s$. Ezeket a paramétereket az alábbi

c_1	a_{11}	a_{12}	$\cdots$	a_{1s}
c_2	a_{21}	a_{22}	$\cdots$	a_{2s}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
c_s	a_{s1}	a_{s2}	$\cdots$	a_{ss}
	b_1	b_2	$\cdots$	b_s

ún. Butcher-tablóban tároljuk.

- A (2.5)-(2.6) s -lépcsős Runge-Kutta módszerrel ekvivalens felírás

$$k_i = f \left(t_{n-1} + c_i \tau, y_{n-1} + \tau \sum_{j=1}^s a_{ij} k_j \right), \quad i = 1, \dots, s, \quad (2.7)$$

$$y_{n+1} = y_n + \tau \sum_{i=1}^s b_i k_i. \quad (2.8)$$

A fenti (2.7)-(2.8) felírás a (2.5)-(2.6) felírással azért tekinthető ekvivalensnek, mert azok pontossági feltételei megegyeznek. A (2.7)-(2.8) alak a deriváltfüggvényt közelíti köztes pontokban a megoldásfüggvény helyett.

2.3.2. Megjegyzés. Könnyen látható, hogyha a módszer A mátrixa szigorúan alsóháromszög mátrix, akkor a Runge-Kutta módszer explicit.

A következőkben nézzünk meg néhány nevezetesebb Runge–Kutta típusú módszert, illetve, hogy ezeket milyen közelítésekből kaphatjuk meg.

Ha a (2.3)-es integrált az

$$\int_{t_n}^{t_{n+1}} f(t, u(t)) dt \approx \tau f\left(t_{n+\frac{1}{2}}, y\left(t_{n+\frac{1}{2}}\right)\right)$$

középponti szabály segítségével szeretnénk közelíteni, akkor az implicit részt explicit módon közelítve (lásd [1]) a

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f\left(y_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}k_1\right), \\ y_{n+1} &= y_n + \tau k_2 \end{aligned}$$

ún. *explicit középponti módszerhez*, vagy *javított Euler módszerhez* jutunk.

Az egyik legismertebb és leggyakrabban alkalmazott Runge–Kutta módszer az ún. RK4, amely a Simpson–szabályon alapul. Ennek alakja:

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f\left(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}k_1\right), \\ k_3 &= f\left(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}k_2\right), \\ k_4 &= f(t_n + \tau, y_n + \tau k_3), \\ y_{n+1} &= y_n + \tau \left(\frac{k_1}{6} + \frac{2k_2}{6} + \frac{2k_3}{6} + \frac{k_4}{6}\right). \end{aligned}$$

Lentebb meadjuk az eddig definiált explicit Runge–Kutta módszerekhez tartozó Butcher tablókat.

- Explicit Euler

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

- Javított Euler

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

- RK4

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

2.3.2. Runge–Kutta módszerek rendfeltételei

Az újonnan származtatott módszerek esetében nemtriviális többváltozós Taylor-sorfejtés után adódik a magasabb konzisztenciarend. A lépcsőszámok növelésével természetesen adódik, hogy a magasabb konzisztencia rendet garantáló *rendfeltételek* száma is növekszik.

Egyfajta konvencióként alapvető elvárás lehet a Runge–Kutta módszerekkel szemben, hogy a konstans függvény integráljára pontos legyen, azaz, hogy teljesüljön az

$$\int_0^1 c \, dt = c = \sum_{i=1}^s b_i f(c_i) = \sum_{i=1}^s b_i c$$

összefüggés. Ebből adódik, hogy a $\sum_{i=1}^s b_i = 1$ feltételnek teljesülnie kell. Ez a legelső rendfeltétel, melynek vektoralakja

$$b^T e,$$

ahol e a megfelelő egységvektor. A Runge–Kutta módszerek másik alapvető struktúrális feltétele, hogy a c vektor elemei megegyeznek az A mátrix sorösszegeivel, azaz

$$Ae = c.$$

Ez nem klasszikus rendfeltétel, hanem a módszer koherenciáját biztosító összefüggés. Ha ez nem teljesül, akkor a módszer nem értelmezhető standard Runge–Kutta módszerként.

A

$$c^k = [c_1^k, c_2^k, \dots, c_s^k]^T \in \mathbb{R}^s \quad \text{és} \quad C = \text{diag}[c_1, c_2, \dots, c_s] \in \mathbb{R}^{s \times s}$$

jelöléseket bevezetve a 2.1. táblázatban felsoroljuk a $p = 4$ renddel bezárólag a Runge–Kutta módszerek rendfeltételeit.

p	rendfeltételek
1	$b^T e = 1$
2	$b^T c = \frac{1}{2}$
3	$b^T c^2 = \frac{1}{3}, \quad b^T A c = \frac{1}{6}$
4	$b^T c^3 = \frac{1}{4}, \quad b^T C A c = \frac{1}{8},$ $b^T A c^2 = \frac{1}{12}, \quad b^T A^2 c = \frac{1}{24}$

2.1. táblázat. A Runge–Kutta módszerek rendfeltételei negyedrenddel bezárólag.

Ezek alapján ellenőrizhető, hogy a javított Euler-módszer másodrendben, míg az RK4 negyedrendben konzisztens.

Megfigyelhető, hogy az eddig vizsgált Runge–Kutta módszereknél a lépcsők száma megegyezik a renddel. Ez azonban csak $p = s = 4$ -ig igaz. Ez $s \geq 5$ -nél nem teljesül. Általában nem ismert, hogy egy tetszőleges p -edrendű módszerhez hány darab lépcső szükséges. A 2.2. táblázat összefoglalja, hogy az adott rendű módszerhez hány feltétel, lépcső és paraméter tartozik.

rend	1	2	3	4		5	6		7		8
feltételek száma	1	2	4	8		17	37		85		200
lépcsők száma	1	2	3	4	5	6	7	8	9	10	11
paraméterek száma	1	3	6	10	15	21	28	36	45	55	66

2.2. táblázat. A Runge–Kutta módszerek rendfeltételeinek, lépcsőinek és paramétereinek száma.

2.3.3. Megjegyzés. A 2.2. táblázat magyarázatot ad arra is, hogy például az ötlépcsős Runge–Kutta módszer, miért nem lehet ötödrendű. Ahogy látható a szabad paramétereinek a száma 15, azonban az ötödrendűséghez 17 feltételt kell kielégítenie, így nem lehetséges olyan paraméterezés, amely minden szükséges rendfeltételt teljesítene.

2.3.4. Megjegyzés. Jogosan merül fel a kérdés, hogy melyik Runge–Kutta módszert érdemes alkalmazni egy adott probléma esetén. A többlépcsős módszerek általában nagyobb pontosságot és jobb konzisztenciát kínálnak, azonban a paraméterek számának növekedése jelentős számítási többlet terhet is jelent, amely nem minden feladatnál indokolt. Éppen ezért az egyik leggyakrabban alkalmazott módszer a klasszikus négy lépcsős Runge–Kutta módszer (RK4), amely sok gyakorlati esetben jó kompromisszumot nyújt a számítási költség és a pontosság között. Ugyanakkor előfordulhat, hogy még az RK4 is csak kis lépéshossz mellett szolgáltat kellően pontos eredményt, ezért a megfelelő módszer kiválasztása mindig az adott probléma sajátosságaitól függ.

2.3.3. Implicit Runge–Kutta módszerek

Amennyiben a Runge–Kutta módszert definiáló $A \in \mathbb{R}^{s \times s}$ mátrix nem szigorúan alsóháromszög mátrix, akkor a módszert *implicit Runge–Kutta (IRK) módszernek* nevezzük. A nemlineáris algebrai egyenletrendszer megoldásában könnyebbé teszi, ha az A alsó háromszög mátrix. Ezek az ún. *diagonálisan implicit Runge–Kutta (DIRK) módszerek*. A számításmatematikai költségei ellenére a következő tulajdonságaik miatt szeretnek implicit módszereket alkalmazni:

- ugyanolyan lépcsőszám mellett magasabb rendben pontosak;
- általában jobb stabilitási tulajdonságokkal rendelkeznek.

Az explicit Runge–Kutta (ERK) módszerekhez hasonlóan itt megadjuk a már említett implicit módszerek Butcher-táblóit.

- Implicit Euler

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

- Implicit középponti módszer

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

További magasabbrendű IRK módszereket hasonlóan az explicit módszerekhez szoftizáltabb kvadratúrákból származtathatunk. Ezek megtalálásához ún. kollokációs módszert használunk, amelyek a Gauss alappontokon alapulnak (lásd [1], 4.7.1.).

- A *Radau-módszernél* olyan alappontokat választunk, ahol az intervallum egyik végpontja (vagyis $c_1 = 0$ vagy $c_s = 1$) hozzátartozik az alappontokhoz. A módszerek pontossága $p = 2s - 1$. Attól függően, hogy a kezdő- vagy végpontot választjuk megkülönböztetünk RadauI és RadauII módszereket.
- A *Lobatto-módszer* olyan alappont megválasztású módszer, amikor az intervallum mindkét végpontja hozzátartozik az alappontokhoz ($c_1 = 0$ és $c_s = 1$). A módszer pontossága $p = 2s - 2$.
- Amennyiben az intervallum végpontjai nem alappontok, akkor a Gauss alappontok megválasztásával a *Gauss–Legendre-módszer* pontossága $p = 2s$.

2.4. Többlépéses módszerek

A Runge–Kutta módszereknél eddig úgy értünk el magasabb konzisztenciarendet, hogy egy részintervallumon belül a jobboldalt több ponton is kiértékeljük. Létezik azonban egy másik lehetőség is: nemcsak az aktuális lépés előtti rácpontban vett közelítést használjuk fel, hanem az azt megelőző értékeket is bevonjuk a számításba. A legelső 1855-ös többlépéses módszer Augustus De Morgan nevéhez fűződik. Az egyik legismertebb többlépéses módszert Francis Bashforth publikálta 1883-as cikkében [4], azonban a módszereket J. C. Adams dolgozta ki.

Az általános definíció megadása előtt vezessük be az $f_n = f(t_n, y_n)$ jelölést.

2.4.1. Definíció. Az adott $\alpha_0, \alpha_1, \dots, \alpha_k$ és $\beta_0, \beta_1, \dots, \beta_k$ együtthatók mellett az

$$\alpha_0 y_{n+1} + \alpha_1 y_n + \dots + \alpha_k y_{n+1-k} = \tau [\beta_0 f_{n+1} + \beta_1 f_n + \dots + \beta_k f_{n+1-k}], \quad (2.9)$$

iterációt, ahol $n + 1 = k, k + 1, \dots$, lineáris, k -lépéses módszernek nevezzük.

2.4.1. Megjegyzés.

- A továbbiakban feltesszük, hogy $\alpha_0 \neq 0$, hiszen ellenkező esetben nem tudnánk y_{n+1} értékét meghatározni. Konvenció szerint feltesszük, hogy $\alpha_0 = 1$.
- A $\beta_0 = 0$ választásával explicit módszert kapunk, egyébként a többlépéses módszer implicit.
- Egy (2.9) többlépéses módszert az α_j és β_j paramétereivel definiálunk. Ha két módszer együtthatói konstansszorosai egymásnak, akkor a két módszert nem különböztetjük meg.

A Runge–Kutta módszerekhez hasonlóan itt is felmerül a kérdés, hogy a módszerek rendjét itt is tudják-e garantálni a különböző együtthatók közötti összefüggések. Tekintsük először a módszerek képlethibáját:

$$g_n(\tau) = \sum_{j=0}^k [\alpha_j u(t_{n+1-j}) - \tau \beta_j f(t_{n+1-j}, u(t_{n+1-j}))].$$

Az $u'(t_{n+1-j}) = f(t_{n+1-j}, u(t_{n+1-j}))$ összefüggés mellett adódik, hogy

$$g_n(\tau) = \sum_{j=0}^k [\alpha_j u(t_{n+1-j}) - \tau \beta_j u'(t_{n+1-j})].$$

Ezután az $u(t)$ és $u'(t)$ függvényeket p -ed, illetve $p - 1$ -ed renddel bezárólág a $t = t_n$ pont körüli Taylor-sorfejtéssel adódik, hogy

$$g_n(\tau) = d_0 u(t_{n+1}) + \tau d_1 u'(t_{n+1}) + \tau^2 d_2 u''(t_{n+1}) + \dots + \tau^p d_p u^{(p)}(t_{n+1}) + \mathcal{O}(\tau^{p+1}),$$

ahol

$$\begin{aligned} d_0 &= \sum_{j=0}^k \alpha_j, \\ d_1 &= -\sum_{j=0}^k j \alpha_j + \beta_j, \\ d_2 &= \sum_{j=0}^k \frac{1}{2} j^2 \alpha_j + j \beta_j, \\ &\vdots \\ d_p &= (-1)^p \sum_{j=0}^k \frac{1}{p!} j^p \alpha_j + \frac{1}{(p-1)!} j^{p-1} \beta_j. \end{aligned}$$

Ezen levezetésnek köszönhetően adódik, hogy egy többlépéses módszer p -edrendű, ha $d_0 = d_1 = \dots = d_p = 0$. Azaz az alábbi tételt bizonyítottuk be.

2.4.1. Tétel. *A k -lépéses módszer p -edrendű, ha a módszert definiáló paraméterekre teljesülnek az alábbi feltételek:*

$$\begin{aligned} \alpha_0 &= 1, \quad \sum_{j=0}^k \alpha_j = 0, \\ \frac{1}{s} \sum_{j=0}^k j^s \alpha_j + \sum_{j=0}^k j^{j-1} \beta_j &= 0, \quad s = 1, 2, \dots, p. \end{aligned}$$

2.4.1. Adams-módszerek

A lineáris többlépéses módszerek egyik nagy családja az *Adams-módszerek*, amelyek a kezdetiérték-feladatok baloldali deriváltját az előre felé haladó $\frac{y_{n+1} - y_n}{\tau}$ véges differenciával közelítik. Azaz az általános k -lépéses felírásban $\alpha_0 = 1, \alpha_1 = -1, \alpha_2 = \alpha_3 = \dots = \alpha_k = 0$.

2.4.2. Definíció. *Az*

$$y_{n+1} - y_n = \tau \sum_{j=0}^k \beta_j f_{n+1-j}$$

alakú módszereket k -lépéses Adams-módszereknek nevezzük.

2.4.2. Megjegyzés.

- Az Adams-módszereknek két nagy családja az *explicit Adams–Bashforth-módszerek*, valamint az *implicit Adams–Moulton-módszerek*.
- Az egyik legismertebb explicit Adams módszer a kétlépéses Adams–Bashforth-módszer, melynek alakja:

$$y_{n+1} = y_n + \tau \left(\frac{3}{2} f_n - \frac{1}{2} f_{n-1} \right).$$

- Az általános levezetésünkéből adódik, hogy a k -lépéses Adams–Bashforth-módszerek konzisztenciarendje $p = k$.
- Az Adams–Moulton-módszereknél a legelterjedtebb az egylépéses verziója, mely az explicit és implicit Euler-módszerek $1/2$ súllyal vett lineáris kombinációja, az ún. trapéz-módszer, melynek alakja:

$$y_{n+1} = y_n + \tau \left(\frac{1}{2}f_n + \frac{1}{2}f_{n+1} \right)$$

- Az általános levezetésünkéből adódik, hogy a k -lépéses Adams–Moulton-módszerek konzisztenciarendje $p = k + 1$.

Gyakorlatban az implicit Adams–Moulton-módszereket explicit módon implementálják. Először egy Adams–Bashforth-módszerrel kiszámolt y_n értékkel megjósoljuk a t_n időrétegen a közelítő értéket, majd az Adams–Moulton-módszerrel javítjuk a numerikus megoldást a következő módon: a jobboldalon szereplő $\beta_0 f_n$ implicit f_n tag helyett $f_n^* = f(t_n, y_n^*)$ értékkel dolgozunk. Az így kapott módszereket *prediktor-korrektor* (jósló-javító) módszernek nevezzük.

2.4.2. BDF-módszerek

A másik nevezetes többlépéses módszercsalád a *retrográd differencia módszerek* (angolul: *Backward Differentiation Formulas* ~ BDF). Amíg az Adams-típusú módszerekben a deriváltat előre felé haladó véges differenciahányadossal közelítettük, valamint a jobboldalt egy interpoláló polinommal approximáltuk a korábbi f értékeken, addig a BDF-módszerek esetén ennek fordítottját alkalmazzák. Itt a deriváltakat visszafelé haladó véges differenciahányadosok segítségével közelítjük, azaz több korábbi y értéket vonunk be a baloldalra. A módszer általános alakjában a jobboldalon csak az aktuális $f(t_{n+1}, y_{n+1})$ szerepel, azaz egyedül $\beta_0 \neq 0$.

2.4.3. Definíció. Legyen $\nabla^j y_{n+1}$ a visszafelé haladó differencia operátor j -szer alkalmazva y -ra a t_{n+1} pontban. A

$$\sum_{j=0}^k \frac{1}{j} \nabla^j y_{n+1} = \beta_0 \tau f(t_{n+1}, y_{n+1})$$

alakú többlépéses módszert BDF-módszernek nevezzük.

2.4.3. Megjegyzés.

- A $k = 1$ esetre vonatkozó BDF-módszer az implicit Euler-módszert adja.
- Az általános levezetésünkéből következik, hogy a BDF-módszerek konzisztenciarendje $p = k$.
- A stabilitási tulajdonságok egyik fontos következménye, hogy gyakorlatban a $k > 6$ lépésű BDF-módszereket gyakorlatban nem alkalmazzák.

3. fejezet

Numerikus módszerek stabilitása

Az eddigiekben Runge–Kutta és többlépéses módszerek különböző családjait és ahhoz tartozó konzisztenciarendjeit ismerhettük meg. Ebben a fejezetben a módszerek stabilitási tulajdonságait ismerhetjük meg.

3.1. A Dahlquist-féle tesztfeladat

A numerikus módszerek alkalmazásakor fontos cél a pontos megoldás közelítése mellett, hogy a módszer használatakor fellépő hibák ne vezessenek a megoldás eltorzulásához. Egy módszert intuitívan akkor nevezhetünk *stabilnak*, ha a képlethiba által szolgáltatott hiba korlátosan akkumulálódik.

Tekintsük az

$$\begin{aligned}u'(t) &= \lambda u(t), \quad \lambda \in \mathbb{C}, \\u(0) &= 1\end{aligned}$$

ún. *Dahlquist-féle tesztfeladatot*, melynek megoldása $u(t) = e^{\lambda t}$. Ez $t \rightarrow \infty$ esetén különböző $\lambda \in \mathbb{C}$ értékek mellett más minőségű megoldást eredményez. Egy numerikus megoldásnál célunk, hogy az is lekövesse a függvény viselkedését. Annak van értelme ebben az esetben, ha a megoldás korlátos marad, azaz amikor $\operatorname{Re}(\lambda) \leq 0$.

Alkalmazzuk az explicit és implicit Euler-módszereket a tesztfeladatra. Ekkor az alábbi összefüggések adódnak:

- Explicit Euler-módszer

$$y_{n+1} = (1 + \tau\lambda)y_n,$$

- Implicit Euler-módszer

$$y_{n+1} = \frac{1}{1 - \tau\lambda}y_n.$$

Általánosságban az egylépéses módszerek esetében a tesztfeladatra való alkalmazás után az

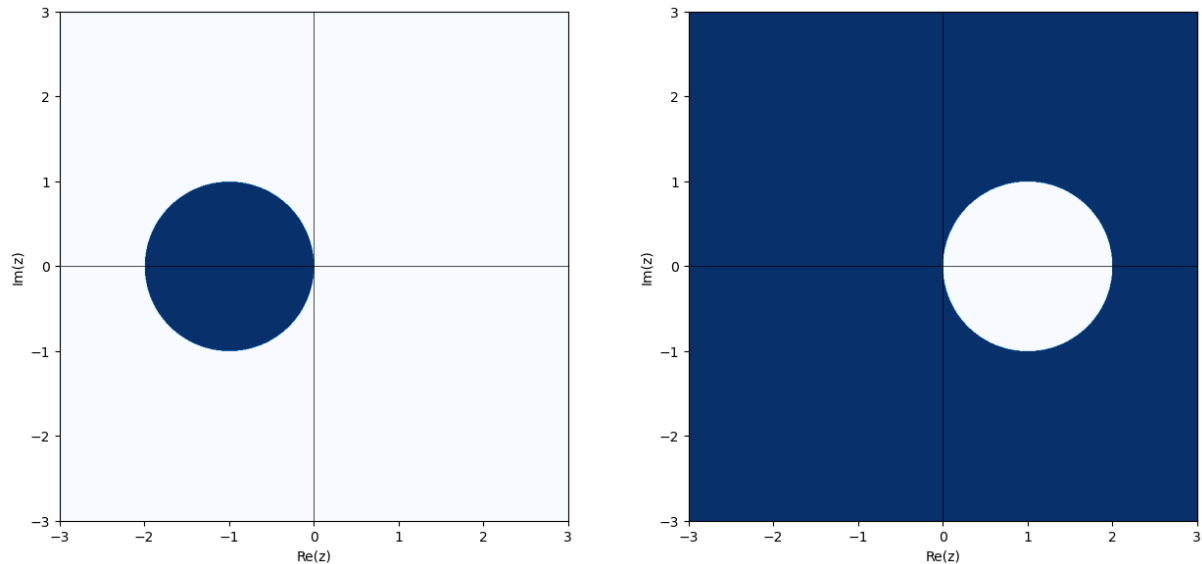
$$y_{n+1} = R(z)y_n \tag{3.1}$$

összefüggés adódik, ahol $z = \tau\lambda \in \mathbb{C}$. A (3.1) összefüggésben szereplő $R(z)$ függvényt a numerikus módszer *stabilitási függvényének* nevezzük. A kívánatos $\operatorname{Re}(\lambda) \leq 0$ feltétel az

$$|R(z)| \leq 1 \tag{3.2}$$

feltétel teljesülését vonja maga után. esetén fog teljesülni. Azon $z \in \mathbb{C}$ komplex számok halmazát, amelyre (3.2) fennáll, a módszer *abszolút stabilitási tartományának* nevezzük.

A 3.1. ábrán nyilvánvaló válik, hogy számításmatematikai költségessége ellenére az implicit Euler-módszer sok esetben a nagy stabilitási tartománya következtében preferálandó az explicit Euler-módszerrel szemben.



3.1. ábra. Az explicit (bal) és implicit (jobb) Euler-módszerek stabilitási tartományai.

3.2. A-stabilitás és merevség

3.2.1. Definíció. Azt mondjuk, hogy egy numerikus módszer *A-stabil*, ha a stabilitási tartománya tartalmazza a $\mathbb{C}_0^- = \{z \in \mathbb{C} : R(z) \leq 0\} \subset \mathbb{C}$ komplex félsíkot.

Mint ahogyan azt a 3.1. ábrán is láthatjuk, az implicit Euler módszer rendelkezik az A-stabilitás tulajdonságával. Ez a gyakorlatban azt jelenti, hogy a módszer jól tudja követni a korlátos megoldás viselkedését tetszőleges lépésköz választása esetén. Intuíció alapján egy explicit módszer nem lesz A-stabil, míg egy implicit módszer az lehet. Vizsgáljuk meg, hogy magasabbrendű módszereknél is igaz lesz-e ez az intuíció.

Runge–Kutta módszerek

Kezdeti lépésként határozzuk meg a módszerek *stabilitási függvényét*. Egy tetszőleges p -edrendű Runge–Kutta módszert a tesztfeladatra adódik, hogy a stabilitási függvénye

$$R(z) = 1 + zb^T(I - zA)^{-1}e.$$

Elegendően kicsi τ mellett az $(I - zA)^{-1}$ Neumann sorát felírva, majd ebbe helyettesítve a p -edrendű módszer rendjét felhasználva adódik, hogy

$$R(z) = 1 + \sum_{k=1}^p z^k \frac{1}{k!} + \sum_{k=p+1}^{\infty} z^k (b^T A^{k-1} e).$$

Explicit Runge–Kutta (ERK) esetben kihasználva az A mátrix szigorúan alsóhármászög tulajdonságát adódik, $A^k = 0$, $k \geq s + 1$. Ekkor az ERK stabilitási függvénye

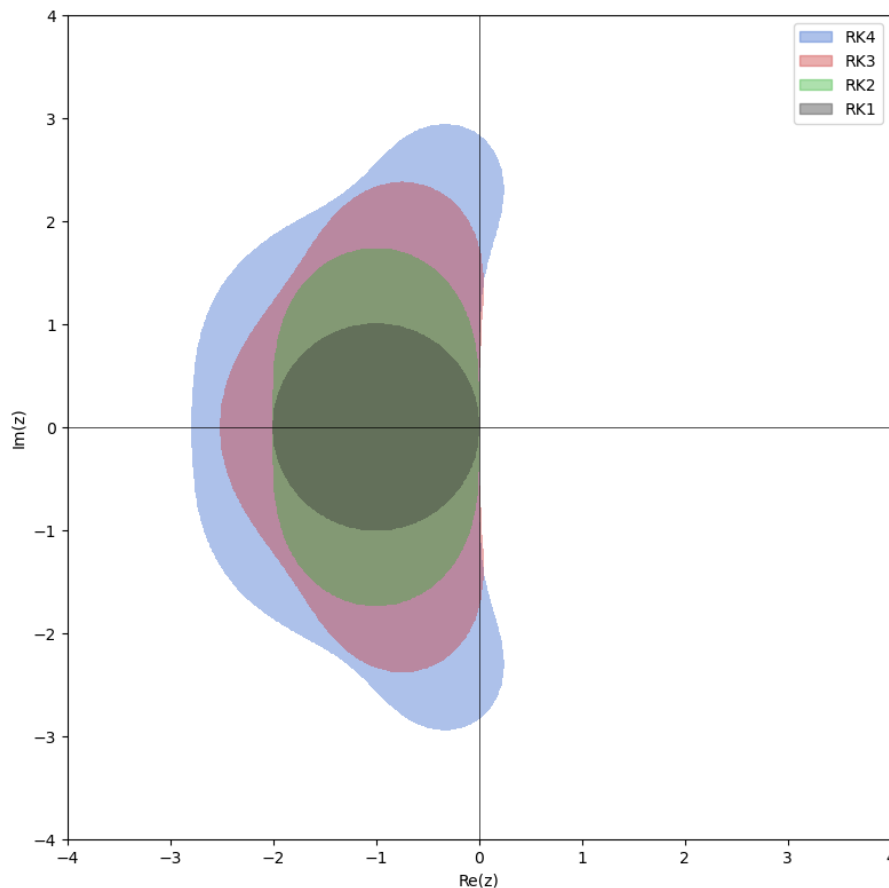
$$R(z) = 1 + \sum_{k=1}^p z^k \frac{1}{k!} + \sum_{k=p+1}^s z^k (b^T A^{k-1} e).$$

Az ERK módszerek stabilitási függvénye egy legfeljebb s -edfokú polinom. Ennek következtében ez $|z| \rightarrow \infty$ esetén szintén tart a végtelenbe. Emiatt az ERK módszerek *nem lesznek A -stabilak*, hiszen a stabilitási tartományuk nem tartalmazza a teljes $\mathbb{C}_0^-$ komplex félsíkot.

Mivel $k = 1, 2, 3, 4$ -re a módszerek rendje megegyezik a lépcsők számával, így ezek stabilitási függvénye

$$R(z) = 1 + \sum_{k=1}^p z^k \frac{1}{k!}.$$

Stabilitási tartományukat a 3.2. ábrán láthatjuk.



3.2. ábra. Az ERK1–4 módszerek stabilitási tartományai.

A következőkben vizsgáljuk meg a Runge–Kutta módszerek stabilitási tulajdonságait.

3.2.1. Tétel. *Egy tetszőleges Runge–Kutta módszer stabilitási függvénye*

$$R(z) = \frac{\det(I - zA + ze \cdot b^T)}{\det(I - zA)}. \quad (3.3)$$

Bizonyítás. A (2.5)-(2.5) s -lépcsős Runge–Kutta módszer képletét a tesztfeladatra alkalmazva adódik az

$$Y = y_n e + \tau A \lambda Y = y_n e + z A Y$$

összefüggés, azaz

$$Y_i - z \sum_{j=1}^s a_{ij} Y_j = y_n \quad i = 1, \dots, s. \quad (3.4)$$

Másrészt

$$y_{n+1} = (1 + z b^T (I - z A)^{-1} e) y_n,$$

így

$$y_{n+1} - z \sum_{j=1}^s b_j Y_j = y_n. \quad (3.5)$$

Ekkor (3.4)-(3.5) egy lineáris algebrai egyenletrendszert jelent az $Y_1, Y_2, \dots, k_s, y_{n+1}$ összesen $s + 1$ darab ismeretlenre nézve. Jelölje $y \in \mathbb{R}^{s+1}$ ezen ismeretlenek vektorát. Ekkor az egyenletrendszer felírható $By = f$ alakban, ahol a rendszer $f \in \mathbb{R}^{s+1}$ jobb oldala $f = y_n e$ alakú, míg a rendszer $B \in \mathbb{R}^{(s+1) \times (s+1)}$ együtthatómátrixa

$$B = \begin{pmatrix} 1 - za_{11} & -za_{12} & \cdots & -za_{1s} & 0 \\ -za_{21} & 1 - za_{22} & \cdots & -za_{2s} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -za_{s1} & -za_{s2} & \cdots & 1 - za_{ss} & 0 \\ -zb_1 & -zb_2 & \cdots & -zb_s & 1 \end{pmatrix}.$$

Ebből a rendszerből az y_{n+1} ismeretlent a Cramer-szabállyal meghatározva adódik, hogy

$$y_{n+1} = \frac{\det(B_f)}{\det(B)},$$

ahol

$$B_f = \begin{pmatrix} 1 - za_{11} & -za_{12} & \cdots & -za_{1s} & y_n \\ -za_{21} & 1 - za_{22} & \cdots & -za_{2s} & y_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -za_{s1} & -za_{s2} & \cdots & 1 - za_{ss} & y_n \\ -zb_1 & -zb_2 & \cdots & -zb_s & y_n \end{pmatrix}.$$

Könnyen látható, hogy $\det(B)$ értékét az utolsó oszlop szerinti kifejtéssel közvetlenül meghatározhatjuk, és ekkor

$$\det(B) = \det(I - zA).$$

Továbbá $\det(B_f)$ értékének meghatározásához vonjuk ki az utolsó sorát rendre a felette lévő sorokból. Ezután az utolsó oszlop szerinti kifejtéssel már látható, hogy

$$\det(B_f) = y_n \det(I - zA + ze \cdot b^T).$$

A két determináns értéket helyettesítve adódik az állítás. □

A (3.3) racionális törtfüggvénynek az e^z pontos megoldást kell közelítenie, amit az IRK módszereknél Padé-típusú közelítéssel tudunk megtenni. A részleteket mellőzve az alábbi stabilitási eredmények igazolhatók:

- Egy s -lépcsős Radau-féle IRK módszer stabilitási függvénye az exponenciális függvény $R_{s-1,s}(z)$ maximális rendű Padé-típusú approximációja, így ezek a módszerek A-stabilak.
- Egy s -lépcsős Lobatto-féle IRK módszer stabilitási függvénye az exponenciális függvény $R_{s-1,s-1}(z)$ maximális rendű Padé-típusú approximációja, így ezek a módszerek A-stabilak.
- Egy s -lépcsős Gauss–Legendre-féle IRK módszer stabilitási függvénye az exponenciális függvény $R_{s,s}(z)$ maximális rendű Padé-típusú approximációja, így ezek a módszerek A-stabilak.

Többlépéses módszerek

A (2.9) lineáris többlépéses módszerek (LTM-ek) stabilitásvizsgálata a Runge–Kutta módszerekhez képest összetettebb feladat és terjedelme korlátok következtében mindössze egy tömör összefoglalással élünk. A stabilitáshoz kötődő egyik kulcsfogalom a *gyökkritérium* lesz.

3.2.2. Definíció. A

$$\rho(\xi) = \sum_{j=0}^k \alpha_j \xi^{k-j} \quad (3.6)$$

k -adfokú polinomot a m -lépéses LTM első karakterisztikus polinomjának nevezzük.

3.2.3. Definíció. Azt mondjuk, hogy egy LTM-re teljesül a gyökkritérium, ha a (3.6) első karakterisztikus polinom ξ_j gyökeire teljesülnek, hogy

- $|\xi_j| \leq 1$, $j = 1, \dots, k$, illetve
- a $|\xi_j| = 1$ tulajdonságú gyökök egyszeresek.

Ezen definíciók mellett bizonyítás nélkül ismertetjük a LTM-ekre vonatkozó konvergencia állítást.

3.2.2. Tétel ([5], 5.2. Tétel). Ha egy LTM konzisztens és érvényes rá a gyökkritérium, akkor konvergens is.

LTM-ek esetén az A -stabilitás bizonyítása nemtriviális és sajnálatos módon mindössze néhány módszer és csak alacsony konzisztenciarenddel bír a kívánatos tulajdonsággal. Ezek az ún. *első és második Dahlquist korlátok*, melyek szerint

- az explicit LTM-ek nem lehetnek A -stabilak;
- az A -stabil LTM-ek rendje nem lehet kettőnél nagyobb.

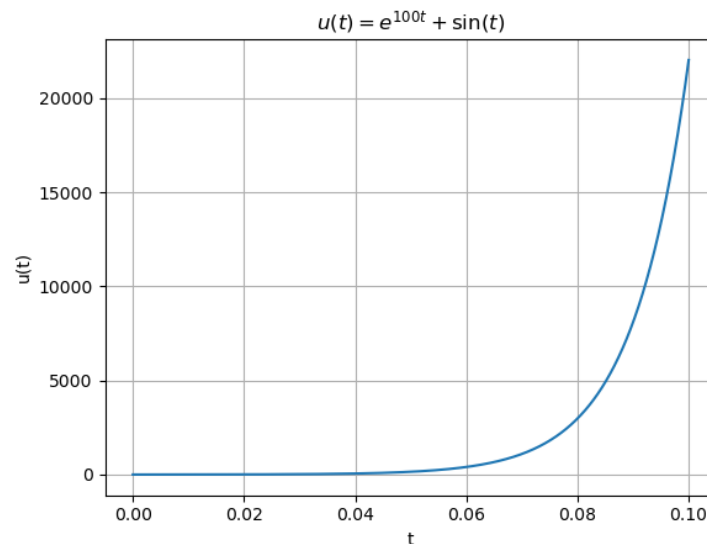
Merevség

Sok modell esetén előfordul, hogy a rendszer különböző komponensei eltérő időskálán változnak, azaz egyes részek gyorsan, míg mások lassan módosulnak. A merevség definiálása sokáig intuitív alapú volt. Többnyire úgy értelmezték, hogy olyan feladatok, amelyeken az explicit módszerek szignifikánsan rosszabbul működnek az implicit módszerekhez képest. Erre a leírásra az ad magyarázatot, hogy a feladatban jellemző gyorsabb fázist egy explicit módszer csak nagyon kis lépéshossz mellett képes követni.

Példa gyanánt tekintsük a következő Cauchy-feladatot:

$$\begin{aligned} u'(t) &= 100(u(t) - \sin(t)) + \cos(t), \quad t \in [0, 1], \\ u(0) &= 1. \end{aligned}$$

Könnyen látható, hogy itt a $100(u(t) - \sin(t))$ tag gyorsan, míg a $\cos(t)$ tag lassan változik az időben. A megoldás más skálán történő léte a megoldás ábrázolása során is nyilvánvaló, melyet a 3.3. ábrán láthatunk.



3.3. ábra. A példa megoldása: $u(t) = e^{100t} + \sin(t)$.

A merev feladatok többféle értelemben is kihívás elé állít egy numerikus módszert. A merevség mértékének megállapítása és definiálása egy közel hatvan éves folyamatot vett igénybe és szofisztikált logaritmikus Lipschitz normákon alapuló technikákat követel [11].

A következő fejezetben a merev feladatok megoldására egy hatékony numerikus módszerrel, az ún. *implicit-explicit* (IMEX) típusú módszer családdal foglalkozunk.

4. fejezet

IMEX-módszerek

A fizika, kémia és biológia területén gyakran felmerülő advekcio-diffúzió, vagy reakció-diffúzió parciális differenciálegyenletek (PDE-k) tipikusan merev és nem merev tagokból állnak. Ennek numerikus megoldására elterjedt módszer az implicit-explicit (IMEX) módszerek használata.

Az ilyen problémáknál a diffúziós tagok inkább merevek, míg a konvekciós, vagy reakciós tagok nem merevek. Ebből kifolyólag térbeli diszkretizáció után célszerű időben az előbbieket impliciten, míg az utóbbiakat explicit módon diszkretizálni.

Tekintsünk egy időfüggő parciális differenciálegyenletet, melyben a térbeli deriváltakat valamilyen klasszikus térbeli módszerrel (például véges differencia sémákkal, vagy spektrál módszerrel) diszkretizáltunk. Ekkor a kapott szemidiszkrét, mégpedig időben folytonos közönséges differenciálegyenlet (KDE) rendszert az alábbi általános alakban írhatjuk fel:

$$u' = f(u) + \nu g(u) \quad (4.1)$$

Itt f a konvekciós, vagy reakciós tagot reprezentálja, ami általában a nem merev nem-lineáris tag és így ezt érdemes explicit módszerrel kezelni. A $\nu g(u)$ azonban egy lineáris diffúziós tag ν nemnegatív paraméterrel, így erre a tagra merevsége miatt érdemes implicit módszert alkalmazni, mely jobb stabilitási tulajdonságokkal bír egy explicit módszerhez képest.

A továbbiakban a (4.1) feladatra alkalmazott speciális IMEX módszercsaládok konkrét sémáival, valamint ezek stabilitási tulajdonságaival foglalkozunk. Két fő módszerosztályt különböztetünk meg:

- többlépéses IMEX módszerek;
- IMEX Runge–Kutta módszerek.

Mindkét osztály sajátos előnyökkel és kihívásokkal rendelkezik, és más-más alkalmazási területeken bizonyulhat hatékonynak. Összeségében elmondható, hogy a többlépéses IMEX módszerek univerzálisabbak, azonban vannak problémák, amelyekre az IMEX Runge–Kutta módszerek jobb teljesítményt nyújtanak. A következő alfejezetekben részletesen bemutatjuk a két módszercsalád jellemzőit, példákat adunk gyakran használt sémákra, és elemezzük azok stabilitási tulajdonságait.

4.1. Többlépéses IMEX-módszerek

Először tekintsük a lineáris többlépéses IMEX módszereket, illetve azok általános alakját. Legyen τ a lépésköz és y_n a $t_n = \tau n$ időpontban lévő numerikus megoldás. A (4.1) feladatra alkalmazott k -lépéses IMEX séma alakja

$$\frac{1}{\tau}y_{n+1} + \frac{1}{\tau} \sum_{j=0}^{k-1} a_j y_{n-j} = \sum_{j=0}^{k-1} b_j f(y_{n-j}) + \nu \sum_{j=-1}^{k-1} c_j g(y_{n-j}), \quad (4.2)$$

ahol $c_{-1} \neq 0$. Elegendően sima $u(t)$ megoldásfüggvény mellett a (4.2) módszerhez tartozó lokális approximációs hibára $t = t_n$ pont körüli Taylor-sorfejtéssel és közvetlen számolással adódnak a p -edrendű rendfeltételek:

$$\begin{aligned} 1 + \sum_{j=0}^{k-1} a_j &= 0, \\ 1 - \sum_{j=1}^{k-1} j a_j &= \sum_{j=0}^{k-1} b_j = \sum_{j=-1}^{k-1} c_j, \\ \frac{1}{2} + \sum_{j=1}^{k-1} \frac{j^2}{2} a_j &= - \sum_{j=1}^{k-1} j b_j = c_{-1} - \sum_{j=1}^{k-1} j c_j, \\ &\vdots \\ \frac{1}{p!} + \sum_{j=1}^{k-1} \frac{(-j)^p}{p!} a_j &= \sum_{j=1}^{s-1} \frac{(-j)^{p-1}}{(p-1)!} b_j = \frac{c_{-1}}{(p-1)!} + \sum_{j=1}^{k-1} \frac{(-j)^{p-1}}{(p-1)!} c_j. \end{aligned} \quad (4.3)$$

Bizonyítás nélkül közöljük az alábbi állítást.

4.1.1. Tétel ([2], Tétel 2.1.). *A (4.2) k -lépéses IMEX módszerekre az alábbi állítások igazak:*

- (a) *A (4.3) rendfeltételben lévő $2p+1$ feltétel lineárisan független, így létezik k -adrendű k -lépéses IMEX módszer.*
- (b) *Egy k -lépéses IMEX módszer rendje nem lehet magasabb, mint k .*
- (c) *Egy k -lépéses k -adrendű IMEX módszer k paraméterrel rendelkezik.*

Próbljunk meg egy egylépéses elsőrendű módszert konstruálni. A 4.1.1. Tétel értelmében ez megvalósítható egy paraméter mellett, melyet jelöljünk γ -val. A nagy csonkolási hiba elkerülése érdekében legyen $0 \leq \gamma \leq 1$. Ekkor adódik, hogy

$$y_{n+1} - y_n = \tau f(y_n) + \nu \tau ((1 - \gamma)g(y_n) + \gamma g(y_{n+1})). \quad (4.4)$$

Ha (4.4)-ben $\gamma = 0$, akkor az explicit Euler-módszert kapjuk, mely teljes explicit mivolta miatt, nem ad hozzá a vizsgálandó feladathoz. Ugyanakkor, ha $\gamma = 1$, akkor az

$$y_{n+1} - y_n = \tau f(y_n) + \tau \nu g(y_{n+1}) \quad (4.5)$$

numerikus módszer a legegyszerűbb implicit, ún. *IMEX Euler-módszerhez* vezet. A (4.5) módszer a merev tagot implicit, a nem merev tagot pedig explicit Euler-módszerrel

kezeli és olykor az irodalomban szokás szemi-implicit BDF (SBDF) módszernek is nevezni.

Az IMEX Euler-módszer nyilvánvaló előnye, hogy viszonylag egyszerű és gyorsan kiszámítható. Ugyanakkor, amint azt a stabilitás vizsgálat során látni fogjuk, bizonyos alkalmazásokban instabilitás léphet fel.

Másodrendű kétlépéses IMEX módszerek esetén a 4.1.1. Tétel alapján két paraméter áll majd rendelkezésünkre: γ és c . Ha a leendő másodrendű sémát a $n + \gamma$ időpontra centráljuk, akkor az alábbi

$$\frac{1}{\tau} \left[\left(\gamma + \frac{1}{2} \right) y_{n+1} - 2\gamma y_n + \left(\gamma - \frac{1}{2} \right) y_{n-1} \right] = (\gamma + 1)f(y_n) - \gamma f(y_{n+1}) + \nu \left[\left(\gamma + \frac{c}{2} \right) g(y_{n+1}) + (1 - \gamma - c)g(y_n) + \frac{c}{2}g(y_{n-1}) \right] \quad (4.6)$$

családot származtatjuk. A (4.6) módszercsalád tartalmazza a leggyakrabban használt többlépéses IMEX módszereket.

- A $(\gamma, c) = (1/2, 0)$ választás az ún. *CNAB* sémát eredményezi. Az elnevezést az inspirálja, hogy a módszer az implicit részben az ún. Crank–Nicolson formulát, míg az explicit részben a másodrendű Adam–Bashforth módszert használja.
- A CNAB módszerhez hasonló, azonban jobb csillapítási tulajdonságokkal rendelkező $(\gamma, c) = (1/2, 1/8)$ változatot *módosított CNAB (MCNAB)* módszernek nevezzük.
- A $(\gamma, c) = (0, 1)$ megválasztásával a módszer az ún. *leap frog* (magyarul bakugrás) sémát használja expliciten, míg az implicit részre a Crank–Nicolson formulát. Emiatt ezen megválasztással a módszer *CNLF* néven ismert.
- Végül a $(\gamma, c) = (1, 0)$ megválasztással egy másodrendű SBDF módszer adódik.

4.1.1. Stabilitásvizsgálat

A stabilitásvizsgálathoz tekintsük a (4.1) szemidiszkrét KDE rendszer komplex skaláris esetét, mint tesztegysenletet:

$$u' = i\beta u + \alpha u. \quad (4.7)$$

A fenti (4.7) tesztfeladat természetes választás egy advekcio-diffúzió PDE feladat esetén térbeli diszkretizáció (klasszikus véges differencia technika és diszkrét Fourier transzformáció) után. A β konstans az advekcios, míg az α konstans a diffúziós taghoz tartozik.

A (4.7) tesztegysenletre a (4.2) k -lépéses IMEX módszert alkalmazva adódik, hogy

$$\frac{1}{\tau} y_{n+1} + \frac{1}{\tau} \sum_{j=0}^{k-1} a_j y_{n-j} = \sum_{j=0}^{k-1} b_j (i\beta) y_{n-j} + \sum_{j=-1}^{k-1} c_j \alpha y_{n-j}. \quad (4.8)$$

A (4.8) egy állandó együtthatós lineáris differenciaegyenlet, melynek megoldása

$$y_{n+1} = p_1 \xi_1^n + p_2 \xi_2^n + \cdots + p_k \xi_k^n$$

alakban adható meg, ahol ξ_j a j -edik gyöke a

$$\Phi(z) = (1 - c_{-1}\alpha\tau)z^k + \sum_{j=0}^{k-1} (a_j - b_j i\beta\tau - c_j \alpha\tau) z^{k-j-1}$$

karakterisztikus egyenletnek, valamint p_j konstans ξ_j egyszeres gyök esetén, míg polinomiális n -ben egyébként. Azaz stabilitásról akkor beszélhetünk, ha $|\xi_j| \leq 1$ esetén a ξ_j gyökök egyszeresek, illetve $|\xi_j| < 1$ esetén a ξ_j gyökök többszörösek.

A (4.1) tesztfeladatra történő (4.2) módszer alkalmazása során adódik az

$$y_{n+1} = R(\alpha, \beta)y_n$$

összefüggés, ahol

$$R(\alpha, \beta) = \frac{1 + \tau\alpha(1 - \gamma) + i\tau\beta}{1 - \tau\gamma\alpha}. \quad (4.9)$$

A módszer stabilitási tartománya azon (α, β) párokból áll, amelyekre (4.9)-ben fennáll, hogy $|R(\alpha, \beta)| \leq 1$.

Vegyük észre, hogy ezek az egy lépéses sémák triviálisan kezelik a változó időlépést, és viszonylag kevés tároló kapacitást igényelnek. A $\gamma = 1$ választás különösen vonzó, mivel erős csillapítás lép fel, ha α nagy és negatív. Azonban az elsőrendű IMEX-sémák hátránya, hogy instabillá válnak a nem nulla β -tengely közelében, ugyanis

$$|R(0, \beta)| = |1 + i\tau\beta| = \sqrt{1 + \tau^2\beta^2} > 1.$$

Ebből kifolyólag érdemes magasabbrendű módszereket használni. A térbeli deriváltak klasszikus másodrendű közelítése miatt, a fókuszunk a másodrendű módszereken lesz. A (4.6) másodrendű módszert a (4.1) tesztfeladatra történő alkalmazása után az alábbi karakterisztikus függvényhez juthatunk:

$$\Phi(z) = \left[\gamma + \frac{1}{2} - \alpha\tau \left(\gamma + \frac{c}{2} \right) \right] z^2 - [2\gamma + i\beta\tau(\gamma + 1) + \alpha\tau(1 - \gamma - c)]z + \gamma - \frac{1}{2} + i\beta\tau\gamma - \alpha\tau\frac{c}{2}. \quad (4.10)$$

Könnyen látható, hogy ezek a módszerek $(\alpha, \beta) = (0, 0)$ -ra stabilak lesznek, hiszen a (4.10) karakterisztikus egyenlet gyökei 1 és $\frac{2\gamma-1}{2\gamma+1}$, azaz legfeljebb 1 értékűek és így ezek a sémák stabilak az origóban.

Mivel (4.10)-ben α értéke rendre c -vel van szorozva, így a c értéket $|\alpha| \gg 1/k$ esetére vonatkozó stabilitási tulajdonságok alapján választjuk meg. Ekkor felírhatunk egy

$$\left(\gamma + \frac{c}{2} \right) \xi^2 + (1 - \gamma - c)\xi + \frac{c}{2} = 0$$

közelítő karakterisztikus egyenlet, melynek gyökei

$$\xi_{1,2} = \frac{\gamma + c - 1 \pm \sqrt{(1 - \gamma)^2 - 2c}}{2\gamma + c}.$$

Bármely (γ, c) esetén a

$$D_{\gamma,c} := \max(|\xi_1|, |\xi_2|) \leq 1 \quad (4.11)$$

kiértékelése becslést ad a nagy ν esetén fellépő nagyfrekvenciás csillapításra. Ne feledjük, hogy a vizsgálatink során ν be van ágyazva az α értékbe.

A c szerinti minimalizálás meghatározza azt a módszert, amely adott γ mellett a legerősebb aszimptotikus nagyfrekvenciás csillapítást biztosítja. Ez [10] alapján:

$$c = \begin{cases} \frac{(1-\gamma)^2}{2}, & \text{ha } \gamma > 0, \\ \frac{1}{2}, & \text{ha } \gamma = 0. \end{cases} \quad (4.12)$$

Továbbá a

$$c = 1 - \gamma \quad (4.13)$$

választás szintén hasznosnak bizonyul (lásd [10]). Az SBDF, MCNAB és CNLF sémák kielégítik a (4.12) feltételt, az SBDF és CNLF sémák a (4.13) feltételt, azonban a CNAB egyiknek sem felel meg. Továbbá, a (4.11) összefüggés γ és c szerinti minimalizálása alapján az SBDF séma rendelkezik a legerősebb aszimptotikus csillapítással a másodrendű módszerek között.

A stabilitási tulajdonságok mélyebb megértéséhez az $\alpha\beta$ -sík stabilitási kontúrjainak vizsgálata szükséges. Terjedelmi okok miatt erre nem teszünk kísérletet.

4.2. IMEX Runge–Kutta módszerek

A többlépcsős IMEX sémákat sújtó fő nehézség annak köszönhető, hogy viszonylag kis időlépések szükségesek a stabilitási korlátozások miatt. Az ilyen korlátozások súlyosbodnak, ha magasabbrendű többlépcsős sémákkal dolgozunk, ahogy azt [2] is taglalja. Másrészt láthattuk, hogy az ERK módszerek stabilitási tartománya még magasabb (három és négy) lépcsőszám mellett is moderált és csak kis mértékben növekszik. Ellenpólusként megvizsgáljuk a megfelelő lépcsőszámú IMEX Runge–Kutta módszerek hatékonyságát.

Először tekintsük, hogyan tudjuk a $\nu = 1$ melletti (4.1) szemidiszkrét rendszert diszkrétizálni a Runge–Kutta módszerek filozófiája szerint.

A merev g tagra alkalmazzunk egy s -lépcsős DIRK módszert, melyet az $A \in \mathbb{R}^{s \times s}$, $c, b \in \mathbb{R}^s$ együtthatók jellemeznek. A nem merev f tagra alkalmazzunk egy $\sigma = s + 1$ lépcsős ERK módszert, amelyben $\hat{c}^T = [0 \ c]$ és $\hat{A} \in \mathbb{R}^{\sigma \times \sigma}$, $\hat{b} \in \mathbb{R}^\sigma$. Ahhoz, hogy a DIRK módszer is σ lépcsős legyen a Butcher-tablójába beírunk egy csupa nulla sort és oszlopot ("dummy" sort és oszlopot). Ekkor a következő

$$\begin{array}{c|cccccc} 0 & 0 & 0 & 0 & \cdots & 0 \\ c_1 & 0 & a_{11} & 0 & \cdots & 0 \\ c_2 & 0 & a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & 0 & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & 0 & b_1 & b_2 & \cdots & b_s \end{array}$$

felfűjt Butcher-tablót kapjuk. Az így kapott új mátrixot és vektorokat jelölje $\tilde{A} \in \mathbb{R}^{\sigma \times \sigma}$, $\tilde{b}, \tilde{c} \in \mathbb{R}^\sigma$. Ez a jelölés egyszerűsíti a rendfeltételek felírását, mivel az explicit és implicit részek szorosabb összehangolását teszi lehetővé.

Tekintsük az IMEX Runge–Kutta módszer egy lépését. Legyen $\hat{K}_1 = f(y_n)$. Ekkor minden $i = 1, \dots, s$ -re számítsuk ki K_i -t az alábbi módon:

$$K_i = g(y_i),$$

$$\text{ahol } y_i = y_n + \tau \sum_{j=1}^i a_{ij} K_j + \tau \sum_{j=1}^i \hat{a}_{i+1,j} \hat{K}_j.$$

Ezután $\hat{K}_{i+1} = f(y_i)$ kiértékelése után megkapjuk az

$$y_{n+1} = y_n + \tau \sum_{j=1}^s b_j K_j + \tau \sum_{j=1}^{\sigma} \hat{b}_j \hat{K}_j \quad (4.14)$$

s -lépcsős IMEX Runge–Kutta módszert. Az általános (4.14) sémán kívül két módszer alcsaládot különböztetünk meg:

- Ha a $\hat{b} = \tilde{b}$, ahol $\hat{b}_1 = 0$, akkor (4.14) az alábbi módon módosul:

$$y_{n+1} = y_n + \tau \sum_{j=1}^s b_j (K_j + \hat{K}_{j+1})$$

- Amennyiben $\hat{b}_{s+1} = 0$, akkor $\hat{K}_{s+1}$ -et nem kell kiértékelni. Továbbá, ha $b_j = a_{sj}$ és $\hat{b}_j = \hat{a}_{s+1,j}$ minden $i = 1, \dots, s$ -re, akkor az $u_{n+1} = u_s$ összefüggés adódik.

Megjegyezzük továbbá, hogy az explicit séma most már előállítható egy általános, s -lépcsős explicit sémából, amely a $c_1, \dots, c_{s-1}$ abszcisszákon alapul.

0	0	0	0	$\dots$	0
c_1	$\hat{a}_{21}$	0	0	$\dots$	0
c_2	$\hat{a}_{31}$	$\hat{a}_{32}$	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
c_{s-1}	$\hat{a}_{s1}$	$\hat{a}_{s2}$	$\hat{a}_{s3}$	$\dots$	0
	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\dots$	$\hat{b}_s$

A továbbiakban nézzük pár példát IMEX Runge–Kutta módszerekre. Alkalmazzuk legegyszerűbb ERK és DIRK módszereket, azaz az implicit és explicit Euler-módszerrel. Ezek felfűjt tablóí

0	0	0	és	0	0	0
1	0	1		1	1	0
	0	1			1	0

Az ezek alapján kapott séma

$$y_{n+1} = y_n + \tau(f(y_n) + g(y_{n+1})). \quad (4.15)$$

Vezessük be az (s, σ, p) jelölést, ahol s az implicit módszer lépcsőszámát, σ az explicit módszer lépcsőszámát, míg p a kombinált módszer rendjét jelöli. Ekkor az (4.15) variáns az ún. $(1, 1, 1)$ IMEX Euler-módszer.

A fenti $(1, 1, 1)$ IMEX Euler-módszer nem elégíti ki a $\hat{b} = \tilde{b}$ összefüggést. Az a módszer, amely ezt kielégíti a következő módosított táblókkal rendelkezik:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0 & 1 \\ \hline & 0 & 1 \end{array} \quad \text{és} \quad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 0 & 1 \end{array}.$$

A kapott $(1, 2, 1)$ IMEX Euler-módszer a séma a következő alakot ölti:

$$y_{n+1} = y_n + \tau(g(y_1) + f(y_1)), \text{ ahol} \\ y_1 = y_n + \tau(g(y_1) + f(y_n)).$$

A mostani verzió eggyel több f kiértékelést von maga után. A fenti két IMEX Euler-módszerhez képest másodrendű pontosságú séma úgy állítható elő, hogy az explicit részhez az explicit középponti módszert, az implicit részhez pedig az implicit középponti módszert alkalmazzuk. Ekkor nyerjük a

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \hline & 0 & 1 \end{array} \quad \text{és} \quad \begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

felfűjt Butcher-tablójú $(1, 2, 2)$ IMEX közepéépponti módszert. A módszer másodrendű, mivel mind az explicit, mind az implicit rész külön-külön is másodrendű, továbbá $\tilde{c} = \hat{c}$. Ez az IMEX séma jellemzően hasonló teljesítményt nyújt, mint a széles körben használt CNAB-módszer, ugyanakkor szimmetriatulajdonságai kedvezőbbek [3].

A legjobb csillapítási tulajdonságokkal rendelkező kétlépcsős, harmadrendű DIRK séma az

$$\begin{array}{c|cc} \eta & \eta & 0 \\ 1 - \eta & 1 - 2\eta & \eta \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Butcher-tablójú séma [6], ahol $\eta = (3 + \sqrt{3})/6$. Ez más sémákkal összehasonlítva jelentős előnyt mutat a merevségi határ közelében. Például az $(1, 2, 2)$ IMEX középponti módszer nem biztosít csillapítást a $c\tau \rightarrow \infty$ esetben, míg ez a módszer jól viselkedik ebben az esetben is. Ehhez a DIRK sémához társítható egy

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \eta & \eta & 0 & 0 \\ 1 - \eta & 1 - \eta & 2(1 - \eta) & 0 \\ \hline & 0 & \frac{1}{2} & \frac{1}{2} \end{array}$$

harmadrendű ERK, amely ugyan háromlépcsős, de hasonló tulajdonságokat mutat, mint a korábban ismert háromlépcsős explicit módszerek. Összességében ez a kombinált $(2, 3, 3)$ IMEX módszer nemcsak hogy harmadrendű pontosságot ér el, de a stabilitási tartománya kedvező, ugyanis csillapít bizonyos képzetes irányú komponenseket, valamint a stabilitási függvénye a merevségi határon is megfelelő csillapítást biztosít.

A kombinált $(2, 3, 3)$ IMEX módszerhez képest még jobb tulajdonságú, egy ún. merev pontos (lásd [7]) másodrendű módszer DIRK módszer felel az implicit részért, melynek Butcher táblója

$$\begin{array}{c|cc} \eta & \eta & 0 \\ 1 & 1 - \eta & \eta \\ \hline & 1 - \eta & \eta \end{array},$$

ahol $\eta = (2 - \sqrt{2})/2$. Az ehhez tartozó másodrendű ERK módszer Butcher-tablója

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \eta & \eta & 0 & 0 \\ 1 & \rho & 1 - \rho & 0 \\ \hline & 0 & 1 - \eta & \eta \end{array}.$$

A módszer stabilitási tulajdonságainak javítása érdekében egy ρ paramétert úgy választanak meg, hogy a stabilitási függvény illeszkedjen az exponenciális függvény harmadrendű Taylor-sorához. Ez garantálja a módszer harmadrendű pontosságát és disszipatív (csillapító) viselkedését. A megfelelő stabilitási tartomány $\rho = -2\sqrt{2}/3$ választással adódik. A kombinált módszer egy $(2, 3, 2)$ IMEX módszerhez vezet.

Módosítsuk az előző esetet úgy, hogy a $\hat{b} = \tilde{b}$ feltétel teljesülése helyett a $b_j = a_{sj}$ és $\hat{b}_j = \hat{a}_{s+1,j}$ feltétel teljesüljön. A DIRK rész nem változik, azoban az ERK módszer Butcher-tablója a következőképpen változik:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \eta & \eta & 0 & 0 \\ 1 & \rho & 1 - \rho & 0 \\ \hline & \rho & 1 - \rho & 0 \end{array},$$

ahol $\rho = 1 - 1/(2\eta)$. A kombinált módszer $(2, 2, 2)$ IMEX DIRK módszert eredményezi.

4.2.1. Stabilitás kérdése

Az előző részben megismerkedhettük, hogyan tudunk egyszerűbb IMEX Runge–Kutta módszereket származtatni. A következőkben tömören megnézzük, hogy bizonyos esetekben melyik módszert célszerű alkalmazni és melyeket kerüljük egyes problémák megoldásánál.

Tekintsünk a $\nu = 1$ melletti (4.1) tesztegyenletet. Ha erre alkalmazzuk a (4.14) IMEX Runge–Kutta módszerek általános sémáját, akkor felírhatjuk a séma egy időlépését, mint

$$y_{n+1} = y_n \tau (\alpha + \beta i) \sum_{j=1}^s b_j y_j,$$

azaz $y_{n+1} = R(\alpha, \beta) y_n$ alakban, ahol

$$y_i = \frac{(1 + \tau \beta i \hat{a}_{i+1,1}) y_n + \tau \sum_{j=1}^{i-1} (\alpha a_{ij} + \beta i \hat{a}_{i,j+1}) y_j}{1 - \tau \alpha a_{ii}}, \quad i = 1, \dots, s.$$

Az $(1, 1, 1)$ IMEX Euler-módszer stabilitási függvényére

$$R(\alpha, \beta) = \frac{1 + \tau \beta i}{1 - \tau \alpha}$$

adódik. Az $(1, 2, 1)$ IMEX Euler-módszer stabilitási függvénye:

$$R(\alpha, \beta) = 1 + \tau (\alpha + \beta i) \cdot \frac{1 + \tau \beta i}{1 - \tau \alpha}.$$

A technikai elemzést elhagyva [2] alapján elmondható, hogy ha a diffúzió dominál, akkor az $(1, 1, 1)$ IMEX Euler-módszer prerefált, míg advekción dominált esetben pedig az $(1, 2, 1)$ IMEX Euler-módszer.

4.2.2. Numerikus kísérlet

Tekintsünk egy konkrét reakció-diffúzió típusú PDE feladatot. Nevezetesen

$$u_t = D u_{xx} + r u(1 - u),$$

ahol $D > 0$ a diffúziós együttható, $r > 0$ pedig a logisztikus növekedési tag szorozója. Az egyenlet a Fisher-Kolmogorov típusú modellek családjába tartozik, amelyeket széles körben alkalmaznak például populációdinamikai, ökológiai és kémiai reakciófolyamatok leírására. Ezen belül is a Fisher-Kolmogorov-Petrovsky-Piskunov modell utazó hullám megoldásait szeretnénk a $(2, 2, 2)$ IMEX DIRK módszerrel meghatározni, valamint összehasonlítani másodrendű, de standard Runge-Kutta módszerekkel. Ezek az explicit javított Euler és az implicit középponti módszerek voltak.

A térbeli tartomány az egyszerűség kedvéért legyen $[0, 1]$. A kezdeti feltétellel és Dirichlet-peremmel ellátott feladat ekkor:

$$\begin{aligned} u_t(x, t) &= \nu u_{xx}(x, t) + r u(x, t)(1 - u(x, t)), \quad 0 < x < 1, \quad t > 0, \\ u(x, 0) &= \frac{1}{1 + e^{10(x-0.3)}}, \\ u(0, t) &= 1, \quad u(1, t) = 0. \end{aligned}$$

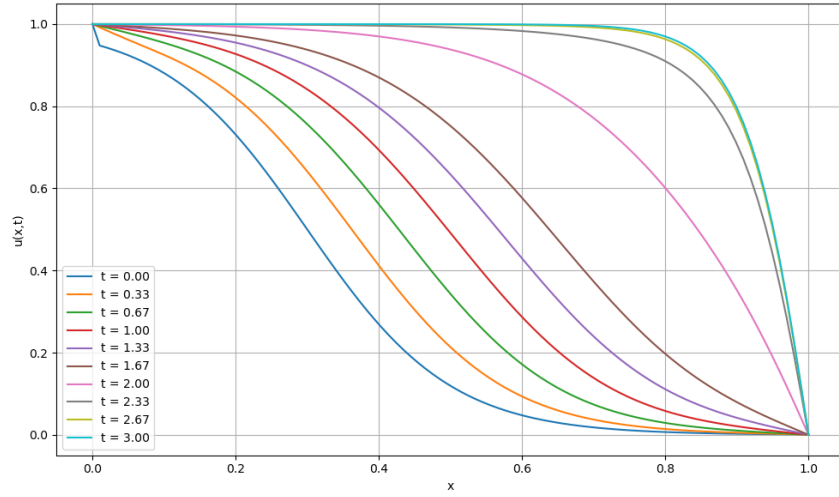
A diffúziós tagot implicit módon, míg a nemlineáris reakcióért felelő tagot explicit módon kezeljük. Ahhoz, hogy a problémát IMEX módszerrel tudjuk vizsgálni, először a térbeli változók mentén kell diszkretizálnunk az egyenletet. Ennek legegyszerűbb módja, ha a második deriváltat másodrendű centrális véges differenciával közelítjük. A diszkretizáció eredményeként adódik, hogy

$$u' = f(u) + g(u),$$

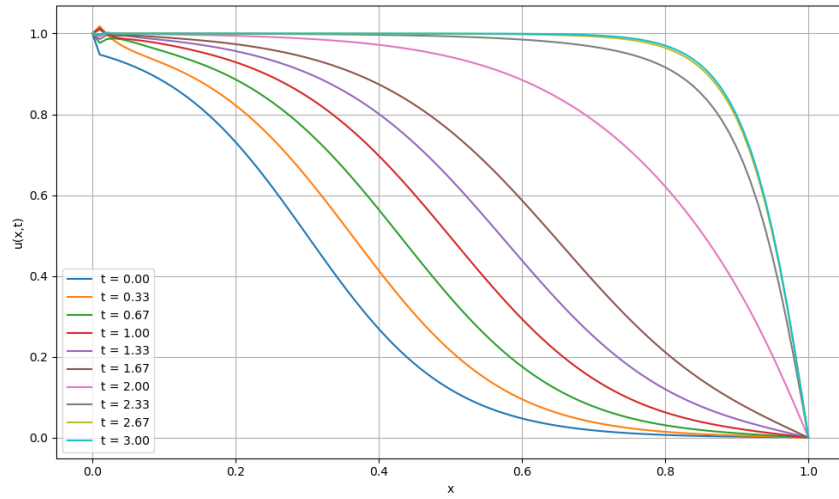
ahol $f(u) = ru(1 - u)$ és $g(u) = \nu Au$. Az A mátrix Laplace differenciáloperátor diszkrét megfelelőjeként tekinthető triadiagonális mátrix lesz egy másodrendű centrális véges differencia séma alkalmazása után.

A diffúziós együttható $\nu = 0.01$, a reakciós tagot $r = 4$ volt. A térbeli intervallumot 100 részre osztottuk, míg az időintervallum a $[0, 3]$ volt.

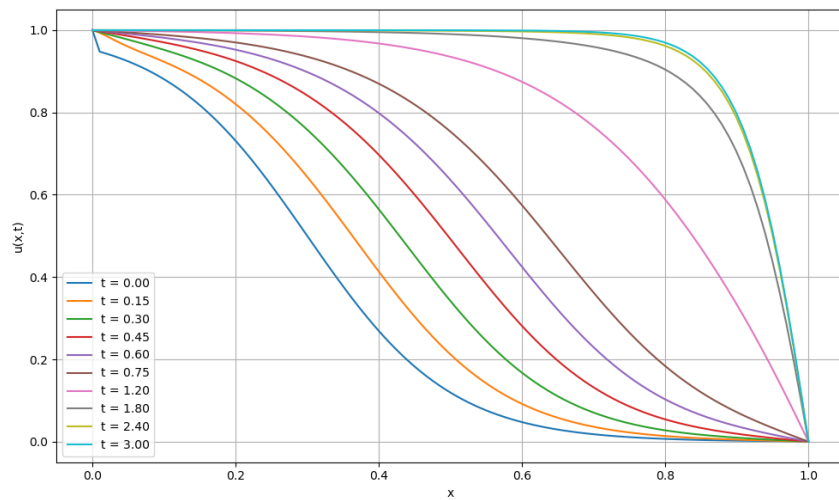
A 4.1. ábrán látható eredmény eléréshez, azaz, hogy az explicit javított Euler módszer módszer ne szálljon el az A -stabilitás hiánya miatt 600 időbeli lépésre volt szükség. Ezzel szemben az implicit középponti módszer és a $(2, 2, 2)$ IMEX DIRK módszer már kevesebb, mindössze 10 időlépéssel is jól teljesített köszönhetően a jó stabilitási tulajdonságaiknak (lásd 4.2. és 4.3. ábrák). Ugyanakkor az ábrákon az is jól látható, hogy az implicit középponti módszer megoldása a térrészekenél zajos, valamint ennél a példánál nem jelentős, de 0,2 másodperccel lassabb futási eredményt produkált a többlépcsős $(2, 2, 2)$ IMEX DIRK módszerhez képest. Összességében az IMEX módszer a legstabilabb eredményt és leggyorsabb futási időt szolgáltatotta a Fisher-Kolmogorov feladatra.



4.1. ábra. Az explicit javított euler megoldása a fenti Fisher–Kolmogorov feladatra.



4.2. ábra. Az implicit középponti módszer megoldása a fenti Fisher–Kolmogorov feladatra.



4.3. ábra. A (2,2,2) IMEX DIRK módszer a fenti Fisher–Kolmogorov feladatra.

5. fejezet

Összegzés

Az implicit-explicit (IMEX) sémák népszerűsége nem véletlen: elegáns keretrendszer adnak merev és nem merev tagokat tartalmazó parciális differenciálegyenletek numerikus megoldására, amely képes ötvözni az explicit eljárások egyszerűségét és az implicit lépések stabilitását. Minden egyes családnak, sőt, minden konkrét sémának megvan a maga erőssége: egyesek különösen stabilak domináns diffúzió, vagy erősen nemlineáris tag mellett, míg mások könnyen implementálhatók és csekély memóriaigényűek, megint mások pedig magas pontosságot kínálnak viszonylag nagy lépésközök mellett.

Ugyanakkor az is látható, hogy univerzális IMEX séma nem létezik. Könnyen találhatók olyan feladatok, ahol egy teljesen explicit, vagy teljesen implicit eljárás –megfelelő adaptív lépésvezérléssel– hatékonyabban teljesít, mint bármely kompromisszumos IMEX megoldás. A merev problémák körében az IMEX módszerek ugyan kézenfekvőek, de a konkrét sémaválasztás messze nem egyértelmű: a stabilitási tartomány, a csillapítási tulajdonságok, a pontosság és a számítási költség gyakran egymás ellen dolgozó tulajdonságok, melyeknél a priorizálás elengedhetetlen.

A többlépéses IMEX sémák széles körben elterjedtek, mert egyszerű a struktúrájuk és relatíve alacsony a memória használatuk, viszont lépésköz és rácsfinomsági korlátjaik gyakran túl szigorúvá válnak a domináns nemlineáris tagok, vagy nagyon finom térbeli felbontást igénylő szituációkban. Ezzel szemben az IMEX Runge–Kutta módszerek nagyfokú rugalmasságot kínálnak: magasabb rendre skálázhatók, és paramétereik révén finomhangolhatók a kívánt stabilitási viselkedésre. Ára azonban a jelentősebb számítási költség, továbbá egyes erősen diffúzív problémákon mutató instabilitásuk.

Az IMEX módszerek népszerűségét nem egy mindent megoldó "varázs-séma" adja, hanem a szemléletmód, amely elvezet a megfelelő megoldáshoz. A merevség szerkezetének felismerése, a tagok megfelelő implicit–explicit szétválasztása, valamint a feladat sajátosságaihoz illeszkedő lépésköz- és sémaválasztás teszi lehetővé, hogy a numerikus megoldás egyszerre legyen stabil, hatékony és pontos. A kutatás és a gyakorlat előtt továbbra is nyitott feladat marad új, adaptív és probléma specifikus IMEX variánsok kifejlesztése, amelyek a jelenlegi kompromisszumokat még kedvezőbb arányban oldják fel.

A dolgozatban lévő ábrák kódjai nyilvánosan elérhetők a következő GitHub repozitóriumban: https://github.com/zhorlik/Thesis_code.

Irodalomjegyzék

- [1] U. M. Ascher, L. R. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM, 1998
- [2] U. M. Ascher, S. J. Ruuth, R. J. Spiteri, *Implicit-explicit Runge–Kutta methods for time-dependent partial differential equations*, Applied Numerical Mathematics, 25(2–3), 151–167, 1997
- [3] U. M. Ascher, S. J. Ruuth, B. T. R. Wetton, Implicit-explicit methods for time-dependent PDEs, *SIAM Journal on Numerical Analysis*, 32(3):797–823, 1995
- [4] F. Bashforth, *An Attempt to Test the Theories of Capillary Action by Comparing the Theoretical and Measured Forms of Drops of Fluid*, Cambridge University Press, 1883
- [5] István Faragó, *Numerikus modellezés és közönséges differenciálegyenletek numerikus megoldási módszerei*, ELTE TTK jegyzet, Budapest, 2013
- [6] E. Hairer, S. P. Norsett, G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*, Springer, New York, 1993
- [7] E. Hairer, G. Wanner, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, Springer, New York, 1991
- [8] M. W. Kutta, *Beitrag zur näherungsweise Integration totaler Differentialgleichungen*, Zeitschrift für Mathematik und Physik, **46**, 435–453, 1901
- [9] C. Runge, *Über die numerische Auflösung von Differentialgleichungen*, Mathematische Annalen, **46**, 167–178, 1895
- [10] S. Ruuth. *Implicit-explicit methods for time-dependent PDE's*, Master's thesis, Institute of Applied Mathematics, University of British Columbia, Vancouver, 1993
- [11] G. Söderlind, L. Jay, M. Calvo: *Stiffness 1952-2012: Sixty years in search of a definition*, BIT Numerical Mathematics 55 (2), 531-558, 2015
- [12] W. Walter.: *Ordinary Differential Equations*. Springer, Graduate Texts in Mathematics, 1998

Alulírott Horlik Zalán Zoltán nyilatkozom, hogy szakdolgozatom elkészítése során az alább felsorolt feladatok elvégzésére a megadott MI alapú eszközöket alkalmaztam:

Feladat	Felhasznált eszköz	Felhasználás helye	Megjegyzés
LaTeX kód generálása	GPT-4o	main.tex fájl	Szükséges csomagok impotálása, kezdetleges környezet megteremtése
Dolgozatvázlat készítése	GPT-4o	Részlegesen a dolgozat felépítése során	Az AI által javasolt felépítést csak részben tartottam meg

A felsoroltakon túl más MI alapú eszközt nem használtam.