Numerikus modellezés és közönséges differenciálegyenletek numerikus megoldási módszerei

Faragó István

2013.02.15.

Tartalomjegyzék

1. Bevezetés **4**

2. Közönséges differenciálegyenletek kezdetiérték-feladata **6**
2.1. Elméleti összefoglaló 6
2.1.1. Közönséges differenciálegyenletek kezdetiérték-feladata és megoldhatósága 6
2.1.2. Differenciálegyenletek stabilitása 10
2.1.3. Merev feladatok 19
2.2. Időfüggő parciális differenciálegyenletek és modellezésük közönséges differenciálegyenlet-rendszerrel 25
2.2.1. A parciális differenciálegyenletek alapfogalmai 25
2.2.2. Megmaradási törvények 29
2.2.3. A korlátos térbeli tartományon kitűzött hővezetési egyenlet és szemidiszkretizációjának tulajdonságai 45

3. Bevezetés a kezdetiérték-feladatok egylépéses numerikus módszereibe **64**
3.1. Bevezetés az egylépéses módszerekbe 64
3.1.1. Taylor-sorba fejtéses módszer 65
3.1.2. Néhány nevezetes egylépéses módszer 71
3.1.3. Az explicit Euler-módszer 72
3.1.4. Az implicit Euler-módszer 76
3.1.5. A trapéz-módszer 77
3.1.6. Az egylépéses módszerek konvergenciája 78
3.2. Az egylépéses módszerek általános alakjának vizsgálata 84
3.2.1. Egylépéses módszerek operátoros alakban 85
3.2.2. Egylépéses módszerek konzisztenciája 89
3.2.3. Egylépéses módszerek konvergenciája és 0-stabilitása 90
3.2.4. Egylépéses módszerek alaptétele és alkalmazása az explicit Euler-módszerre 92
3.3. Abszolút stabilitás 96
3.3.1. Az explicit Euler-módszer abszolút stabilitási tartománya 96

3.3.2. További módszerek abszolút stabilitási tartománya 104
3.3.3. Az A-stabilitás fogalma 108

4. Runge-Kutta típusú módszerek kezdetiérték-feladatok megoldására 112
4.1. A Runge-Kutta típusú módszerek alapjai 112
4.1.1. További másodrendű módszerek 113
4.2. Explicit, magasabb rendű módszerek 117
4.2.1. A Runge-Kutta típusú módszerek rendje 122
4.2.2. Az explicit Runge-Kutta típusú módszerek konvergenciája 131
4.2.3. A Runge-Kutta típusú módszerek lépéshosszának megválasztása. Hibaanalízis. 133
4.2.4. Az explicit Runge-Kutta típusú módszerek abszolút stabilitása . . 138
4.3. Az implicit Runge-Kutta típusú módszerek 141
4.3.1. Bevezetés az implicit Runge-Kutta típusú módszerekbe 142
4.3.2. A kollokációs módszeren alapuló implicit Runge-Kutta típusú módszerek 147
4.3.3. Egylépéses módszerek a tesztfeladaton 152
4.3.4. Az implicit Runge-Kutta típusú módszerek abszolút stabilitása . . 156

5. Többlépéses módszerek 161
5.1. Adams-típusú módszerek 163
5.2. A retrográd differencia módszerek 170
5.3. A kezdeti közelítések megválasztása 173
5.4. Általános alakú lineáris többlépéses módszerek rendje 174
5.5. Az általános alakú lineáris többlépéses módszerek konvergenciája 179
5.5.1. Bevezetés a lineáris többlépéses módszerek stabilitásába 180
5.5.2. A lineáris többlépéses módszerek konvergenciája 181

6. Közönséges differenciálegyenletek kétpontos peremérték-feladata 198
6.1. Bevezetés, motiváció 198
6.2. Közönséges differenciálegyenletek peremérték-feladatának megoldhatósága 200
6.3. A lineáris peremérték-feladat megoldhatósága 202

7. Közönséges differenciálegyenletek kétpontos peremérték-feladatának numerikus megoldása 206
7.1. Peremérték-feladat numerikus megoldása Cauchy-feladatra való visszavezetéssel 206
7.1.1. A belövéses módszer 207
7.1.2. Lineáris peremérték-feladatok numerikus megoldása 211
7.2. A peremérték-feladat numerikus megoldása véges differenciák módszerével 213
7.2.1. Véges differenciás approximáció 214

7.2.2. Lineáris peremérték-feladatok approximációja véges differenciák módszerével 216
7.3. A lineáris peremérték-feladat numerikus megoldásának általános vizsgálata 223

8. Közönséges differenciálegyenletek numerikus megoldása Matlab programmal 232
8.1. A Matlab alapjai és a kezdetiérték feladatok megoldása 232
8.2. A peremérték-feladatok megoldása Matlab segítségével 242
8.2.1. A modellfeladat: stacionárius hőeloszlás homogén vezetékben . . . 243
8.2.2. A tesztfeladat numerikus megoldása Matlab segítségével 244

1. fejezet

Bevezetés

A differenciálegyenletek gyakori eszközei a természettudományos, műszaki, közgazdasági folyamatok leírásának, azaz a folytonos matematikai modelleket többnyire ezek segítségével lehetséges (és szokásos) leírni. Az ilyen modellek vizsgálatával a közönséges (illetve parciális) differenciálegyenletek elmélete foglalkozik. Ezek a vizsgálatok elsősorban a különböző jellegű feladatok megoldhatóságával foglalkoznak, elsősorban azt vizsgáljuk, hogy a kitűzött feladat milyen feltételek mellett lesz korrekt kitűzésű. A megoldás konkrét előállítása zárt alakban (azaz megadása olyan képletek segítségével, amelyek ismert és könnyen kiértékelhető függvényeket tartalmaznak) csak ritkán lehetséges. Ezért gyakorlati szempontból megkerülhetetlen az a megközelítés, aminek során a megoldást valamilyen *numerikus módszer* segítségével *közelítő alakban* keressük. Mint látni fogjuk, ezek a módszerek lehetővé teszik a numerikus megoldás nagy pontosságú és megbízható előállítását. Ez utóbbi azt jelenti, hogy becslést tudunk adni az ismeretlen pontos megoldás és az alkalmazott numerikus módszerrel nyert numerikus megoldás közötti eltérésre.

A könyv alapvető célja bevezetni az Olvasót a közönséges differenciálegyenletek numerikus módszereinek alapjaiba, illetve ismertetni azokat az eljárásokat, amelyeket a programcsomagok is használnak. Ezen a területen érezhetően intenzív a fejlődés, amelyről a megjelent cikkek és monográfiák viszonylag nagy száma is árulkodik.

A könyv mindazoknak készült, akik valamilyen szinten a numerikus módszereket szeretnék alkalmazni, és nem érik be azzal, hogy egy programcsomag, általuk nem ismert módon kiszámolt eredményét használják, hanem a dolgok hátterét és megbízhatóságát is szeretnék látni. A témához kapcsolódó szükséges numerikus előismeretek nagy részét ismertetjük, és a további ismeretek egyéb ismeretek megszerzésére javasoljuk az általános numerikus analízis kérdéseibe bevezető és elektronikusan hozzáférhető [5] könyvet.

A könyv alapvetően az Eötvös Loránd Tudományegyetem alkalmazott matematikus mesterszakon tartott kurzusaimon alapul. Ezért köszönettel tartozom hallgatóimnak, akik az előadások során kérdéseikkel, észrevételeikkel az egyes részek újragondolására serkentettek, és a könyv több részén a közös gondolkodásunk eredménye található meg. Nagyon hálás vagyok bírálómnak, Mincsovics Miklósnak, aki lelkiismeretesen átnézve a

kéziratot számos javító, a könyv lényegét jobban kiemelő észrevételt tett. Köszönettel tartozom Csörgő Gábornak, aki a könyv elkészítésének számos technikai részét magára vállalta. A legnagyobb köszönet azonban családomat illeti, akik elnézték a könyv írása miatti elfoglaltságomat.

Budapest, 2013 február.

Faragó István

2. fejezet

Közönséges differenciálegyenletek kezdetiérték-feladata

Ebben a fejezetben megismerkedünk a numerikus módszereink tárgyát képező közönséges differenciálegyenletekkel, illetve kezdetiérték-feladataival, és ezek elméleti összefoglalásával foglalkozunk. Megvizsgáljuk, hogy milyen feltételek mellett létezik egyértelmű megoldás. A fejezet jelentős részét képezi a parciális differenciálegyenletekkel való kapcsolat vizsgálata. Megmutatjuk, hogy az ún. szemidiszkretizáció segítségével a parciális differenciálegyenletek (amelyek jobban leírják a vizsgált jelenséget), szintén jól kezelhetők közönséges differenciálegyenletekkel. Foglalkozunk a speciális tulajdonságú rendszerek (az ún. merev rendszerek) tulajdonságaival is.

2.1. Elméleti összefoglaló

Először a kezdetiérték-feladatnak a megfogalmazásával és megoldhatóságával foglalkozunk, külön kitérve a numerikus szempontból fontos szerepet játszó stabilitásra is.

2.1.1. Közönséges differenciálegyenletek kezdetiérték-feladata és megoldhatósága

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

$$\frac{d\boldsymbol{u}(\cdot)}{dt} = \boldsymbol{f}(\cdot, \boldsymbol{u}), \quad \boldsymbol{u}(t_0) = \boldsymbol{u}_0 \tag{2.1}$$

feladatot kezdetiérték-feladatnak*, avagy más szóval* Cauchy-feladatnak *nevezzük.*

A könnyebb áttekinthetőség kedvéért írjuk ki a (2.1) feladatot koordinátánként! Jelölje $u_i(\cdot)$ az ismeretlen $\mathbf{u}(\cdot)$ vektorértékű függvény i-edik koordináta-függvényét, $f_i : G \to \mathbb{R}$ az $\mathbf{f}$ és u_{0i} $(i = 1, 2, \ldots, d)$ pedig az $\mathbf{u}_0$ vektor koordinátáit. Ekkor a Cauchy-feladat felírható a következő ún. koordinátánkénti alakban:

$$\begin{aligned} \frac{du_i(\cdot)}{dt} &= f_i(\cdot, u_1, \ldots u_d), \\ u_i(t_0) &= u_{0i} \end{aligned} \tag{2.2}$$

ahol $i = 1, 2, \ldots, d$.

Egy Cauchy-feladat megoldása azt jelenti, hogy meghatározzuk az összes olyan $\mathbf{u} : \mathbb{R} \to \mathbb{R}^d$ függvényt, amely valamely $I \subset \mathbb{R}$ intervallum pontjaiban egyrészt behelyettesíthető a (2.1) feladatba, másrészt pedig azt ki is elégíti.

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

- $\{(t, \boldsymbol{u}(t)) : \quad t \in I\} \subset G$ *és* $t_0 \in I$*;*
- $\dfrac{d\boldsymbol{u}(t)}{dt} = \boldsymbol{f}(t, \boldsymbol{u}(t)), \quad$ *minden* $\quad t \in I,$
- $\boldsymbol{u}(t_0) = \boldsymbol{u}_0$

a (2.1) Cauchy-feladat megoldásának nevezzük.

Amikor a (2.1) Cauchy-feladat egy természettudományos, műszaki, közgazdasági folyamat matematikai modellje, akkor alapvető követelmény, hogy létezzen egyértelmű megoldása. Ennek biztosítására vezessük be a $H(t_0, \mathbf{u}_0) = \{(t, \mathbf{u}) : |t - t_0| \le \alpha, \|\mathbf{u} - \mathbf{u}_0\| \le \beta\} \subset G$ jelölést. (Tehát $H(t_0, \mathbf{u}_0)$ egy $(t_0, \mathbf{u}_0)$ közepű, zárt, $d + 1$-dimenziós téglalap.) Mivel $\mathbf{f}$ folytonos a zárt $H(t_0, \mathbf{u}_0)$ halmazon, ezért értelmes az $M = \max_{H(t_0, \mathbf{u}_0)} \|\mathbf{f}(t, \mathbf{u})\|$ valós szám bevezetése. Ekkor minden olyan t esetén, amelyre $|t - t_0| \le \min\{\alpha, \beta/M\}$, a (2.1) Cauchy-feladatnak létezik $\mathbf{u}(t)$ megoldása. Ha emellett a $H(t_0, \mathbf{u}_0)$ halmazon az $\mathbf{f}$ függvény *a második változójában lipschitzes*, azaz valamely $L > 0$ állandó mellett minden $(t, \mathbf{u}_1), (t, \mathbf{u}_2) \in H(t_0, \mathbf{u}_0)$ pontban teljesül a

$$\|\mathbf{f}(t, \mathbf{u}_1) - \mathbf{f}(t, \mathbf{u}_2)\| \le L\|\mathbf{u}_1 - \mathbf{u}_2\| \tag{2.3}$$

ún. *Lipschitz-féle feltétel*, akkor ez a megoldás egyértelmű is.

A továbbiakban a (2.1) feladatra feltesszük, hogy létezik olyan $H(t_0, \mathbf{u}_0) \subset G$ részhalmaz, amelyen $\mathbf{f}$ folytonos és a második változójában lipschitzes, azaz létezik egyértelmű megoldása az $I_0 := \{t \in I : |t - t_0| \le T\}$ intervallumon, ahol $T = \min\{\alpha, \beta/M\}$.

Mivel a t változó az időt jelöli, ezért egy Cauchy-feladat megoldása azt írja le, hogy a rendszer időben hogyan változik. Mi a gyakorlati problémák vizsgálata során általában arra vagyunk kíváncsiak, hogy időben hogyan fejlődik a rendszer. Ez azt jelenti, hogy ha ismerjük a rendszer állapotát egy rögzített időpontban, akkor abból hogyan fejlődik tovább, azaz az $\mathbf{u}(t)$ függvényt a $t > t_0$ értékekre szeretnénk meghatározni. A $t = t_0$ időpontot *kezdőpontnak*, a megoldásfüggvény ezen pontbeli értékét pedig *kezdeti feltételnek* nevezzük. Nyilvánvalóan nem jelent megszorítást, ha a kezdőpontot $t_0 = 0$ értéknek vesszük. Így a (2.1) feladat megoldásának értelmezési tartománya a $[0, T] \subset I$ intervallum, és ekkor feladatunk a következő alakot ölti:

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{f}(t, \mathbf{u}(t)), \quad t \in [0, T], \tag{2.4}$$

$$\mathbf{u}(0) = \mathbf{u}_0. \tag{2.5}$$

Célunk a továbbiakban ezen $\mathbf{u}(t)$ függvény meghatározása.

2.3. Megjegyzés Az $\mathbf{f}$ folytonossága esetén, (azaz $\mathbf{f} \in C(H)$ mellett) a Cauchy-feladat $\mathbf{u}(t)$ megoldása egyszeresen folytonosan differenciálható, tehát $\mathbf{u} \in C^1[0, T]$. Ugyanakkor, ha $\mathbf{f}$ magasabb rendben sima, akkor a megoldás is simábbá válik: ha $\mathbf{f} \in C^p(H)$, akkor $\mathbf{u} \in C^{p+1}[0, T]$, ahol $p \in \mathbb{N}$. Így $\mathbf{f}$ megfelelő simaságával a megoldás szükséges simasága mindig biztosítható. Ezért tehát nem jelent lényeges megszorítást, ha a továbbiakban – ahol ez szükséges – feltesszük, hogy a megoldás *megfelelően sima*.

A numerikus módszereket – a könnyebb áttekinthetőség kedvéért – a skaláris egyenletekre fogalmazzuk meg, azaz a $d = 1$ esetet tekintjük. Legyen $Q_T := [0, T] \times \mathbb{R} \subset \mathbb{R}^2$, $f : Q_T \to \mathbb{R}$. A továbbiakban a

$$\frac{du}{dt} = f(t, u), \quad u(0) = u_0 \tag{2.6}$$

feladatot nevezzük Cauchy-feladatnak, ahol mindvégig feltesszük, hogy $f \in C(Q_T)$ és a második változójában lipschitzes függvény, azaz

$$|f(t, u_1) - f(t, u_2)| \leq L\,|u_1 - u_2|\,, \quad \forall (t, u_1), (t, u_2) \in Q_T, \tag{2.7}$$

továbbá $u_0 \in \mathbb{R}$ adott szám. Tehát feladatunk a következő: keressük azon megfelelően sima $u : [0, T] \to \mathbb{R}$ függvényt, amelyre

$$\frac{du(t)}{dt} = f(t, u(t)), \quad \forall t \in [0, T], \quad u(0) = u_0. \tag{2.8}$$

2.4. Megjegyzés Felmerülhet a kérdés: van-e kapcsolat valamely $g : \mathbb{R}^2 \to \mathbb{R}$ függvény folytonossága és a második változójában való lipschitzessége között? A válasz nemleges,

ugyanis, mint azt a következő két példa is mutatja, ezek egymástól független feltételek. Legyen először $g(x,y) = y^2$. Ez a függvény nyilván folytonos a $G = \mathbb{R}^2$ síkon, de nem lipschitzes, ugyanis

$$|g(x,y_1) - g(x,y_2)| = \left|y_1^2 - y_2^2\right| = |y_1 + y_2|\,|y_1 - y_2|\,,$$

és így a (2.7) feltétel nem teljesülhet, hiszen y_1 és y_2 tetszőlegessége miatt $|y_1 + y_2|$ nem lehet felülről korlátos valamely L állandóval.
Legyen most $g(x,y) = D(x)y$, ahol $D(x)$ a jól ismert Dirichlet-függvény[1]. Ekkor g sehol sem folytonos, viszont

$$|g(x,y_1) - g(x,y_2)| = |D(x)|\,|y_1 - y_2| \le |y_1 - y_2|\,,$$

azaz $L = 1$ értékkel a (2.7) összefüggés érvényes a $G = \mathbb{R}^2$ síkon.

2.5. Megjegyzés Hogyan biztosítható a lipschitzesség? Tegyük fel, hogy valamely $g : \mathbb{R}^2 \to \mathbb{R}$ függvény az értelmezési tartományának valamely nyílt H_g részhalmazán a második változójában korlátos deriválttal rendelkezik. Ekkor a Lagrange-középértéktétel értelmében valamely $\tilde{y} \in (y_1, y_2)$ érték mellett $g(x,y_1) - g(x,y_2) = \partial_2 g(x,\tilde{y})(y_1 - y_2)$, azaz a (2.7) feltétel teljesül az $L = sup_{H_g}\left(|\partial_2 g(x,y)|\right) < \infty$ állandóval.

Ezen megjegyzés következménye: ha a (2.8) Cauchy-feladat f függvénye a Q_T halmazon folytonos, és a második változójában korlátos parciális deriválttal rendelkezik, akkor létezik egyértelmű megoldása a $[0,T]$ intervallumon.

2.6. Megjegyzés Rendszerek esetén hasonló állítás nyerhető. Legyen ugyanis $\mathbf{g} : \mathbb{R}^{d+1} \to \mathbb{R}^d$ egy megfelelő simaságú függvény. Ekkor

$$\mathbf{g}(x,\mathbf{y}_1) - \mathbf{g}(x,\mathbf{y}_2) = \mathbf{J}(x,\tilde{\mathbf{y}})(\mathbf{y}_1 - \mathbf{y}_2),$$

ahol $\tilde{\mathbf{y}} \in \mathbb{R}^d$ egy vektor, amelynek koordinátáira $\tilde{\mathbf{y}}_i \in ((\mathbf{y}_1)_i, (\mathbf{y}_2)_i)$, $\mathbf{J}(x,\tilde{\mathbf{y}}) \in \mathbb{R}^{d\times d}$ a $\mathbf{g}$ függvény Jacobi-mátrixa az $(x,\tilde{\mathbf{y}})$ pontban, azaz $\mathbf{J}(x,\tilde{\mathbf{y}})_{i,j} = \dfrac{\partial g_i}{\partial \mathbf{y}_j}(x,\tilde{\mathbf{y}}_i)$. Ezért ha a (2.5) feladatban a $G \subset \mathbb{R}\times\mathbb{R}^d$ halmazon az $\mathbf{f}(t,\mathbf{u})$ függvény folytonos, az $\mathbf{u}$ változójában folytonosan differenciálható, valamint

$$L := \sup_{(t,\mathbf{u})\in G} \left\| \frac{\partial \mathbf{f}}{\partial \mathbf{u}}(t,\mathbf{u}) \right\| < \infty,$$

akkor létezik egyértelmű megoldása.

[1] A Dirichlet-függvény definíciója: $D(x) = 1$, ha x racionális, és $D(x) = 0$, ha x irracionális. Ez a függvény minden pontban szakad.

2.7. Megjegyzés Ha a (2.5) feladatban $\mathbf{f}(t, \mathbf{u})$ függvény nem függ a $t \in \mathbb{R}$ változótól, azaz $\mathbf{f} = \mathbf{f}(\mathbf{u}) : \mathbb{R}^d \to \mathbb{R}^d$, akkor a Cauchy-feladatot *autonóm rendszernek* nevezzük. Megjegyezzük, hogy egy tetszőleges nem autonóm rendszer mindig átírható egy egyel több ismeretlent tartalmazó autonóm rendszerré. Ugyanis bevezetve az $\mathbf{u}_{d+1} = t$ újabb változót, a (2.2) alakú koordinátánkénti egyenletek az alábbi módon írhatók fel:

$$\begin{aligned} \frac{du_i(\cdot)}{dt} &= f_i(u_{d+1}, u_1, \dots u_d), \\ u_i(t_0) &= u_{0i}, \\ \frac{du_{d+1}(\cdot)}{dt} &= 1, \quad u_{d+1}(t_0) = t_0, \end{aligned} \tag{2.9}$$

ahol $i = 1, 2, \dots, d$. Megjegyezzük továbbá, hogy a magasabb rendű egyenletek az ún. átviteli elv segítségével átírhatók (a 2.5) alakú elsőrendű rendszerré.

2.1.2. Differenciálegyenletek stabilitása

Könyvünk célja *valós feladatok* matematikai és numerikus modellezésének vizsgálata. Ezért megvizsgáljuk, hogy ha egy kezdetiérték-feladatban a kezdeti feltétel hibával terhelt (ami a mérések, a számítógépes számábrázolás, és számos más tényező miatt szükségszerűen előfordul), akkor ez az eltérés hogyan hat ki az eredeti ("pontos") megoldásra, azaz mennyire tér el attól az időbeli haladás esetén? Megvizsgáljuk, hogy milyen esetben garantálható a kezdeti feltétel megfelelő pontosságának biztosításával, hogy egy előre megadott értéknél kisebb mértékben térjen csak el egymástól a két megoldás. Mivel ezen kérdéskör vizsgálata nem tartozik szorosan a könyv tárgyához, ezért a továbbiakban csak néhány, a numerikus módszereknél később előforduló alapfogalmat emelünk ki, a részletes elméleti vizsgálatoktól eltekintünk. (Az érdeklődőknek a magyar nyelvű irodalomból ajánljuk a [17] könyvet, illetve a [15] elektronikus jegyzetet.)

A stabilitáselméleti alapfogalmak szemléltetésére tekintsük meg először az alábbi egyszerű fizikai példát. Képzeljünk el egy golyót, valamint egy $\delta > 0$ mélységű gödröt, illetve dombot.

A golyó kezdeti elhelyezkedésére vonatkozóan három esetet vizsgálunk meg. Nevezetesen, a golyó

1. a gödör alján helyezkedik el,
2. a domb tetején található,
3. egy vízszintes sík felületen van.

Mindhárom helyzetben egyensúlyban van a golyó, azaz, ha nem mozdítjuk meg, helyben marad. Azonban ha kicsit elmozdítjuk, majd elengedjük, akkor mindhárom esetben

más történik. (Ez azt jelenti, hogy a mozgását leíró kezdetiérték-feladatban a kezdeti feltételt megváltoztatjuk, más szóval, azt *perturbáljuk*.) A rendszer viselkedése az egyes esetekben nyilvánvalóan a következőek.

1. Az első esetben a golyó visszagurul a gödör aljára, ha az elmozdított kezdeti állapotának magassága δ értékénél nem jobban tér el a gödör aljától. (Azaz, nem vettük ki a gödörből.)

2. A második esetben a golyó legurul a domboldalon, egyre jobban eltávolodik az eredeti kezdeti helyzetétől.

3. A harmadik esetben a golyó ott marad az elmozdítás helyén, azaz nem tér vissza az eredeti helyzetébe, de nem is távolodik el onnan.

Tehát az első esetben a kezdeti feltétel -bizonyos határokon belüli- perturbációjára érzéketlen a megoldás, azaz, a pontos, (”perturbálatlan”) megoldás és ”perturbált” megoldás távolsága korlátos marad, sőt, időben oda mindig vissza is tér, azaz a távolságuk nullához tart. A második esetben ez nem áll fenn: a két állapot közötti távolság kinő, és kellően nagy idő elteltével tetszőleges nagy lesz. A harmadik esetben, az első esethez hasonlóan, korlátos marad a megoldások különbsége, viszont a távolság nullához való csökkenése ebben az esetben nem teljesül.

Tekintsük most az

$$u'(t) = \lambda u(t), \quad t > 0, \tag{2.10}$$
$$u(0) = u_0, \tag{2.11}$$

skaláris kezdetiérték-feladatot, ahol $\lambda \in \mathbb{C}$ adott komplex szám. Ennek megoldása az $u(t) = e^{\lambda t}u_0$ függvény. Ezért, ha a (2.11) helyett az $u(0) = \tilde{u}_0$ kezdeti feltételt adjuk meg a (2.10) egyenletre, akkor a perturbált egyenlet megoldása $\tilde{u}(t) = e^{\lambda t}\tilde{u}_0$, és így a két megoldás eltérésére az

$$u(t) - \tilde{u}(t) = e^{\lambda t}(u_0 - \tilde{u}_0)$$

összefüggést kapjuk. Ezért

$$|u(t) - \tilde{u}(t)| = |e^{\lambda t}||(u_0 - \tilde{u}_0)| = e^{(\operatorname{Re}\lambda)t}|(u_0 - \tilde{u}_0)|.$$

Ezért tehát $\operatorname{Re}\lambda \leq 0$ esetén a megoldások távolsága korlátos marad, és $\operatorname{Re}\lambda < 0$ esetén a távolság $t \to \infty$ esetén nullához tart. Viszont $\operatorname{Re}\lambda > 0$ esetén a kezdeti távolság növekszik, és $t \to \infty$ esetén a végtelenhez tart. Ez azt jelenti, hogy ha a (2.10)-(2.11) feladatot tekintjük a pontos feladatnak, akkor a kezdeti érték perturbációja csak $\operatorname{Re}\lambda \leq 0$ esetén okoz korlátos hibát, ellenkező esetben a kezdeti érték hibája tovább növekszik és kinő a végtelenbe.

Fogalmazzuk át a fenti eredményünket! Legyenek $u_1(t)$ és $u_2(t)$ a (2.10) egyenlet két tetszőleges megoldása. Ekkor

$$|u_1(t) - u_2(t)| = e^{(\operatorname{Re}\lambda)t}|(u_1(0) - u_2(0))|.$$

Tehát az egyenlet két tetszőleges megoldásának a távolsága csak $\operatorname{Re}\lambda \leq 0$ esetén marad korlátos.

A továbbiakban megfogalmazzuk a fenti tulajdonságot a (2.4)-(2.5) általános alakú

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{f}(t, \mathbf{u}(t)), \quad t > t_0 \tag{2.12}$$

$$\mathbf{u}(t_0) = \mathbf{u}_0. \tag{2.13}$$

kezdetiérték-feladatra. Jelölje a feladat megoldását $\mathbf{u}(t; t_0, \mathbf{u}_0)$, ahol $t \geq t_0$.[2]

2.8. Definíció *Azt mondjuk, hogy a (2.12)-(2.13) feladat $\boldsymbol{u}(t; t_0, \boldsymbol{u}_0)$ megoldása stabil, ha minden $\varepsilon > 0$ számhoz létezik olyan $\delta > 0$ szám, amelyre érvényes, hogy minden olyan $\overline{\boldsymbol{v}}_0 \in \mathbb{R}^d$ kezdeti vektorra, amelyre*

$$\|\boldsymbol{u}_0 - \overline{\boldsymbol{v}}_0\| < \delta, \tag{2.14}$$

teljesülnek a következők.

- *az $\boldsymbol{u}(t; t_0, \overline{\boldsymbol{v}}_0)$ megoldás létezik a $[t_0, \infty)$ intervallumon,*
- *minden $t > t_0$ esetén*

$$\|\boldsymbol{u}(t; t_0, \boldsymbol{u}_0) - \boldsymbol{u}(t; t_0, \overline{\boldsymbol{v}}_0)\| < \varepsilon. \tag{2.15}$$

2.9. Definíció *Azt mondjuk, hogy a (2.12)-(2.13) kezdetiérték-feladat $\boldsymbol{u}(t; t_0, \boldsymbol{u}_0)$ megoldása aszimptotikusan stabil, ha*

- *ez a megoldás stabil,*
- *érvényes a*

$$\lim_{t\to\infty} \|\boldsymbol{u}(t; t_0, \boldsymbol{u}_0) - \boldsymbol{u}(t; t_0, \overline{\boldsymbol{v}}_0)\| = 0 \tag{2.16}$$

egyenlőség.

[2] Ebben a részben több helyen is külön jelezzük a megoldás kezdeti feltételtől (azaz az $(t_0, \mathbf{u}_0)$ ponttól) való függését.

Ha egy megoldás nem stabil, akkor instabilnak nevezzük. A fenti stabilitási fogalmakat szokásos Ljapunov-féle stabilitásnak is nevezni.[3]

A definíciók alapján tehát a (2.10)-(2.11) kezdetiérték-feladat $u(t) = e^{\lambda t}u_0$ megoldása $\operatorname{Re}\lambda \leq 0$ esetén stabil, $\operatorname{Re}\lambda < 0$ esetén aszimptotikusan stabil, és $\operatorname{Re}\lambda > 0$ esetén pedig instabil.

Fontos eset a lineáris rendszerek esete, azaz amikor (a 2.12) egyenletben $\mathbf{f}(t, \mathbf{u}(t)) = \mathbf{A}(t)\mathbf{u}(t)$ alakú, ahol $\mathbf{A}(t)$ minden $t \in [t_0, \infty)$ esetén egy $\mathbb{R}^{d\times d}$ mátrix. Tehát feladatunk a

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{A}(t)\mathbf{u}(t), \quad t > t_0 \tag{2.17}$$
$$\mathbf{u}(t_0) = \mathbf{u}_0. \tag{2.18}$$

feladat megoldása. A stabilitás vizsgálatát kezdjük azzal a speciális esettel, amikor a mátrix nem függ t-től, azaz $\mathbf{A}(t) = \mathbf{A}$ minden $t \geq t_0$ esetén, azaz a

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{A}\mathbf{u}(t), \quad t > t_0 \tag{2.19}$$
$$\mathbf{u}(t_0) = \mathbf{u}_0 \tag{2.20}$$

kezdetiérték-feladatot vizsgáljuk. Ekkor a megoldás $\mathbf{u}(t) = \exp((t - t_0)\mathbf{A})\mathbf{u}_0$ alakú. Tegyük fel, hogy $\mathbf{A}$ diagonalizálható, azaz létezik olyan $\mathbf{T}$ reguláris mátrix, amellyel a hasonlósági transzformáció után

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \operatorname{diag}[\lambda_1, \ldots, \lambda_d] = \Lambda, \tag{2.21}$$

ahol λ_i az $\mathbf{A}$ mátrix sajátértékei. Bevezetve $\mathbf{w}(t) = \mathbf{T}^{-1}\mathbf{u}(t)$ új változót, a (2.19) egyenlet átírható

$$\frac{d\mathbf{w}(t)}{dt} = \Lambda\mathbf{w}(t), \quad t > t_0$$

alakra. Ennek megoldása a $\mathbf{w}(t) = \exp((t - t_0)\Lambda)\mathbf{w}(t_0)$ $[t_0, \infty) \to \mathbb{R}^d$ függvény, ahol $\mathbf{w}(t_0) = \mathbf{T}^{-1}\mathbf{u}_0$ a kezdeti feltételből ismert vektor. Mivel egy diagonális mátrix exponenciálisa egy olyan, szintén diagonális mátrix, amelynek diagonálisában az eredeti mátrix főátlójának exponenciálisai állnak, ezért a $\mathbf{w}(t)$ vektorfüggvény i-dik koordinátafüggvénye $\mathbf{w}_i(t) = e^{(t-t_0)\lambda_i}\left(\mathbf{w}(t_0)\right)_i$ minden $i = 1, 2, \ldots, d$ esetén. Ez alapján közvetlenül meghatározhatjuk a (2.19)-(2.20) kezdetiérték-feladat stabilitásának feltételeit.

2.10. Tétel (Állandó együtthatós lineáris rendszer stabilitása.) *Tegyük fel, hogy $\boldsymbol{A}$ diagonalizálható mátrix. Ekkor a (2.19)-(2.20) kezdetiérték-feladat $\boldsymbol{u}(t) = \exp((t - t_0)\boldsymbol{A})\boldsymbol{u}_0$ megoldása*

[3] Ljapunov, Alexander Michajlovich (1857-1918) orosz matematikus és fizikus. Legjelentősebb eredményei a dinamikai rendszerek stabilitásához, a matematikai fizikához és a valószínűségszámításhoz fűződnek.

- *pontosan akkor stabil, amikor az* $\boldsymbol{A}$ *mátrix mindegyik sajátértékének valós része nem pozitív, azaz* $\operatorname{Re}\lambda_i \leq 0$ *minden* $i = 1, 2, \ldots d$ *esetén;*
- *pontosan akkor aszimptotikusan stabil, amikor az* $\boldsymbol{A}$ *mátrix mindegyik sajátértékének valós része negatív, azaz* $\operatorname{Re}\lambda_i < 0$ *minden* $i = 1, 2, \ldots d$ *esetén;*
- *minden egyéb esetben instabil, azaz az* $\boldsymbol{A}$ *mátrixnak létezik pozitív valós részű sajátértéke, vagyis létezik olyan* $i \in \{1, 2, \ldots, d\}$ *index, amelyre* $\operatorname{Re}\lambda_i > 0$.

2.11. Megjegyzés A fenti állítás kiterjeszthető tetszőleges $\mathbf{A} \in \mathbb{R}^{d\times d}$ mátrix esetére is. Ekkor a stabilitást a sajátértékek valós részének előjele és a hozzájuk tartozó Jordan-blokkok mérete alapján lehet megadni. Nevezetesen, egy állandó együtthatós lineáris rendszer pontosan akkor stabil, amikor az $\mathbf{A}$ mátrix λ_i sajátértékeire

1. $\operatorname{Re}\lambda_i \leq 0$ minden i-re;
2. ha $\operatorname{Re}\lambda_i = 0$, akkor λ_i egyszeres sajátérték,

és akkor aszimptotikusan stabil, amikor λ_i sajátértékeire $\operatorname{Re}\lambda_i < 0$ minden i-re. Az egyéb esetekben a feladat instabil.

2.12. Példa *Egyszerű eset a rezgő húr egyenletének, azaz az* $-u''(t) + u(t) = 0$ *egyenlet stabilitásának vizsgálata. Bevezetve a* $w_1(t) = u(t)$ *és a* $w_2(t) = u'(t)$ *jelöléseket, a* $\boldsymbol{w}(t) = [w_1(t), w_2(t)]^\top$ *vektorfüggvényre az egyenlet felírható* $\boldsymbol{w}'(t) = \boldsymbol{A}\boldsymbol{w}(t)$ *alakban, ahol*

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Mivel $\boldsymbol{A}$ *sajátértékei* $\lambda_{1,2} = \pm 1$*, ezért ezen feladat bármely kezdeti feltétel melletti megoldása instabil.*

Térjünk át a (2.17)-(2.18) változó együtthatós (nem-autonóm) rendszer kezdetiérték-feladat stabilitásának vizsgálatára! Jelölje $\mathbf{Y}(t) \in \mathbb{R}^{d\times d}$ a (2.17) feladat alapmátrixát, azaz legyen megoldása a

$$\frac{d\mathbf{Y}(t)}{dt} = \mathbf{A}(t)\mathbf{Y}(t), \quad t > t_0 \tag{2.22}$$

$$\mathbf{Y}(t_0) = \mathbf{I} \tag{2.23}$$

kezdetiérték-feladatnak.[4] Ekkor a megoldás $\mathbf{u}(t) = \mathbf{Y}(t - t_0)\mathbf{u}_0$ alakú. A stabilitás vizsgálatához a két különböző kezdeti feltételhez tartozó megoldás különbségének időbeli viselkedése szükséges. Jelölje az adott $\mathbf{u}_1(t_0)$ és $\mathbf{u}_2(t_0)$ kezdeti értékekhez tartozó megoldásokat $\mathbf{u}_1(t)$ és $\mathbf{u}_2(t)$! Ekkor

$$\|\mathbf{u}_1(t) - \mathbf{u}_2(t)\| = \|\mathbf{Y}(t)\left(\mathbf{u}_1(0) - \mathbf{u}_2(0)\right)\| \leq \|\mathbf{Y}(t)\|\|\mathbf{u}_1(0) - \mathbf{u}_2(0)\|. \tag{2.24}$$

[4] $\mathbf{I} \in \mathbb{R}^{d\times d}$ az egységmátrixot jelöli.

Vezessük be a

$$\mathcal{K} = \sup_{t\in[t_0,\infty)} \|\mathbf{Y}(t)\| \tag{2.25}$$

jelölést! Ekkor a stabilitási fogalmaink alapján nyilvánvalóan érvényes a következő állítás.

2.13. Tétel (Változó együtthatós lineáris rendszer stabilitása.) *Tekintsük a (2.22)-(2.23) kezdetiérték-feladatot! Az* $\boldsymbol{u}(t) = \boldsymbol{Y}(t-t_0)\boldsymbol{u}_0$ *megoldás*

- *pontosan akkor stabil, amikor* $\mathcal{K} < \infty$*;*
- *pontosan akkor aszimptotikusan stabil, amikor* $\lim_{t\to\infty} \|\boldsymbol{Y}(t)\| = 0$*;*
- $\mathcal{K} = \infty$ *esetén instabil.*

2.14. Megjegyzés Vegyük észre, hogy a változó együtthatós rendszer stabilitása közvetlenül nem függ az $\mathbf{A}(t)$ mátrix sajátértékeitől. (Tehát nem arról van szó, hogy az $\mathbf{A}(t)$ mátrix $\lambda_i(t)$ sajátértékeire kell teljesülnie az állandó együtthatós lineáris rendszer stabilitási feltételeinek, tetszőleges $t > t_0$ esetén!) Tekintsük például az $u'(t) = (\cos(t))u(t)$ $(t > 0)$ skaláris egyenletet, ahol tehát $\mathbf{A}(t) = \cos(t) \in \mathbb{R}^{1\times 1}$. Ekkor természetesen az $\mathbf{A}$ "mátrix" egyetlen sajátértéke $\lambda_1(t) = \cos(t)$, míg $\mathbf{Y}(t) = e^{\sin(t)}$.[5] Mivel

$$\mathcal{K} = \sup_{t\in[t_0,\infty)} \|Y(t)\| = \sup_{t\in[t_0,\infty)} |e^{\sin(t)}| = e < \infty,$$

ezért a megoldás stabil. Ugyanakkor a $\lambda_1(t) = \cos(t)$ sajátérték pozitív értékeket is felvesz.

Tekintsük a (2.17)-(2.18) változó együtthatós rendszer kezdetiérték-feladatban az általánosabb alakú *inhomogén egyenletet*, azaz valamely adott $\mathbf{f}(t) \in \mathbb{R}^{d\times d}$ függvény mellett vizsgáljuk a

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{A}(t)\mathbf{u}(t) + \mathbf{f}(t), \quad t > t_0 \tag{2.26}$$

$$\mathbf{u}(t_0) = \mathbf{u}_0 \tag{2.27}$$

kezdetiérték-feladatt. Ennek megoldása az

$$\mathbf{u}(t) = \mathbf{Y}(t-t_0)\left(\mathbf{u}_0 + \int_{t_0}^{t} \mathbf{Y}^{-1}(s)\mathbf{f}(s)ds\right)$$

[5]Ez könnyen ellenőrizhető, ugyanis $\left(e^{\sin(t)}\right)' = \cos(t)e^{\sin(t)}$, azaz jelöléseink mellett valóban $\mathbf{Y}'(t) = \mathbf{A}(t)\mathbf{Y}(t)$. Emellett az $e^{\sin(t)}$ függvény a $t = 0$ helyen egyel egyenlő, azaz $\mathbf{Y}(0) = 1$.

függvény. Ezért a (2.26) egyenlet bármely két különböző kezdeti feltételhez tartozó megoldására érvényben marad a (2.24) becslés. Tehát egy inhomogén feladat megoldása pontosan akkor stabil (aszimptotikusan stabil, instabil), amikor a megfelelő homogén feladat megoldása stabil (aszimptotikusan stabil, instabil).

Mint láttuk, a lineáris elméletben a kezdetiérték-feladat valamely megoldásának stabilitása független volt a kezdeti feltétel konkrét megválasztásától, azaz az csak magától az egyenlettől függött. Tehát valójában a stabilitás nem a kezdetiérték-feladat megoldásától, hanem magától az egyenlettől függ. (Tehát nem lehetséges, hogy ha ugyanazon egyenletre van két, különböző kezdeti feltételekhez tartozó megoldásunk, akkor az egyik megoldás stabil, míg a másik instabil.) Ezért a lineáris esetben szokásos nem egy megoldás stabilitásáról, hanem *a differenciálegyenlet stabilitásáról* beszélni.

A továbbiakban röviden összefoglaljuk a nemlineáris feladatok stabilitásának elméletét. Felhasználva a 2.7. megjegyzést, elegendő a

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{f}(\mathbf{u}(t)), \quad t > t_0 \tag{2.28}$$

$$\mathbf{u}(t_0) = \mathbf{u}_0. \tag{2.29}$$

autonóm rendszer vizsgálata. Fontos megjegyeznünk, hogy az $\mathbf{u}(t) = \mathbf{u}^\star = \mathit{constans} \in \mathbb{R}^d$ időben nem változó megoldást *egyensúlyi pontnak* nevezzük. Nyilvánvalóan ezekre a megoldásokra $\mathbf{u}^{\star\prime} = 0$, azaz a (2.28) egyenlet alapján az egyensúlyi pontokat az $\mathbf{f}(\mathbf{u}) = 0$ algebrai egyenletrendszer megoldásaiból nyerjük.

2.15. Példa *Tekintsük*

$$u' = u(1-u) \tag{2.30}$$

autonóm egyenletet! Az $u(1-u) = 0$ *algebrai egyenlet megoldásaiból azt kapjuk, hogy két egyensúlyi pont van, nevezetesen az* $u_1^\star = 0$ *és az* $u_2^\star = 1$ *pontok. (Ez tehát azt jelenti, hogy ha a (2.30) egyenlethez az* $u(0) = 0$ *illetve az* $u(0) = 1$ *kezdeti feltételeket adjuk meg, akkor a két feladat megoldása az* $u_1^\star = 0$ *és az* $u_2^\star = 1$ *konstans függvények lesznek, azaz a kezdeti állapot időben nem változik meg.)*

2.16. Megjegyzés Az autonóm rendszerek egyensúlyi pontjai jól jellemzik a rendszer tetszőleges, időben változó megoldásait is. Nevezetesen, ha egy időben változó megoldás $t \to \infty$ esetén tart egy véges határértékhez, akkor ez a határérték feltétlenül a rendszer egyensúlyi pontja. Ugyanakkor, mivel a feladatnak egyetlen megoldása van, ezt az egyensúlyi pontot véges idő alatt nem érheti el. Összefoglalva, legyen $\mathbf{u}(t)$ a (2.28)-(2.29) feladat megoldása.

- Ha létezik $\lim_{t\to\infty} \mathbf{u}(t) = \mathbf{u}^\star$, akkor $\mathbf{u}^\star$ egyensúlyi pont.
- Ha valamely $t^\star$ pontban $\mathbf{u}(t^\star) = \mathbf{u}^\star$, akkor $\mathbf{u}(t) = \mathbf{u}(t^\star)$ minden $t \geq t_0$ esetén.

Számunkra igazából az egyensúlyi pontok stabilitása a fontos, azaz annak a kérdésnek a megvizsgálása, hogy ha valamely egyensúlyi pont környezetéből indul egy megoldás, akkor a perturbált megoldás eltávolodik-e az egyensúlyi helyzetétől, és ha igen, akkor hogyan.

2.17. Példa *Tekintsük a 2.15. példa (2.30) autonóm egyenletét! Mint az közvetlenül látható, hogy az $u(0) = 0$ kezdeti feltétel esetén a megoldás a $u(t) = 0$ egyensúlyi állapot, míg az $u(0) \neq 0$ esetén a megoldás az*

$$u(t) = \frac{1}{1 - \left(1 + \frac{1}{u(0)}\right) e^{-t}}$$

alakú függvény. Tetszőleges $u(0)$ esetén ezen függvényre $\lim_{t\to\infty} u(t) = 1$ és a konvergencia monoton. Tehát, az $u_2^\star = 1$ egyensúlyi megoldás stabil (a definícióban $\delta = \varepsilon$ megválasztással), sőt, aszimptotikusan is stabil. Ugyanakkor az $u_1^\star = 0$ egyensúlyi megoldás instabil, hiszen akármilyen közel is indítjuk a megoldást az $u(0) = 0$ értékhez, az az azonosan egy megoldáshoz tart, tehát eltávolodik az azonosan nulla megoldástól.

Mint az a 2.17. példából látszik, nemlineáris esetben már a stabilitás valóban egy adott megoldásra vonatkozó fogalom: ebben az esetben valóban előfordulhat, hogy egyes megoldások stabilak, míg mások instabilak. Vegyük észre továbbá, hogy homogén lineáris esetben az $\mathbf{u} = \mathbf{0}$ egyensúlyi pont, tehát elegendő az $\mathbf{u}(t) = \mathbf{0}$ megoldás stabilitását vizsgálni.

Hogyan vizsgálhatjuk meg általános esetben egy nemlineáris rendszer stabilitását / instabilitását? Tekintsük a (2.12)-(2.13) feladatot, amelynek a megoldása $\mathbf{u}(t)$. Legyen $\tilde{\mathbf{u}}(t)$ a (2.12) egyenlet egy másik kezdőpontból induló megoldása.

Fejtsük Taylor-sorba a (2.28) egyenlet jobb oldalát az $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}(t)$ helyen az $\mathbf{u} = \mathbf{u}(t)$ pont körül! Ekkor

$$\mathbf{f}(\tilde{\mathbf{u}}) = \mathbf{f}(\mathbf{u}) + \mathbf{f}'(\mathbf{u})(\tilde{\mathbf{u}} - \mathbf{u}) + \mathbf{r}(\tilde{\mathbf{u}} - \mathbf{u}), \tag{2.31}$$

ahol $\mathbf{f}'(\mathbf{u}) \in \mathbb{R}^{d\times d}$ az $\mathbf{f}$ függvény $\mathbf{u}$ helyen vett Jacobi-mátrixa, az $\mathbf{r}(\tilde{\mathbf{u}} - \mathbf{u})$ hibatagra pedig fennáll a

$$\lim_{\tilde{\mathbf{u}}\to\mathbf{u}} \frac{\mathbf{r}(\tilde{\mathbf{u}} - \mathbf{u})}{\|\tilde{\mathbf{u}} - \mathbf{u}\|} = \mathbf{0} \tag{2.32}$$

összefüggés.

A továbbiakban felhasználjuk az alábbi állítást.

Tekintsük az

$$\frac{d\mathbf{y}}{dt} = \mathbf{A}\mathbf{y} + \mathbf{g}(t, \mathbf{y}) \tag{2.33}$$

közönséges differenciálegyenlet-rendszert, ahol $\mathbf{A} \in \mathbb{R}^{d\times d}$ adott állandó mátrix, $\mathbf{g}(t,\mathbf{y})$ olyan folytonos függvény, amelyre a

$$\lim_{\mathbf{y}\to\mathbf{0}} \frac{\mathbf{g}(t,\mathbf{y})}{\|\mathbf{y}\|} = \mathbf{0} \tag{2.34}$$

határérték t-ben egyenletesen érvényes. Ekkor az állandó együtthatós

$$\frac{d\mathbf{y}}{dt} = \mathbf{A}\mathbf{y} \tag{2.35}$$

közönséges differenciálegyenlet-rendszert a (2.33) egyenlet *első közelítésének* nevezzük. Az alábbi állítás kapcsolatot teremt a (2.33) és a (2.35) feladatok stabilitása között.

2.18. Lemma *Ha az $\boldsymbol{A}$ mátrix mindegyik sajátértékének valós része negatív, akkor a (2.33) egyenlet $\boldsymbol{y} = \boldsymbol{0}$ megoldása aszimptotikusan stabil. Ha az $\boldsymbol{A}$ mátrixnak létezik pozitív valós részű sajátértéke, akkor az $\boldsymbol{y} = \boldsymbol{0}$ megoldása instabil.*

A fenti lemma kapcsán megjegyezzük a következőket.

- Nyilván a (2.34) feltétel következtében $\mathbf{g}(t,\mathbf{0}) = \mathbf{0}$, azaz $\mathbf{y} = 0$ tényleg (egyensúlyi) megoldása a (2.33) egyenletnek.
- Ha $\mathbf{g}(t,\mathbf{0}) = \mathbf{0}$ és $\dfrac{\partial \mathbf{g}}{\partial \mathbf{y}}(t,\mathbf{0}) = \mathbf{0}$, akkor a (2.34) feltétel automatikusan teljesül.
- A fenti a 2.18. lemma speciális esete az ún. Poincaré-Ljapunov tételnek, amely az

  $$\frac{d\mathbf{y}}{dt} = \mathbf{A}\mathbf{y} + \mathbf{B}(t)\mathbf{y} + \mathbf{g}(t,\mathbf{y})$$

  egyenletet vizsgálja a 2.18. lemma feltételei és a $\lim_{t\to\infty} \|\mathbf{B}(t)\| = 0$ feltétel mellett. A tétel állításai megegyeznek a 2.18. lemma állításával.
- Ha az $\mathbf{A}$ mátrix sajátértékei között van nulla valós részű (azaz tiszta képzetes) sajátérték is, és a többi sajátérték valós része negatív, akkor a megoldás egyaránt lehet (aszimptotikusan) stabil és instabil is, tehát ebben az esetben a (2.35) első közelítéssel nem lehet eldönteni a (2.33) egyenlet stabilitását.

Térjünk vissza a (2.28)-(2.29) nemlineáris autonóm rendszer $\mathbf{u}^\star$ egyensúlyi pont körüli stabilitásának vizsgálatára! Felhasználva a (2.31) sorfejtést, egy tetszőleges $\mathbf{u}(t)$ megoldás esetén az egyenlet jobb oldala felírható

$$\mathbf{f}(\mathbf{u}) = \mathbf{f}(\mathbf{u}^\star) + \mathbf{f}'(\mathbf{u}^\star)(\mathbf{u}-\mathbf{u}^\star) + \mathbf{r}(\tilde{\mathbf{u}}-\mathbf{u})$$

alakban. Mivel $\mathbf{u}^\star$ egyensúlyi pont, ezért $\mathbf{f}(\mathbf{u}^\star) = \mathbf{0}$. Így

$$\mathbf{f}(\mathbf{u}) = \mathbf{f}'(\mathbf{u}^\star)(\mathbf{u} - \mathbf{u}^\star) + \mathbf{r}(\tilde{\mathbf{u}} - \mathbf{u}). \tag{2.36}$$

Ezt figyelembe véve a (2.28)-(2.29) feladat az

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{f}'(\mathbf{u}^\star)(\mathbf{u} - \mathbf{u}^\star) + \mathbf{r}(\tilde{\mathbf{u}} - \mathbf{u}), \quad t > t_0 \tag{2.37}$$

$$\mathbf{u}(t_0) = \mathbf{u}_0 \tag{2.38}$$

alakot ölti. Vezessük be az $\mathbf{e}(t) = \mathbf{u}(t) - \mathbf{u}^\star$ függvényt! Mivel $\mathbf{u}^\star$ állandó, ezért a t szerinti deriváltja nulla, azaz a (2.37) egyenlet alapján érvényes a

$$\frac{d\mathbf{e}(t)}{dt} = \mathbf{f}'(\mathbf{u}^\star)\mathbf{e}(t) + \mathbf{r}(\mathbf{e}(t)), \quad t > t_0 \tag{2.39}$$

egyenlet. Vezessük be a $\mathbf{J}(\mathbf{u}^\star) = \mathbf{f}'(\mathbf{u}^\star)$ jelölést a Jacobi-mátrixra! Ekkor tehát az $\mathbf{e}(t)$ függvényre, amely az $\mathbf{u}(t)$ függvény $\mathbf{u}^\star$ egyensúlyi ponttól való eltávolodását írja le, a

$$\frac{d\mathbf{e}(t)}{dt} = \mathbf{J}(\mathbf{u}^\star)\mathbf{e}(t) + \mathbf{r}(\mathbf{e}(t)), \quad t > t_0 \tag{2.40}$$

egyenletet kapjuk. Mivel $\mathbf{r}(\mathbf{e}(t))$ függvényre a (2.34) feltétel teljesül, ezért alkalmazhatjuk a 2.18. lemmát. Eredményként az alábbi állítást nyerjük.

2.19. Tétel *Tegyük fel, hogy az $\boldsymbol{f}$ függvény kétszer folytonosan differenciálható, és legyen $\boldsymbol{u}^\star$ egy egyensúlyi pontja. Ha a $\boldsymbol{J}(\boldsymbol{u}^\star)$ Jacobi-mátrix mindegyik sajátértékének valós része negatív, akkor $\boldsymbol{u}^\star$ aszimptotikusan stabil egyensúlyi pont. Ha a $\boldsymbol{J}(\boldsymbol{u}^\star)$ mátrixnak létezik pozitív valós részű sajátértéke, akkor az $\boldsymbol{u}^\star$ egyensúlyi pont instabil.*

2.20. Példa *Térjünk vissza a 2.15. példára! A feladatban szereplő f függvény Jacobi-mátrixa skalár, és a szokásos derivált segítségével számolható. Tehát $J(u^\star) = f'(u^\star) = 1 - 2u^\star$. Ezért $J(u_1^\star) = J(0) = 1$, és $J(u_2^\star) = J(1) = -1$. Tehát az $u_1^\star = 0$ egyensúlyi pont instabil, míg az $u_2^\star = 1$ egyensúlyi pont aszimptotikusan stabil. Ez teljes összhangban van a 2.17. példa eredményével.*

2.1.3. Merev feladatok

A fizikai, műszaki és kémiai folyamatok matematikai modellezésénél gyakran adódnak olyan esetek, amelyeknél a keresett függvények skálázása jelentősen eltér. Például az egyik függvénynek megfelelő folyamat jóval gyorsabban zajlik le, mint a másik. Abban az esetben, amikor a két folyamat együttesen zajlik és egymásra kölcsönhatással vannak, akkor, mint azt a későbbiekben látni fogjuk, a leíró differenciálegyenletrendszer numerikus megoldására alkalmazott módszer stabil (legnagyobb megengedhető) lépésközének értékét a gyorsabb változás szabja meg. Így a megválasztható lépésköz esetenként

túlságosan is kicsi lesz, és akár a gépi nulla alá kerülhetünk. Ez az adott numerikus módszert alkalmazhatatlanná teszi. Az ilyen típusú feladatot *merev feladatnak* (angolul: stiff problem) nevezzük. Mint rövidesen látni fogjuk, ez a probléma skaláris feladatban is megfigyelhető.

2.21. Példa *Tekintsük ismételten a (2.10)-(2.11)*

$$u'(t) = \lambda u(t), \quad t > 0, \tag{2.41}$$
$$u(0) = u_0, \tag{2.42}$$

skaláris kezdetiérték-feladatot, ahol $\lambda \in \mathbb{C}$ adott komplex szám, amelyre $\operatorname{Re}\lambda < 0$. A feladat megoldása $u(t) = e^{\lambda t}u_0$, vagyis $|u(t)| = e^{\operatorname{Re}\lambda t}|u_0|$. Ezért $u(t)$ korlátos $t \to \infty$ esetén. Oldjuk meg numerikusan a feladatot a $h > 0$ egyenközű rácshálón az ismert explicit Euler-módszerrel, azaz az $u(t_n)$ közelítését az

$$y_n = y_{n-1} + h\lambda y_{n-1} = (1 + h\lambda)\, y_{n-1}, \quad n = 1, 2, \ldots, \quad y_0 = u_0$$

iterációval határozzuk meg.[6] *Ekkor $y_n = (1 + h\lambda)^n\, u_0$. Így a numerikus megoldás $n \to \infty$ csak akkor maradhat korlátos, amikor $|1+h\lambda| \le 1$, azaz a $h \le 2/\operatorname{Re}(-\lambda)$ feltétel teljesül. Nagy abszolút értékű,negatív valós részű λ esetén ez a korlát h megválasztására nézve a gyakorlatban nem valósítható meg.*

Bár a merevség fogalmának nincs pontos és egyértelmű definíciója, jellemezzük a jelenséget néhány matematikai feladat segítségével.

Először tekintsük az

$$u'(t) = \lambda u(t) - \lambda F(t) + F'(t)$$

differenciálegyenletet, ahol F egy adott, viszonylag lassan változó, folytonosan differenciálható függvény, $\lambda \ll 0$ adott szám. Az egyenlet megoldása

$$u(t) = F(t) + e^{\lambda t}\left(u(0) - F(0)\right).$$

Ezért valamely $h > 0$ esetén

$$u(t+h) = F(t+h) + e^{\lambda(t+h)}\left(u(0) - F(0)\right) = F(t+h) + e^{\lambda h}e^{\lambda t}\left(u(0) - F(0)\right)$$
$$= F(t+h) + e^{\lambda h}\left(u(t) - F(t)\right).$$

Mivel $\lambda \ll 0$, ezért $u(t+h) \simeq F(t+h)$, azaz a feladat megoldása, függetlenül az $u(0)$ kezdeti értékétől, az $F(t)$ függvény lesz. Ez egy extra stabilitási tulajdonság, és azt jelenti, hogy bármilyen nagy kezdeti hibával is indul a megoldás (azaz akárhonnan is indul a folyamat), egy kis idő elteltével már a pontos megoldás környezetében lesz a

[6] Az explicit Euler-módszerről a következő fejezetekben részletesen írunk.

megoldás. Tehát a stabilitási definíció úgy módosul, hogy a t_0-hoz közeli alkalmas t_1 esetén a definícióban $\delta > 0$ független ε-tól, azaz bármely $\overline{\mathbf{v}}_0$ esetén

$$\|\mathbf{u}(t; t_0, \mathbf{u}_0) - \mathbf{u}(t; t_0, \overline{\mathbf{v}}_0)\| < \varepsilon$$

minden $t > t_1$ esetén.

Vegyük észre, hogy az $u(t)$ megoldás két függvény összegéből tevődik össze: egy gyorsan változó (gyorsan lecsengő) tagból (azaz az $e^{\lambda t}\,(u(0) - F(0))$ tagból) és egy lassan változó tagból (azaz a $F(t)$ tagból). Az első, gyorsan lecsengő tagot a megoldás merev összetevőjének, a második (lassan változó) tagot pedig nem merev összetevőnek nevezzük. Nyilvánvalóan, ha az egyenletünket az $u(0) = u_0$ kezdeti feltétellel kiegészítve integráljuk ki a $[0, T]$ intervallumon, akkor ezen intervallum nagy részén a megoldást csak a nem merev tagja határozza meg. Azt a szakaszt, ameddig a merev összetevő kioltódik, *tranziens szakasznak*, azt a szakaszt, amikor már csak a nem merev összetevő dominál, *sima szakasznak*szokásos nevezni. (Természetesen a sima szakaszban a megoldás deriváltja jóval kisebb, mint a tranziens szakaszban.)

A fenti speciális stabilitási tulajdonság bizonyos értelemben előnyös, hiszen a megoldás gyakorlatilag érzéketlen a bemenő adatokra, és ezzel néhány hibaforrás (pl. az öröklött hiba) nincs hatással. Ugyanakkor érdekes (és az első ránézésre talán meglepő) módon éppen ez az a tulajdonság, amely a legfőbb problémát okozza a merev rendszerek numerikus megoldása során. Ugyanis, mint azt a fenti példában láttuk, λ konkrét értéke nincs kihatással a megoldásra a tranziens szakasz után, ugyanakkor a tranziens szakaszon ennek értékét figyelembe kell venni, és nagy $\mathrm{Re}(-\lambda)$ esetén h értéke csak nagyon kicsinek választható. (Pl. az explicit Euler-módszer esetén $h \leq 2/(-\operatorname{Re}\lambda)$.) Így a sima szakaszon meglehetősen sok lépést kell tennünk a teljes $[t_1, T]$ intervallumon való numerikus integráláshoz, és ennek hatásaként, elsősorban a hibák felhalmozódása miatt, a numerikus megoldás rendszerint eltorzul.

További példaként tekintsük az $u'' + (\gamma + 1)u' + \gamma u = 0$ másodrendű differenciálegyenletet az $u(0) = 1$ és $u'(0) = \gamma - 2$ kezdeti feltételekkel. Ekkor a feladat a szokásos módon átírható egy kétismeretlenes, elsőrendű, állandó együtthatós rendszerre, ahol az együtthatómátrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\gamma & -(\gamma+1) \end{bmatrix},$$

a kezdeti vektor pedig $\mathbf{c} = [1, \gamma - 2]$. Az $\mathbf{A}$ mátrix karakterisztikus egyenlete

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \lambda^2 + (\gamma + 1)\lambda + \gamma = 0.$$

Ezért az $\mathbf{A}$ mátrix sajátértékei $\lambda_1 = -1$ és $\lambda_2 = -\gamma$. A pontos megoldás tehát

$$\begin{aligned} u_1(t) &= \quad 2\exp(-t) - \exp(-\gamma t) \\ u_2(t) &= -2\exp(-t) + \gamma\exp(-\gamma t). \end{aligned} \tag{2.43}$$

Tegyük fel, hogy $\gamma >> 1$. Ekkor a pontos megoldás (2.43) alakjából látható, hogy az egyes megoldásokban szereplő $\exp(-\gamma t)$ függvények már nagyon kis t_1 mellett $t \geq t_1$ esetén sem játszanak szerepet, és a pontos megoldás gyakorlatilag az $u_1(t) \simeq -u_2(t) \simeq \exp(-t)$ lesz a $[t_1, T]$ intervallumon.

Ezen példa alapján látható (amely a közönséges differenciálegyenlet-rendszerek elméletéből jól ismert tény), hogy az $\mathbf{u}'(t) = \mathbf{A}\mathbf{u}(t)$ ($\mathbf{A} \in \mathbb{R}^{d\times d}$) állandó együtthatós lineáris rendszer általános megoldását az $\mathbf{A}$ mátrix sajátértékei segítségével írjuk fel és a megoldások a sajátértékek exponenciálisainak lineáris kombinációjaként állnak elő. Ezért például, ha az $\mathbf{A}$ mátrix sajátértékeinek valós részei negatívak, és nagyságrendben eltérnek egymástól, akkor a megoldások komponensei olyan függvények összegei, amelyek közül néhány időben gyorsan lecseng. (V.ö. a (2.43) alakkal.) Ezért az állandó együtthatós lineáris rendszerek esetén elfogadott a merev rendszer alábbi definíciója.

2.22. Definíció *Az $\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t) + \boldsymbol{f}(t)$ állandó együtthatós lineáris rendszert merev rendszernek nevezzük, ha az $\boldsymbol{A}$ mátrix λ_i $(i = 1, 2, \ldots, d)$ sajátértékeire teljesülnek az alábbiak.*

1. *Létezik nagy negatív valós részű sajátértéke, azaz van olyan λ_i, amelyre $\operatorname{Re}\lambda_i \ll 0$.*
2. *Létezik olyan sajátértéke is, amely abszolút értékben sokkal kisebb, mint az előző tulajdonságú sajátérték abszolút értéke, azaz van olyan λ_j, amelyre $|\lambda_j| \ll |\lambda_i|$.*
3. *Nincs nagy pozitív valós részű sajátértéke.*
4. *Nincs olyan nagy képzetes részű sajátértéke, amelynek valós részére nem teljesül a $\operatorname{Re}\lambda_i \ll 0$ reláció.*

Néhány megjegyzés a 2.22. definícióhoz.

- Feltesszük, hogy az $\mathbf{f}(t)$ függvény ugyanolyan simasággal rendelkezik, mint a megoldás lassan változó exponenciálisa. (Ellenkező esetben, pl. szakadásos függvény esetén, nem lesz merev a feladat.)
- Vegyük észre, hogy a feladat merevsége megengedi a pozitív valós részű sajátérték létezését is. (Tehát lehet $\operatorname{Re}\lambda_i > 0$ sajátérték is.) Ugyanakkor azt megköveteljük, hogy a gyorsan lecsengő összetevő aszimptotikusan stabil legyen, azaz nagy $|\lambda_i|$ esetén $\operatorname{Re}\lambda_i \ll 0$.
- A definícióból látszik, hogy az inhomogén, állandó együtthatós lineáris rendszer merevségének fogalma nem közvetlenül a megoldás alakjától, hanem az $\mathbf{A}$ együtthatómátrix spektrális tulajdonságaitól függ. Ennek illusztrálásaként tekintsük a fenti feladatot a következő megválasztásokkal.

$$\mathbf{A} = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}, \quad \mathbf{f}(t) = \begin{pmatrix} 2\sin t \\ 2(\cos t - \sin t) \end{pmatrix} \tag{2.44}$$

$$\mathbf{A} = \begin{pmatrix} -2 & 1 \\ 998 & -999 \end{pmatrix}, \quad \mathbf{f}(t) = \begin{pmatrix} 2\sin t \\ 999(\cos t - \sin t) \end{pmatrix} \tag{2.45}$$

Az $\mathbf{u}(0) = [2, 3]^\top$ kezdeti feltétel mellett mindkét feladat megoldása az

$$\mathbf{u}(t) = 2e^{-t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \sin t \\ \cos t \end{pmatrix}$$

függvény. Az első feladatban szereplő mátrix (azaz a (2.44) szerinti $\mathbf{A}$ mátrix) sajátértékei -1 és -3, míg a második feladatban a megfelelő sajátértékek -1 és -1000. Így a (2.44) megválasztású feladat nem merev, míg a (2.45) megválasztású merev. (Ha az explicit Euler-módszert segítségével numerikusan oldjuk meg a fenti két feladatot, akkor láthatjuk a különbséget a gyakorlatban.) Ugyanakkor homogén feladat esetén az $\mathbf{A}$ mátrix egyértelműen meghatározza a megoldást, tehát a megoldás alakja és a merevség definíciója megfelel egymásnak.

- Merev rendszerek esetén tehát

$$S = \frac{\max_i |\operatorname{Re} \lambda_i|}{\min_i |\operatorname{Re} \lambda_i|} \gg 1, \tag{2.46}$$

és az S számot a rendszer merevségi számának nevezzük. (A gyakorlatban már merevnek nevezzük a rendszert, ha $S = \mathcal{O}(10)$.

- Néhány korábbi munkában a merev rendszer definíciójában a $Re\lambda_i < 0$, $i = 1, 2, \ldots, d$ feltétel is szerepelt, azaz a rendszer aszimptotikus stabilitása eleve feltételezve volt. Mint látható, ez a 2.22. definícióban nem szerepel.

Akkor valójában mikor nevezhető merevnek egy rendszer? Lambert [9] három esetet sorol fel a merevség ismérveként.

- Lineáris állandó együtthatós rendszer esetén a sajátértékek valós részei negatívak, és a rendszer merevségi száma nagy.

- A rendszer numerikus integrálásához használt lépésköz megválasztásánál nem elsősorban a pontosság, hanem döntően a numerikus módszer stabilitása (a hibák felhalmozódásának folyamata) a meghatározó.

- A megoldás néhány komponense lényegesen gyorsabban cseng le, mint a többi.

Mint láttuk, az állandó együtthatós lineáris közönséges differenciálegyenlet-rendszerek esetén a 2.22. definíció alapján az $\mathbf{A}$ mátrix spektrumának ismeretében választ tudunk adni a kérdésre. Ugyanakkor a többi esetben a válasz már nem egyértelmű! Mint az alábbi példákban látni fogjuk, a lineáris időfüggő (nemautonóm rendszerek és

a nemlineáris feladatok esetén a spektrum már nem ad megbízható információt a rendszer merevségéről, ezért ezekre az esetekre a fenti a 2.22. definíció általában már nem alkalmazható.

A nemautonóm esetre példaként szolgálhat a következő [20].

2.23. Példa *Tekintsük az*

$$\boldsymbol{u}'(t) = \boldsymbol{A}(t)\boldsymbol{u}(t) \tag{2.47}$$

lineáris közönséges differenciálegyenlet-rendszert, ahol

$$\boldsymbol{A}(t) = \begin{pmatrix} -1 - 9\cos^2(6t) + 6\sin(12t) & 12\cos^2(6t) + \frac{9}{2}\sin(12t) \\ -12\sin^2(6t) + \frac{9}{2}\sin(12t) & -1 - 9\sin^2(6t) - 6\sin(12t) \end{pmatrix}.$$

Közvetlen számítással ellenőrizhető, hogy az $\boldsymbol{A}(t)$ mátrix sajátértékei $\lambda_1 = -1$ és $\lambda_2 = -10$, azaz t-től függetlenek. A fenti (2.47) egyenlet általános megoldása

$$\boldsymbol{u}(t) = C_1 e^{2t} \begin{pmatrix} \cos(6t) + 2\sin(6t) \\ 2\cos(6t) - \sin(6t) \end{pmatrix} + C_2 e^{-13t} \begin{pmatrix} \sin(6t) - 2\cos(6t) \\ 2\sin(6t) + \cos(6t) \end{pmatrix}.$$

Tehát a megoldásnak nincs köze az $\boldsymbol{A}(t)$ mátrix sajátértékeihez, másrészt pedig a $\operatorname{Re}\lambda_i < 0$ feltétel ugyan teljesül, de a megoldás nem cseng le. (Ettől még persze merev a rendszer, hiszen a megoldás második tagja hamar lecseng, és a megoldás tart az első sima taghoz.)

Most tekintsünk egy példát a nemlineáris feladatokra. Kézenfekvőnek tűnik, hogy a stabilitás vizsgálatához hasonlóan a merevséget is a linearizálással határozzuk meg, azaz a nemlineáris rendszer egyensúlyi pontja körüli linearizált rendszer Jacobi-mátrixára alkalmazzuk a 2.22. definíciót. Vizsgáljuk meg ezt a következő klasszikus példán!

2.24. Példa *Tekintsük az ún. Van der Pol egyenletet*

$$u'' - \mu(1 - u^2)u' + u = 0, \tag{2.48}$$

ahol $\mu > 0$ valamely adott paraméter. Bevezetve az $u_1 = u$ és az $u_2 = u'$ jelöléseket, a (2.48) egyenlet felírható

$$\frac{d}{dt}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} u_2 \\ \mu(1 - u_1^2)u_2 - u_1 \end{pmatrix} \tag{2.49}$$

alakban. Az

$$\boldsymbol{f}(u_1, u_2) = \begin{pmatrix} u_2 \\ \mu(1 - u_1^2)u_2 - u_1 \end{pmatrix}$$

függvényre $\boldsymbol{f}(0,0) = \boldsymbol{0}$, azaz $u_1 = u_2 = 0$ egyensúlyi pontja a Van der Pol egyenletnek. Linearizáljuk a feladatot ezen egyensúlyi pont körül! Könnyen látható, hogy az $\boldsymbol{f}$ függvény Jacobi-mátrixa a $\boldsymbol{0} = (0,0)^\top \in \mathbb{R}^2$ pontban

$$\boldsymbol{J}(\boldsymbol{0}) = \begin{pmatrix} 0 & 1 \\ -1 & \mu \end{pmatrix}.$$

Határozzuk meg $\boldsymbol{J}(\boldsymbol{0})$ sajátértékeit! Mivel a karakterisztikus egyenlet $\lambda^2 - \mu\lambda + 1 = 0$, ezért a sajátértékei

$$\lambda_1 = \frac{\mu}{2} + \sqrt{\frac{\mu^2}{4} - 1},$$

$$\lambda_2 = \frac{\mu}{2} - \sqrt{\frac{\mu^2}{4} - 1} = \frac{1}{\lambda_1}.$$

Tehát $\mu \to \infty$ esetén

$$\lambda_1 = \mathcal{O}(\mu), \quad \lambda_2 = \mathcal{O}\left(1/\mu\right). \tag{2.50}$$

Nagy μ esetén tehát a rendszer merevségi száma $\mathrm{S} = \mathcal{O}(\mu^2)$, ami a rendszer merevségét mutatja.

Általánosan, egy nemlineáris rendszer merevségét a rendszer (általában nemlineáris) $\mathbf{f}$ függvényének második változója szerint Lipschitz-állandó nagysága mutatja: nagy L esetén a rendszert merevnek nevezzük.

2.2. Időfüggő parciális differenciálegyenletek és modellezésük közönséges differenciálegyenlet-rendszerrel

A közönséges differenciálegyenletek (illetve a közönséges differenciálegyenlet-rendszerek) egy folyamat időbeli változásának leírására szolgál, ugyanakkor nem alkalmasak a térben és időben egyaránt változó jelenségek matematikai modellezésére. Ilyen esetben az ismeretlen függvényünk többváltozós, nevezetesen, a térbeli és az időbeli változók függvénye. Általában, a differenciálegyenletek közös vonása, hogy a függvény és deriváltjai közötti ismert kapcsolatból kell magát a függvényt meghatározni.

2.2.1. A parciális differenciálegyenletek alapfogalmai

Amikor a keresett függvény egyváltozós, akkor közönséges differenciálegyenletnek nevezzük a problémát. Amikor az ismeretlen függvény többváltozós, és így a kapcsolat az ismeretlen függvény és annak *parciális deriváltjai* között adott, *parciális differenciálegyenletről* beszélünk. Néhány tipikus kétváltozós parciális differenciálegyenlet és elnevezésük:

a. Laplace-egyenlet:

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 0, \tag{2.51}$$

b. Poisson-egyenlet:

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = f(x,y), \tag{2.52}$$

c. hővezetési egyenlet:

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \tag{2.53}$$

d. hullámegyenlet:

$$\frac{\partial^2 u(x,t)}{\partial t^2} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0, \tag{2.54}$$

e. advekciós egyenlet:

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial \left(k(x,t)u(x,t)\right)}{\partial x} = 0, \tag{2.55}$$

f. diffúziós egyenlet:

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial}{\partial x}\left(D(u,x,t)\frac{\partial u(x,t)}{\partial x}\right) = 0, \tag{2.56}$$

g. reakció-diffúziós egyenlet:

$$\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = -u^3(x,t), \tag{2.57}$$

h. biharmonikus egyenlet:

$$\frac{\partial^4 u(x,y)}{\partial x^4} + 2\frac{\partial^4 u(x,y)}{\partial x^2 \partial y^2} + \frac{\partial^4 u(x,y)}{\partial y^4} = 0. \tag{2.58}$$

A fenti példákban x, y a *térbeli változókat*, t az *időbeli változót*, u pedig az *ismeretlen függvényt* jelöli. A parciális differenciálegyenletek elméletében használatos néhány elnevezés. *Rendnek* nevezzük a legmagasabb előforduló parciális derivált számát. Ezért e. elsőrendű, h. negyedrendű, a többi pedig másodrendű parciális differenciálegyenlet. Megkülönböztetjük azokat az eseteket, amikor az ismeretlen függvény parciális deriváltjainak együtthatói *állandóak* avagy *függvények*. Példáinkban a., b., c., d., g. és h. állandó együtthatós parciális differenciálegyenlet, a többi pedig függvényegyütthatós. Fontos osztályozási szempont a linearitás: ha a függvénykapcsolatban az ismeretlen függvény és annak deriváltjai közötti kapcsolat lineáris, akkor *lineáris parciális differenciálegyenletnek*, ellenkező esetben *nemlineáris parciális differenciálegyenletnek* nevezzük a feladatunkat. Példáinkban f. és g. nemlineáris, a többi pedig lineáris parciális differenciálegyenlet. Végezetül, szokásos megkülönböztetni a *homogén* és az *inhomogén* parciális differenciálegyenleteket. Az utóbbi azt jelenti, hogy a függvénykapcsolatban szerepel az

ismeretlen u függvénytől illetve annak parciális deriváltjaitól nem függő összeadandó tag. Példáinkban b. inhomogén, a többi pedig homogén.

Térjünk ki Külön a független változók szerepére. Általában t az időt, x és y pedig a helyet jelöli. Ezek szerepe nem azonos: a folyamatokat általában $t \geq 0$ esetén vizsgáljuk, a térbeli változók viszont akármilyen előjelűek lehetnek. (Emellett a modellekben t mindig növekedő irányban változik, és a jelenségek tipikusan időben nem megfordíthatók, a térbeli változók viszont akármilyen irányban változhatnak.) Az olyan feladatot, amelyben az ismeretlen függvény időben nem változik (azaz u nem függ t-től), *stacionárius feladatnak* nevezzük, ellenkező esetben *időfüggő (instacionárius) feladatról* beszélünk.

Mi a továbbiakban a kétváltozós, másodrendű, lineáris parciális differenciálegyenletekkel foglalkozunk. Tekintsük tehát az $\Omega \subset \mathbb{R}^2$ tartományon az

$$\begin{aligned}(Lu)(x,y) = a(x,y)\frac{\partial^2 u(x,y)}{\partial x^2} + 2b(x,y)\frac{\partial^2 u(x,y)}{\partial x \partial y} + c(x,y)\frac{\partial^2 u(x,y)}{\partial y^2} + \\ d(x,y)\frac{\partial u(x,y)}{\partial x} + e(x,y)\frac{\partial u(x,y)}{\partial y} + g(x,y)u(x,y) = f(x,y)\end{aligned} \tag{2.59}$$

egyenletet, ahol az a, b, c, d, e, g *együtthatófüggvények* és az f *forrás* adottak. (Ezen függvények alkalmas megválasztásával, illetve az $y \sim t$ jelöléssel a korábbi lineáris másodrendű példák (a.-f.) mindegyike felírható.)
A (2.59) egyenletben szereplő L operátor

$$(L_0 u)(x,y) = a(x,y)\frac{\partial^2 u(x,y)}{\partial x^2} + 2b(x,y)\frac{\partial^2 u(x,y)}{\partial x \partial y} + c(x,y)\frac{\partial^2 u(x,y)}{\partial y^2} \tag{2.60}$$

részét az L operátor *főrészének* nevezzük. Az L_0 operátornak megfeleltethetjük a

$$B(\alpha,\beta) = a(x,y)\alpha^2 + 2b(x,y)\alpha\beta + c(x,y)\beta^2 \tag{2.61}$$

kvadratikus alakot, és ez alapján az L operátor alábbi osztályozása lehetséges. Tekintsük valamely rögzített $(x_0, y_0) \in \Omega$ pontban a $B(\alpha,\beta) =$ állandó > 0 egyenlőséggel definiált másodrendű görbéket az (α,β) síkon. Az $a(x_0,y_0), b(x_0,y_0)$ és $c(x_0,y_0)$ értékétől (pontosabban, az $a(x_0,y_0)c(x_0,y_0)$ $-b^2(x_0,y_0)$ kifejezés előjelétől) függően ez egy ellipszist, parabolát vagy hiperbolát határoz meg. Ez motiválja a következő definíciót.

2.1. Definíció *Azt mondjuk, hogy az L operátor (másképpen, a (2.59) egyenlet)*

- elliptikus típusú *az $(x,y) \in \Omega$ pontban, ha $a(x,y)c(x,y) - b^2(x,y) > 0$;*
- parabolikus típusú *az $(x,y) \in \Omega$ pontban, ha $a(x,y)c(x,y) - b^2(x,y) = 0$;*
- hiperbolikus típusú *az $(x,y) \in \Omega$ pontban, ha $a(x,y)c(x,y) - b^2(x,y) < 0$.*

Azt mondjuk, hogy elliptikus (parabolikus, hiperbolikus) típusú a $\Omega_1 \subset \Omega$ tartományon, ha elliptikus (parabolikus, hiperbolikus) típusú a Ω_1 tartomány mindegyik pontjában.

Ha (2.59) állandó együtthatós, akkor a Ω tartományon azonos típusú. Például a Laplace- és a Poisson-egyenletek elliptikus, a hővezetési egyenlet parabolikus, a hullám-egyenlet pedig hiperbolikus típusú a teljes Ω tartományon. Ugyanakkor, az

$$y\frac{\partial^2 u(x,y)}{\partial x^2} + 2x\frac{\partial^2 u(x,y)}{\partial x \partial y} + y\frac{\partial^2 u(x,y)}{\partial y^2} = 0$$

függvényegyütthatós egyenlet a $\Omega_{ell} = \{(x,y) \in \mathbb{R}^2, |y| > |x|\}$ halmazon elliptikus, a $\Omega_{par} = \{(x,y) \in \mathbb{R}^2, |y| = |x|\}$ halmazon parabolikus, a $\Omega_{hip} = \{(x,y) \in \mathbb{R}^2, |y| < |x|\}$ halmazon pedig hiperbolikus típusú.

Célunk olyan matematikai modellek megadása, amelyek *korrekt kitűzésűek*, azaz rendelkeznek az alábbi tulajdonságokkal:

a. létezik megoldása (egzisztencia);

b. ez a megoldása egyetlen (unicitás);

c. a megoldása folytonosan függ a feladatot meghatározó függvényektől (stabilitás).

Ezek a követelmények természetes módon következnek a matematikai modellezés jellegéből és céljából. Vegyük észre, hogy általános esetben a (2.59) alakú parciális differenciálegyenletnek, ha létezik is megoldása, akkor az nem egyértelmű. (Például a Laplace-egyenletnek megoldása az $u(x,y) = ax + by + c$ alakú lineáris függvény tetszőleges a, b és c állandók esetén.) Ezért tehát a parciális differenciálegyenletek önmagukban nem elegendőek a korrekt kitűzés biztosításához. Ehhez további feltételek megadása szükséges. Mivel tipikusan a megoldásfüggvényről a megoldási tartomány határán különböző információkkal rendelkezünk, ezért a kiegészítő feltételeket ezen információk segítségével szokásos megadni.

Amikor az megoldási tartomány az (x,y) térváltozókból álló $\Omega \subset \mathbb{R}^2$ korlátos halmaz, akkor az Ω tartomány Γ peremén írunk elő *peremfeltételeket*. Amikor tér-idő változó egyaránt szerepel, azaz Ω az (x,t) típusú pontokból áll, és a térváltozóban korlátos a tartomány, akkor a $t = 0$ pontban *kezdeti feltételeket*, a térbeli változó határán pedig továbbra is peremfeltételeket adhatunk meg. A peremfeltételek megadásának három típusa van.

- *első (Dirichlet-) típusú peremfeltétel*, ami azt jelenti, hogy a Γ-beli perempontban rögzítjük a megoldásfüggvény értékét;

- *második (Neumann-) típusú peremfeltétel*, ami azt jelenti, hogy a Γ-beli perempontban ismerjük a normál irányú deriváltjának az értékét;

- *harmadik (Robin-) típusú peremfeltétel*, ami azt jelenti, hogy a Γ-beli perempontban előre megadjuk a megoldásfüggvény és annak külső normálvektor irányú deriváltjának valamely lineáris kombinációjának értékét.

A *kezdeti (Cauchy-) feltétel* megadása azt jelenti, hogy $t = 0$ időpontban megadjuk a megoldásfüggvény vagy annak t szerinti deriváltjának az értékét.

Célunk, hogy egy adott parciális differenciálegyenletet olyan kiegészítő feltételekkel lássunk el, amelyekkel a feladat korrekt kitűzésű lesz. Megmutatható ([12, 13]), hogy lineáris esetben

- az elliptikus feladatok a Γ-n megadott első, második vagy harmadik peremfeltétellel,
- a parabolikus feladatok a $t = 0$-ban megadott $u(x, 0)$ kezdeti feltétellel és a térbeli határon megadott első, második vagy harmadik peremfeltételek egyikével,
- a hiperbolikus feladatok a $t = 0$ pontban megadott $u(x, 0)$ és $\dfrac{\partial u}{\partial t}(x, 0)$ kezdeti feltételekkel, valamint a térbeli határon megadott első, második vagy harmadik peremfeltételek valamelyikével

korrekt kitűzésűek. Amikor a térbeli változók Ω tartománya nem korlátos, akkor természetesen peremfeltétel nem értelmezhető. Ilyenkor időfüggő feladatok esetén a megfelelő kezdeti feltételek megadásával lesz a feladatunk korrekt kitűzésű.

Ebben a könyvben az időben változó feladatokkal foglalkozunk. Ezért a továbbiakban az alapvető instacionárius egyenletek származtatásával foglalkozunk.

2.2.2. Megmaradási törvények

Számos fizikailag fontos jelenséget leíró parciális differenciálegyenletet az ún. *megmaradási törvényekből* nyerjük. Jelölje $u(x, t)$ valamely anyag sűrűségfüggvényét az x helyen és a t időpontban. Tegyük fel, hogy ez a koncentráció időben változik, azaz például egy híg folyadékban szuszpenziót[7] hoz létre. Az egyszerűség kedvéért feltesszük, hogy a térbeli változó egydimenziós, például a szuszpenzió egy csőben történik, így x a csőben való távolságot jelenti. Egy további lehetséges modell a szennyezőanyag terjedésének vizsgálata egy patakban, de más példák is felsorolhatók. Feltehető, hogy tetszőleges keresztmetszet mentén a koncentráció állandó, azaz értéke csak x-től függ. Ekkor az

[7] Olyan durva diszperz rendszer, amelyben szilárd halmazállapotú részecskék vannak szétoszlatva egy folyékony halmazállapotú közegben anélkül, hogy feloldódnának abban. A folyékony anyagban a szilárd anyag részecskéi lebegnek, és hosszabb-rövidebb idő után leülepednek.

$u(x,t)$ sűrűségfüggvény segítségével a t időpontban meghatározhatjuk a vizsgált anyag teljes tömegét a cső x_1 és x_2 pontok közötti részében:

$$I(t) = \int_{x_1}^{x_2} u(x,t)\,\mathrm{d}x. \tag{2.62}$$

Megjegyezzük, hogy a sűrűségfüggvény mértékegysége pl. gramm/méter. Mivel valójában u a csőben értelmezett sűrűségfüggvény, ezért a megfelelő mértékegysége gramm/méter^3, és az integrálást a cső keresztmetszetén kell értelmezni.

A transzport jellegű alapvető fizikai folyamatok parciális differenciálegyenletei a következő elvből származtathatók. Tekintsük az $[x_1, x_2]$ részt és határozzuk meg, hogy ebben a részben $I(t)$ hogyan változik az időben. Mivel az *anyagmegmaradás elve* alapján ezen csőszakaszban új anyag nem keletkezik és nem semmisül meg, ezért a tömeg időbeli megváltozása csak két végpontban lévő fluxus[8] hatására történhet. Ezért a fluxus f megadásával felírhatjuk a megfelelő matematikai összefüggést.

Mivel az $[x_1, x_2]$ csőszakaszban lévő tömeg megváltozása csak a végpontokbeli fluxusok hatására történik, ezért

$$\frac{d}{dt}I(t) = f(u(x_1,t)) - f(u(x_2,t)). \tag{2.63}$$

(A negatív jel a fluxus irányultsága miatt van.) Feltéve f és u megfelelő simaságát, ez az egyenlet felírható

$$\frac{d}{dt}I(t) = -\int_{x_1}^{x_2} \frac{\partial}{\partial x} f(u(x,t))\,\mathrm{d}x \tag{2.64}$$

alakban. Mivel u megfelelő simasága esetén

$$\frac{d}{dt}I(t) = \int_{x_1}^{x_2} \frac{\partial}{\partial t} u(x,t)\,\mathrm{d}x, \tag{2.65}$$

ezért a (2.64) és a (2.65) összefüggések alapján

$$\int_{x_1}^{x_2} \left(\frac{\partial}{\partial t} u(x,t) + \frac{\partial}{\partial x} f(u(x,t)) \right) \mathrm{d}x = 0. \tag{2.66}$$

Mivel (2.66) tetszőleges $[x_1, x_2]$ intervallumon érvényes, ezért végeredményként a

$$\frac{\partial}{\partial t} u(x,t) + \frac{\partial}{\partial x} f(u(x,t)) = 0 \tag{2.67}$$

egyenletet kapjuk, amelyet *megmaradási (folytonossági) törvénynek* szokásos nevezni.

[8] A fluxus általában egy adott felületen átáramló anyag vagy energia mennyiségét jelenti.

Alapvető parciális differenciálegyenletek

A fluxus különböző alakjai vezetnek el az egyes jól ismert modellekhez.

1. *Advekciós egyenlet.* Tegyük fel, hogy a folyadék állandó a sebességű, tehát az anyagi részecske ezen állandó a sebességgel mozog. Ha az x pontbeli fluxus a pontbeli sebesség és a lokális sűrűség szorzata, akkor a fluxus függvénye

$$f(u) = au \tag{2.68}$$

alakú. Ekkor a (2.67) képlet alapján a

$$\frac{\partial}{\partial t}u(x,t) + a\frac{\partial}{\partial x}u(x,t) = 0 \tag{2.69}$$

egyenletet kapjuk. Ez az egyenlet, a parciális deriváltak szokásos jelölésével, felírható az egyszerűbb jelölésű

$$u_t + au_x = 0$$

alakban is[9]. Az egyértelmű megoldhatósághoz kezdeti feltétel, és, ha a térbeli változó korlátos, peremfeltételek megadása szükséges. Tekintsük az egyszerűbb nemkorlátos esetet, azaz amikor $x \in (-\infty, \infty)$. Ekkor ugyanis elegendő csak a kezdeti feltétel megadása:

$$u(x,0) = u_0(x), \quad x \in (-\infty, \infty), \tag{2.70}$$

ahol u_0 adott függvény. Könnyen látható, hogy ha u_0 folytonosan differenciálható, akkor az

$$u(x,t) = u_0(x - at), \quad x \in (-\infty, \infty),\ t \geq 0 \tag{2.71}$$

függvény egyértelmű megoldása a (2.69)-(2.70) feladatnak. A megoldás (2.71) alakjából jól látható, hogy az (x,t) sík azon pontjaiban, ahol $x - at$ állandó, (azaz az $x - at = C$ egyenesek pontjaiban) a megoldások ugyanazt az értéket veszik fel.

2. *Diffúziós egyenlet.* Tegyük fel, hogy a folyadék nem mozog, azaz $a = 0$. Ekkor $u_t = 0$, és ezért elvileg az u_0 kezdeti függvény alakja az időben nem változik. Ugyanakkor, ha u_0 nem állandó, a molekuláris mozgás hatására a kezdeti függvény mégis megváltozik. Ez a jelenség az ún. molekuláris diffúzió, amelynek hajtóereje a sűrűségkülönbség. (Ha más erő nem lép fel, a diffúzió képes megszüntetni a sűrűségkülönbséget.) Az anyagáramlás sebessége a sűrűséggradienssel arányos. Ezt a jelenséget Brown-mozgásnak is szokták hívni, melyet a Fick-törvény ír le. Ennek értelmében a fluxus egy adott x pontban egyenesen arányos a sűrűségfüggvény gradiensével, azaz a fluxus függvénye

$$f(u) = -\kappa u_x \tag{2.72}$$

[9]Könyvünkben mindkét jelölést alkalmazzuk.

alakú. Ebben az egyenletben κ lehet állandó (pl. homogén anyagokban), de általános esetben függhet x-től is. Ha κ állandó, akkor (2.72) és (2.67) alapján az

$$u_t = \kappa u_{xx} \tag{2.73}$$

egyenletet nyerjük. A $\kappa = \kappa(x)$ esetben a fenti összefüggések az

$$u_t = (\kappa u_x)_x \tag{2.74}$$

egyenletet eredményezik. A fenti (2.73) és (2.74) egyenleteket összefoglalóan *diffúziós egyenletnek*, a κ együtthatót pedig *diffúziós együtthatónak* nevezzük.

3. *Hővezetési egyenlet.* A hővezetés során a hő a diffúzióhoz hasonló módon terjed[10]. Hővezetés a termodinamika második főtétele szerint mindig a nagyobb hőmérsékletű hely felől a kisebb hőmérsékletű hely felé történik, azaz a hőmérsékleti gradiens irányában. Emellett a hőáramsűrűség (amely lényegében a fluxust jelenti) a Fourier-törvény szerint véges változás esetén egyenesen arányos a hőmérséklet-különbséggel és a szilárd test anyagi minőségére jellemző hővezetési tényezővel, illetve fordítva arányos a távolsággal. Ezért határesetben a (2.72) alakú összefüggést kapjuk. Ezért a (2.73) és (2.74) alakú diffúziós egyenleteket nyerjük, amelyeket szokásos *hővezetési egyenleteknek* is nevezni. A κ együtthatót ebben az esetben *hővezetési együtthatónak* szokásos nevezni.

4. *Advekció-diffúzió egyenlet.* A folyadékmozgás esetén tipikusan egyidejűleg van advekció és diffúzió is, azaz a fluxusra

$$f(u) = au - \kappa u_x \tag{2.75}$$

összefüggés érvényes. Ennek alapján az advekció-diffúzió egyenlet állandó diffúziós együttható és sebesség mellett

$$u_t + au = \kappa u_{xx} \tag{2.76}$$

alakú.

5. *Forrástag az alapegyenletben.* Visszatérve a kiindulási pontunkhoz, nyilvánvalóan $I(t)$ nemcsak a csőszakasz végpontjaiban lévő fluxus hatására változhat meg: ha van forrás vagy nyelő ezen a szakaszon, akkor az össztömeg ennek függvényében is megváltozhat. Jelölje $\Psi(x,t)$ az ilyen forrás /nyelő sűrűségfüggvényét[11]. A Ψ

[10] A hővezetés (konduktív hőátadás) a hőátadás olyan formája, amely a szilárd vagy nyugalomban lévő (nem áramló) folyékony vagy légnemű halmazállapotú rendszerekben, hőmérséklet-különbség hatására jön létre. A hőáramlástól (konvektív hőátadás) abban tér el, hogy nem történik anyagáramlás, hanem a hőátadás a belső energia részecskéről részecskére való átadásával történik.

[11] Pontosabban a belső energia (időegységre jutó) forrás-sűrűségét jelöli, vagyis az időegység alatt térfogategységben keletkező belső energiát jelenti.

függvény előjele mutatja, hogy az (x,t) pontban forrás (pozitív előjel) vagy nyelő (negatív előjel) van. Ennek figyelembevételével a (2.66) egyenlet alakja

$$\int_{x_1}^{x_2} \left(\frac{\partial}{\partial t} u(x,t) + \frac{\partial}{\partial x} f(u(x,t)) \right) \, \mathrm{d}x = \int_{x_1}^{x_2} \Psi(x,t) \, \mathrm{d}x \tag{2.77}$$

alakot ölti. Ezért forrás esetén a (2.67) egyenlet

$$\frac{\partial}{\partial t} u(x,t) + \frac{\partial}{\partial x} f(u(x,t)) = \Psi(x,t) \tag{2.78}$$

alakú. Az f fluxus egyes alakjainak behelyettesítésével megkaphatjuk a fentiekben felsorolt nevezetesebb egyenletek forrástaggal kiegészített változatait.

6. *Reakció-diffúzió egyenlet.* A folyamatokban a forrástagok megjelenésének egyik oka a kémiai reakció: amikor a jelenségben s számú anyag vesz részt, és azok egymással kémiai reakcióba lépnek, akkor ezen anyagok koncentrációi a kémiai átalakulások hatására megváltoznak. Ez a matematikai modellben azt jelenti, hogy ha $\mathbf{u} \in \mathbb{R}^s$ vektor koordinátái az egyes anyagok koncentrációit jelölik, akkor a kinetikai egyenlet $\mathbf{u}_t = R(\mathbf{u})$ alakú. (Ez az egyenlet azt feltételezi, hogy az egyes anyagok tökéletesen elkeverődnek, és értékük csak az időtől függ.) Ha a koncentráció térben is változik, akkor ezt az egyenletet a diffúziós egyenlettel kombináljuk, és eredményül a $\mathbf{u}_t = \kappa \mathbf{u}_{xx} + R(\mathbf{u})$ reakció-diffúzió egyenletet nyerjük. Mivel a diffúziós együttható anyagonként változhat, ezért általában κ egy diagonális mátrix, és nem skalár. Megjegyezzük, hogy a fenti egyenlethez advekciós tag is hozzáadható, például áramló folyadékok esetén.

A hővezetési egyenlet és megoldása

Könyvünkben megkülönböztetett szerepet szánunk a skaláris, térben egyváltozós hővezetési egyenletnek, ugyanis a hővezetési folyamat numerikus modellezésén keresztül jól jellemezhetők az egyes numerikus eljárások. Ezért továbbiakban az

$$u_t = \kappa u_{xx} \tag{2.79}$$

egyenletet vizsgáljuk. Két esetet vizsgálunk a továbbiakban.

1. Az x térbeli változó a teljes számegyenesen változik, azaz $x \in \mathbb{R}$. Ebben az esetben a (2.79) egyenlet nem korlátos tartományon van kitűzve, azaz a korrekt kitűzéshez szükséges és elegendő az
$$u(x,0) = u_0(x), \quad x \in \mathbb{R} \tag{2.80}$$
kezdeti feltétel megadása, ahol u_0 egy megfelelően sima, $\mathbb{R}$-en értelmezett, adott függvény.

2. Az x térbeli változó egy korlátos intervallumon változik, azaz $x \in [0, l]$. Ebben az esetben a (2.79) egyenlet korlátos tartományon van kitűzve, azaz a korrekt kitűzéshez szükséges és elegendő az

$$u(x, 0) = u_0(x), \quad x \in [0, l] \tag{2.81}$$

kezdeti feltétel megadása (ahol u_0 egy megfelelően sima, $\mathbb{R}$-en adott függvény), illetve a két peremen valamilyen típusú peremfeltétel. Tegyük fel, hogy a peremPontokban az első peremfeltétel adott, azaz

$$u(0, t) = \mu_1(t), \quad u(l, t) = \mu_2(t), \quad t > 0, \tag{2.82}$$

ahol $\mu_1(t)$ és $\mu_2(t)$ adott függvények.

Nemkorlátos tartomány esete. A Fourier-transzformáció.

Először a (2.79)-(2.80) feladat megoldásának előállításával foglalkozunk. Egy $v(x) \in L_2(\mathbb{R})$-beli függvény[12] Fourier-transzformáltján[13] azt az $L_2(\mathbb{R})$-en értelmezett operátort értjük, amely egy $v \in L_2(\mathbb{R})$-beli függvényhez a

$$\hat{v}(\xi) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v(x) e^{-i\xi x} \, \mathrm{d}x, \quad \xi \in \mathbb{R} \tag{2.83}$$

módon definiált függvényt rendeli hozzá. Ez a transzformált $\hat{v}(\xi)$ függvény szintén $L_2(\mathbb{R})$-beli, és a Parseval-egyenlőség felhasználásával megmutatható a normatartás is, azaz a

$$\|\hat{v}\|_2 = \|v\|_2 \tag{2.84}$$

reláció. Az $\{e^{i\xi x}, \ \xi \in \mathbb{R}\}$ függvényrendszer bázist alkot az $L_2(\mathbb{R})$ térben, ezért a Fourier-transzformácó segítségével az eredeti $v(x)$ függvényhez hozzárendelt $\hat{v}(\xi)$ értelmezhető $v(x)$ függvény $e^{i\xi x}$ szerinti koordinátájának a különböző $\xi \in \mathbb{R}$ értékekre. Így az inverz Fourier-transzformáció

$$v(x) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{v}(\xi) e^{i\xi x} \, \mathrm{d}\xi, \quad \xi \in \mathbb{R} \tag{2.85}$$

szerinti definíciója kézenfekvő. (Ez megfelel egy vektor - adott bázisban történő - lineáris kombinációként való felírásának.)

[12] $L_2(\mathbb{R})$ azon komplex függvények tere, amelyek négyzetesen integrálhatók Lebesgue-értelemben, azaz az $\int_{\mathbb{R}} |v|^2 \, \mathrm{d}x$ vett integrál (Lebesgue-értelemben) létezik és véges. Mint ismeretes, ez a tér vektortér, és a $(v, w) := \int_{\mathbb{R}} (v\overline{w}) \, \mathrm{d}x$ skaláris szorzattal Hilbert teret alkot. Ezért a megfelelő norma $\|v\|_2 = \left(\int_{-\infty}^{\infty} |v|^2 d\lambda \right)^{1/2}$.

[13] Jean-Baptiste Joseph Fourier (1768–1830) francia matematikus és fizikus

Feladatunkban a parciális differenciálegyenletek ismeretlen függvénye kétváltozós, ezért fontos az $u(x,t)$ függvény Fourier-transzformációjának meghatározására. Mi az x térbeli változó szerinti Fourier-transzformációt határozzuk meg, mégpedig úgy, hogy az $u(x,t)$ függvényt rögzített t mellett tekintjük, és úgy határozzuk meg a Fourier-transzformáltját:

$$\hat{u}(\xi,t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} u(x,t) e^{-i\xi x} \, \mathrm{d}x, \quad \xi \in \mathbb{R}. \tag{2.86}$$

Így az inverz Fourier-transzformáció

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{u}(\xi,t) e^{i\xi x} \, \mathrm{d}\xi, \quad x \in \mathbb{R}. \tag{2.87}$$

Felhasználva a (2.87) képletet, az $u(x,t)$ függvény parciális deriváltjai közvetlenül kifejezhetők, ugyanis a paraméteres integrál deriválhatósága mellett érvényesek az alábbi összefüggések

$$\begin{aligned} u_t(x,t) =& \frac{1}{\sqrt{2\pi}} \frac{\partial}{\partial t} \int_{\mathbb{R}} \hat{u}(\xi,t) e^{i\xi x} \, \mathrm{d}\xi = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{\partial}{\partial t} \left(\hat{u}(\xi,t) e^{i\xi x} \right) \, \mathrm{d}\xi = \\ & \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{u}_t(\xi,t) e^{i\xi x} \, \mathrm{d}\xi \end{aligned} \tag{2.88}$$

$$\begin{aligned} u_x(x,t) =& \frac{1}{\sqrt{2\pi}} \frac{\partial}{\partial x} \int_{\mathbb{R}} \hat{u}(\xi,t) e^{i\xi x} \, \mathrm{d}\xi = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{\partial}{\partial x} \left(\hat{u}(\xi,t) e^{i\xi x} \right) \, \mathrm{d}\xi = \\ & \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{u}(\xi,t) i\xi e^{i\xi x} \, \mathrm{d}\xi. \end{aligned} \tag{2.89}$$

Ezért tehát az $u_t(x,t)$ függvény x-szerinti Fourier-transzformáltja önmaga, míg az $u_x(x,t)$ függvényé az $i\xi\hat{u}(\xi,t)$ függvény. Hasonlóan,

$$\begin{aligned} u_{xx}(x,t) =& \frac{1}{\sqrt{2\pi}} \frac{\partial^2}{\partial x^2} \int_{\mathbb{R}} \hat{u}(\xi,t) e^{i\xi x} \, \mathrm{d}\xi = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{\partial^2}{\partial x^2} \left(\hat{u}(\xi,t) e^{i\xi x} \right) \, \mathrm{d}\xi \\ =& \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{u}(\xi,t)(-\xi^2) e^{i\xi x} \, \mathrm{d}\xi. \end{aligned} \tag{2.90}$$

Tehát az $u_{xx}(x,t)$ függvény Fourier-transzformáltja az $-\xi^2\hat{u}(\xi,t)$ függvény.

Térjünk át a (2.79)-(2.80) feladat megoldására! A (2.79) egyenlet mindkét oldalának Fourier-transzformációját véve, a (2.88) és a (2.90) képletek alapján az

$$\hat{u}_t(\xi,t) = -\kappa\xi^2\hat{u}(\xi,t), \quad \xi \in \mathbb{R} \tag{2.91}$$

egyenletet nyerjük. A (2.80) kezdeti feltétel Fourier-transzformáltja

$$\hat{u}(\xi,0) = \widehat{u_0}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} u_0(x) e^{-i\xi x} \, \mathrm{d}x, \quad \xi \in \mathbb{R}. \tag{2.92}$$

A (2.91)-(2.92) egy közönséges differenciálegyenlet elsőrendű Cauchy-feladata, amelynek megoldása az

$$\hat{u}(\xi,t) = e^{-\kappa\xi^2 t}\hat{u}(\xi,0) = e^{-\kappa\xi^2 t}\widehat{u_0}(\xi), \quad \xi \in \mathbb{R} \tag{2.93}$$

függvény.

Vizsgáljuk meg a hővezetési egyenlet megoldását abban az esetben, amikor a kezdeti függvény az $\bar{x} = 0$ pontra felírt Gauss-típusú eloszlásfüggvény, azaz

$$u_0(x) = e^{-\beta x^2}, \quad x \in \mathbb{R}, \tag{2.94}$$

ahol $\beta \in \mathbb{R}$ adott állandó. Mivel a (2.79)-(2.94) feladat megoldásához a kezdeti függvény Fourier-transzformáltjának ismerete szükséges, határozzuk ezt meg a továbbiakban. Nyilvánvalóan

$$\widehat{u_0}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\beta x^2} e^{-i\xi x}\, \mathrm{d}x. \tag{2.95}$$

Mindkét oldalt deriválva a következő relációk érvényesek:

$$\frac{d\widehat{u_0}}{\mathrm{d}\xi}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\beta x^2}(-ix) e^{-i\xi x}\, \mathrm{d}x. \tag{2.96}$$

A (2.96) jobb oldalán hajtsunk végre egy parciális integrálást a $v = e^{-i\xi x}$ és $w' = -ixe^{-\beta x^2}$ megválasztással! A kiintegrált rész eltűnik az $x = \pm\infty$ helyen, ezért

$$\begin{aligned} \frac{d\widehat{u_0}}{\mathrm{d}\xi}(\xi) &= -\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} i\frac{1}{2\beta} e^{-\beta x^2}(-i\xi) e^{-i\xi x}\, \mathrm{d}x = -\frac{\xi}{2\beta}\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\beta x^2} e^{-i\xi x}\, \mathrm{d}x \\ &= -\frac{\xi}{2\beta}\widehat{u_0}(\xi). \end{aligned} \tag{2.97}$$

A (2.97) egy elsőrendű közönséges differenciálegyenlet, amelynek általános megoldása

$$\widehat{u_0}(\xi) = e^{-\frac{\xi^2}{4\beta}}\widehat{u_0}(0). \tag{2.98}$$

A (2.92) képlet alapján

$$\begin{aligned} \widehat{u_0}(0) =& \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} u_0(x)dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\beta x^2}\, \mathrm{d}x = \\ & \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-s^2} \frac{1}{\sqrt{\beta}}\, \mathrm{d}s = \frac{1}{\sqrt{2\beta}}. \end{aligned} \tag{2.99}$$

(Az utolsó lépésben felhasználtuk a jól ismert $\int_{-\infty}^{\infty} e^{-s^2}\, \mathrm{d}s = \sqrt{\pi}$ összefüggést.) Ekkor a (2.98) és a (2.99) összefüggések alapján

$$\widehat{u_0}(\xi) = \frac{1}{\sqrt{2\beta}} e^{-\frac{\xi^2}{4\beta}}. \tag{2.100}$$

2.2. Megjegyzés Összevetve a (2.94) és a (2.100) formulákat azt láthatjuk, hogy a Fourier-transzformáció egy Gauss-féle eloszlást egy más paraméteres Gauss-féle eloszlásba transzformál.

Felhasználva a (2.93) és a (2.100) összefüggéseket, érvényes az

$$\hat{u}(\xi,t)=\frac{1}{\sqrt{2\beta}}e^{-\kappa\xi^2 t}e^{-\frac{\xi^2}{4\beta}}=\frac{1}{\sqrt{2\beta}}e^{-\frac{\xi^2}{4C}},\quad \xi\in\mathbb{R} \tag{2.101}$$

egyenlőség, ahol

$$C=\frac{1}{4\kappa t+1/\beta}. \tag{2.102}$$

Innen a (2.87) inverz Fourier-transzformáció segítségével előállítható a (2.79)-(2.94) feladat megoldására:

$$\begin{aligned} u(x,t)=&\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\hat{u}(\xi,t)e^{i\xi x}\,\mathrm{d}\xi=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\frac{1}{\sqrt{2\beta}}e^{-\frac{\xi^2}{4C}}e^{i\xi x}\,\mathrm{d}\xi\\ =&\sqrt{C/\beta}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\frac{1}{\sqrt{2C}}e^{-\frac{\xi^2}{4C}}e^{i\xi x}\,\mathrm{d}\xi,\quad x\in\mathbb{R}. \end{aligned} \tag{2.103}$$

Tekintsük a

$$\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\frac{1}{\sqrt{2C}}e^{-\frac{\xi^2}{4C}}e^{i\xi x}\,\mathrm{d}\xi$$

kifejezést! Ez nyilvánvalóan az $\frac{1}{\sqrt{2C}}e^{-\frac{\xi^2}{4C}}$ függvény inverz Fourier-transzformáltja. Az előzőekben beláttuk, hogy az $e^{-\beta x^2}$ függvény Fourier-transzformáltja $\frac{1}{\sqrt{2\beta}}e^{-\frac{\xi^2}{4\beta}}$, azaz ez utóbbi függvény inverz Fourier-transzformáltja az $e^{-\beta x^2}$ függvény. Ezt figyelembe véve tehát

$$\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\frac{1}{\sqrt{2C}}e^{-\frac{\xi^2}{4C}}e^{i\xi x}d\xi=e^{-Cx^2}. \tag{2.104}$$

A (2.103) és a (2.104) képletek alapján tehát a (2.79)-(2.94) feladat megoldása

$$u(x,t)=\sqrt{C/\beta}e^{-Cx^2}, \tag{2.105}$$

ahol C értékét a (2.102) képletből határozhatjuk meg. Mivel $\sqrt{C/\beta}=1/\sqrt{4\kappa\beta t+1}$, ezért a feladat megoldása

$$u(x,t)=\frac{1}{\sqrt{4\kappa\beta t+1}}e^{-\frac{x^2}{\sqrt{4\kappa t+1/\beta}}}. \tag{2.106}$$

Változtassuk meg a kezdeti függvényt! Vegyük észre, hogy a (2.99) alapján a (2.94) alakú kezdeti feltételt leíró függvény integráljára

$$\int_{\mathbb{R}}u_0(x)dx=\int_{\mathbb{R}}e^{-\beta x^2}\,\mathrm{d}x=\sqrt{\frac{\pi}{\beta}}.$$

Ezért az

$$u_0(x) = \sqrt{\frac{\beta}{\pi}} e^{-\beta x^2}, \quad x \in \mathbb{R} \tag{2.107}$$

kezdeti függvény esetén (2.79)-(2.107) egy olyan feladat, amelynek kezdeti eloszlásának integrálja egy, és a megoldása

$$u(x,t) = \frac{1}{\sqrt{4\kappa\pi t + \pi/\beta}} e^{-\frac{x^2}{\sqrt{4\kappa t + 1/\beta}}}. \tag{2.108}$$

Ha az $x = 0$ helyett valamely tetszőleges $x = \bar{x}$ pontra koncentrált Gauss-féle eloszlást tekintjük kezdeti függvényként, azaz

$$u_0(x,\bar{x}) = \sqrt{\frac{\beta}{\pi}} e^{-\beta(x-\bar{x})^2}, \quad x \in \mathbb{R}, \tag{2.109}$$

akkor a hozzátartozó megoldás

$$u(x,t,\bar{x}) = \frac{1}{\sqrt{4\kappa\pi t + \pi/\beta}} e^{-\frac{(x-\bar{x})^2}{\sqrt{4\kappa t + 1/\beta}}} \tag{2.110}$$

alakú.

A kezdeti függvények haranggörbéinek magasságát a β paraméter szabályozza: pozitív β esetén a maximumát az $x = \bar{x}$ pontban veszi fel, és az értéke $\sqrt{\beta/\pi}$. Tekintsük a kezdeti függvények és a hozzátartozó megoldások sorozatát $\beta \to \infty$ esetén! Ekkor minden rögzített $x \neq \bar{x}$ esetén a kezdeti függvények sorozata nullához, az $x = \bar{x}$ pontban pedig a $+\infty$-be tart, azaz a határfüggvény egy olyan $\delta(x)$ "függvény", amelyre

$$\delta(x) = \begin{cases} +\infty, & x = \bar{x} \\ 0, & x \neq \bar{x} \end{cases}, \tag{2.111}$$

és integrálja egy.[14] Ugyanakkor a (2.110) megoldások sorozata $\beta \to \infty$ esetén a

$$G(x,t,\bar{x}) = \frac{1}{\sqrt{4\kappa\pi t}} e^{-\frac{(x-\bar{x})^2}{\sqrt{4\kappa t}}} \tag{2.112}$$

alakú függvényhez tart. Ezt a függvényt a (2.79) hővezetési egyenlet *Green-függvényének* nevezzük. Mint látható, a Green-függvény is egy csökkenő Gauss-típusú eloszlásfüggvény, és egy pontszerű kezdeti állapot időben való terjedését írja le.

[14] Ezt a függvényt az irodalomban Dirac-delta függvénynek szokásos nevezni. (Paul Adrien Maurice Dirac (1902-1984) brit Nobel-díjas fizikus, a kvantummechanika egyik megalapozója.) Az ilyen típusú függvényt *általánosított függvénynek*, más szóval *disztribúciónak* nevezzük. Ez durván azt jelenti, hogy csak egy integrál részeként értelmezzük, mégpedig amely minden alkalmasan megválasztott $\varphi(x)$ függvény esetén az $\int_{\mathbb{R}} \delta(x)\varphi(x)\,\mathrm{d}x = \varphi(\bar{x})$ tulajdonsággal rendelkezik. A disztribúciók részletes tárgyalása nem képezi a könyv tárgyát, az érdeklődőknek javasoljuk a [12] könyvet.

2.3. Megjegyzés Tehát a $G(x,t,\bar{x})$ függvény értéke egy adott helyen azt mutatja meg, hogy az $\bar{x}$ pontbeli kezdeti inhomogenitás időben (és térben) hogyan terjed. (Például, ha egy ideális folyadékba egy pontban tintát cseppentünk.) A Green-függvény alakjából láthatjuk, hogy a kezdeti pontszerű függvény értéke tetszőlegesen $t > 0$ esetén is nem nulla bármelyik $x \neq \bar{x}, x \in \mathbb{R}$ pontban. Mivel a kezdeti függvény ebben a pontban nulla, ezért ez azt jelenti, az $\bar{x}$ nem nulla kezdeti értéke *bármelyik* $x \in \mathbb{R}$ pontban megváltoztatja a kezdeti nulla értéket, azaz a terjedési sebesség ezen modellben végtelen.

Tetszőleges kezdeti függvények esetén a megoldás a Green-függvény segítségével közvetlenül meghatározható. Mivel a Green függvény a rögzített $\bar{x} \in \mathbb{R}$ pontszerű (és integrál-értelemben egységnyi) kezdeti függvényhez tartozó megoldás, ezért az általános alakú $u_0(x)$ kezdeti függvény esetén minden pontban meghatározzuk a hatását, és az eredményeket összegezzük. Ezért a megoldást a Green-függvény és a kezdeti függvény szorzata által definiált függvény integrálja, azaz

$$u(x,t) = \int_{\mathbb{R}} G(x,t,\bar{x}) u_0(\bar{x}) d\bar{x} = \frac{1}{\sqrt{4\kappa\pi t}} \int_{-\infty}^{+\infty} e^{-\frac{(x-\bar{x})^2}{\sqrt{4\kappa t}}} u_0(\bar{x}) d\bar{x} \tag{2.113}$$

alakban áll állítható elő. A (2.113) összefüggést Poisson-féle képletnek[15] is szokásos nevezni. Vegyük észre, hogy az $u(x,t)$ függvény nem más, mint a

$$\widehat{G}(x,t) = \frac{1}{\sqrt{4\kappa\pi t}} e^{-\frac{x^2}{\sqrt{4\kappa t}}} \tag{2.114}$$

és az $u_0(x)$ függvények konvolúciója.[16] Könnyen ellenőrizhető, hogy a $\widehat{G}(x,t)$ függvény kielégíti a (2.79) hővezetési egyenletet, továbbá, hogy

$$\lim_{t \to 0+0} \widehat{G}(x,t) = \lim_{x \pm \infty} \widehat{G}(x,t) = 0.$$

A fenti $\widehat{G}$ függvényt *alapmegoldásnak* is szokás nevezni. Tehát a kezdetiérték-feladat megoldása előállítható az alapmegoldás és a kezdeti függvény konvolúciójaként.

A szakasz befejezéseként röviden áttekintjük az

$$u_t + a u_x = 0, \quad x \in \mathbb{R}, \ t > 0 \tag{2.115}$$

egyenlet

$$u(x,0) = u_0(x), \quad x \in \mathbb{R} \tag{2.116}$$

[15] Siméon Denis Poisson (1781–1840), francia matematikus, geométer és fizikus.

[16] A $(-\infty, \infty)$ intervallumon értelmezett f és g integrálható függvények konvolúcióján az $(f * g)(x) = \int_{-\infty}^{\infty} f(s) g(x-s)\, \mathrm{d}s$ integrált értjük. A konvolúciónak széles körű alkalmazásai vannak a valószínűségszámításban, a Fourier-sorok és a parciális differenciálegyenletek elméletében.

kezdeti feltétel melletti megoldásának előállítását a Fourier-transzformáció segítségével:

A (2.88) és a (2.89) összefüggések alapján az $u_t(x,t)$ függvény Fourier-transzformáltja önmaga, míg $u_x(x,t)$ függvényé az $i\xi\hat{u}(\xi,t)$ függvény. Ezért a (2.115) egyenlet Fourier-transzformációjával az

$$\hat{u}_t(\xi,t) = -i\xi a\hat{u}(\xi,t), \quad \xi \in \mathbb{R}, t > 0 \tag{2.117}$$

elsőrendű, közönséges differenciálegyenletet kapjuk. Ezért

$$\hat{u}(\xi,t) = e^{-i\xi at}\hat{u}(\xi,0). \tag{2.118}$$

Mivel $\hat{u}(\xi,0) = \widehat{u_0}(\xi)$ a (2.92) alapján meghatározható, ezért az inverz Fourier-transzformáció felhasználásával a megoldás közvetlenül előállítható:

$$u(x,t) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-i\xi at}\widehat{u_0}(\xi)e^{i\xi x}\,\mathrm{d}\xi = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \widehat{u_0}(\xi)e^{i\xi(x-at)}\,\mathrm{d}\xi = u_0(x-at). \tag{2.119}$$

(Az utolsó egyenlőség abból adódik, hogy az integrál nem más, mint az $\widehat{u_0}$ függvény inverz Fourier-transzformáltja az $x-at$ pontban.)

2.4. Megjegyzés Mint a (2.118) összefüggésből látható, $|\hat{u}(\xi,t)| = |\widehat{u_0}(\xi)|$ minden $t > 0$ esetén. Ez azt jelenti, hogy a megoldás Fourier-komponense időben nem változik. A hővezetési egyenlet esetén más a helyzet: a (2.93) képlet következtében $|\hat{u}(\xi,t)| = e^{-\kappa\xi^2 t}|\widehat{u_0}(\xi)|$ minden $t > 0$ esetén, azaz $|\hat{u}(\xi,t)|$ az időben exponenciálisan lecseng. A lecsengés sebessége függ a κ együtthatótól, illetve a ξ hullámszámtól. Ezért a magas frekvenciájú, erősen oszcilláló tagok (azaz amikor ξ^2 nagy) lényegesen gyorsabban lecsengenek, mint az alacsony frekvenciájúak. Ennek következtében a kezdeti függvény az időben kisimul. (Ezt szokásos *parabolikus simításnak* is nevezni.)

2.5. Megjegyzés Az előző megjegyzésből látható, hogy a κ konstans hővezetési együtthatóra célszerű kikötni a $\kappa > 0$ előjelfeltételt, bár a megoldás előállításánál ez nem volt szükséges. Ennek alapvető oka, hogy negatív κ esetén a megoldás Fourier-komponense az időben exponenciálisan nem lecsengő, hanem exponenciálisan növekedő lesz. Ez azt jelentené, hogy a magas frekvenciájú tagok (azaz amikor ξ^2 nagy) kinőnek, nevezetesen $\xi \to \infty$ esetén exponenciális sebességgel tart a végtelenhez. Ezért a kezdeti függvény bármilyen kis megváltozása (pl. kerekítési hibája) a megoldás ezen tagjaiban nem korlátos nagyságú megváltozást eredményez, azaz a feladat instabillá válik. Ennek eredményeként tehát a $\kappa < 0$ hővezetési együtthatójú feladat *nem lesz korrekt kitűzésű*. (Könyvünk a nem korrekt kitűzésű feladatokkal nem foglalkozik.)

A korlátos tartomány esete. A változók szétválasztásának módszere.

Tekintsük a hővezetési egyenletet a *térben* egydimenziós, korlátos tartományon. Az egyszerűség kedvéért feltesszük, hogy $x \in (0,1)$, és így az egyenletet a $(0,1) \times (0,T)$ $(T \leq \infty)$ tartományon vizsgáljuk. A korrekt kitűzéshez ekkor kezdeti- és peremfeltétel megadása egyaránt szükséges. Legyen $t^\star \in (0,T)$ egy adott rögzített szám, és jelölje $\Omega_{t^\star} = (0,1) \times (0,t^\star]$ $\mathbb{R}^2$-beli halmazt, ahol a megoldást keressük. $\Gamma_{t^\star}$ pedig a $t = 0$, $x = 0$ és $x = 1$ egyenesekkel határolt ponthalmazt, azaz $\Gamma_{t^\star} = \overline{\Omega}_{t^\star} \backslash \Omega_{t^\star}$. A $\Gamma_{t^\star}$ pontjait *parabolikus peremnek* nevezzük.

Tekintsük tehát a továbbiakban $\Omega_{t^\star}$ pontjaiban a

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad (x,t) \in \Omega_{t^\star} \tag{2.120}$$

egyenletet a $\Gamma_{t^\star}$ parabolikus peremen megadott kiegészítő ("kezdeti+perem") feltételekkel:

$$u(x,0) = u_0(x), \quad x \in (0,1); \qquad u(0,t) = u(1,t) = 0, \quad t \in [0,t^\star]. \tag{2.121}$$

Ha $u_0 = 0$, akkor a feladat megoldása az $u = 0$ függvény. Ezért a továbbiakban feltesszük, hogy $u_0 \neq 0$, így a keresett megoldás $u \neq 0$.

Keressük a megoldást

$$u(x,t) = X(x) \cdot T(t) \tag{2.122}$$

szétválasztható alakban, ahol X és T egyelőre ismeretlen, megfelelően sima függvények.[17] Behelyettesítve a (2.122) alakú u függvényt a (2.120) egyenletbe, a

$$T'(t) \cdot X(x) - X''(x) \cdot T(t) = 0, \quad (x,t) \in \Omega_{t^\star} \tag{2.123}$$

egyenletet nyerjük. Mivel $X(x) \cdot T(t) \neq 0$, ezért (2.123) az

$$\frac{X''(x)}{X(x)} = \frac{T'(t)}{T(t)}, \quad x \in (0,1), \ t \in (0,t^\star) \tag{2.124}$$

azonosságot jelenti. A (2.124) tetszőleges x és t változóra érvényes, ami azt jelenti, hogy mindkét oldal állandó: valamely $\lambda \in R$ szám mellett minden $x \in (0,1)$ és $t \in (0,t^\star]$ esetén érvényes az

$$\frac{X''(x)}{X(x)} = \frac{T'(t)}{T(t)} = \lambda \tag{2.125}$$

egyenlőség. Innen az

$$X''(x) = \lambda X(x), \quad x \in (0,1) \tag{2.126}$$

[17] Szokásos Fourier-módszernek is nevezni ezt az eljárást.

egyenletet kapjuk. Másrészt, behelyettesítve a (2.122) alakot a (2.121) két peremfeltételébe, az

$$X(0) = X(1) = 0 \tag{2.127}$$

feltételt kapjuk. Célunk tehát olyan $\lambda \in \mathbb{R}$ szám meghatározása, amely mellett a (2.126)-(2.127) feladatnak létezik a triviális $X(x) = 0$ függvénytől különböző megoldása.

Mivel (2.126) egy állandó együtthatós, másodrendű közönséges differenciálegyenlet, ezért általános megoldását az $s^2 + \lambda = 0$ karakterisztikus egyenletének gyökeivel határozhatjuk meg. Ez a λ előjelének függvényében az alábbi esetekhez vezet.

- Tegyük fel, hogy $\lambda < 0$. Ekkor a karakterisztikus egyenlet gyökei valósak, és a (2.126) egyenlet általános megoldása

  $$X(x) = C_1 e^{\sqrt{-\lambda}x} + C_2 e^{-\sqrt{-\lambda}x}.$$

  Mivel ekkor az $X(0) = 0$ feltétel a $C_1 + C_2 = 0$ egyenlőséget jelenti, ezért $X(x) = C_1 \left(e^{\sqrt{-\lambda}x} - e^{-\sqrt{-\lambda}x}\right)$ alakú. A másik, $X(1) = 0$ peremfeltételt ide behelyettesítve az $X(1) = C_1 \left(e^{\sqrt{-\lambda}} - e^{-\sqrt{-\lambda}}\right) = 0$ feltételt kapjuk. Mivel $e^{\sqrt{-\lambda}} > e^{-\sqrt{-\lambda}}$, ezért a fenti egyenlőség csak $C_1 = 0$ esetén lehetséges, ami az $X(x) = 0$ megoldást eredményezi. Mivel ez nem megengedett, ezért ez az eset nem lehetséges.

- Tegyük fel, hogy $\lambda = 0$. Ekkor a (2.126) egyenlet $X''(x) = 0$ alakú, így az általános megoldása $X(x) = C_1 x + C_2$. Erre a függvényre, amelynek képe egy egyenes, a (2.127) feltétel csak $C_1 = C_2 = 0$ esetén lehetséges. Ez viszont szintén a nem megengedett $X(x) = 0$ triviális megoldáshoz vezet.

- Tegyük fel, hogy $\lambda > 0$. Ekkor a karakterisztikus egyenlet gyökei $\pm i\sqrt{\lambda}$. Így a (2.126) egyenlet általános megoldása

  $$X(x) = C_1 \sin(\sqrt{\lambda}x) + C_2 \cos(\sqrt{\lambda}x).$$

  Az $X(0) = 0$ feltétel miatt $C_2 = 0$. Tehát az $X(1) = 0$ feltétel az $C_1 \sin\sqrt{\lambda} = 0$ feltételt jelenti. Mivel $C_1 \neq 0$, ezért ez a $\sqrt{\lambda} = k\pi$ $(k = 1, 2, \ldots)$ feltételt jelenti, azaz a lehetséges λ értékekre a

  $$\lambda_k = -k^2\pi^2, \quad k = 1, 2, \ldots \tag{2.128}$$

  értékeket kapjuk. Tehát az

  $$X_k(x) = C_1^k \sin k\pi x, \quad k = 1, 2, \ldots \tag{2.129}$$

  függvények tetszőleges C_1^k állandók mellett megoldásai a $\lambda = \lambda_k$ megválasztású (2.126)-(2.127) feladatnak.

Térjünk át a $T(t)$ függvény meghatározására! A (2.125) egyenlőség felhasználásával a

$$T'(t) = \lambda T(t), \quad t \in (0, t^\star] \tag{2.130}$$

egyenletet kapjuk, ahol (2.128) alapján $\lambda = \lambda_k = -k^2\pi^2$. Az egyenlet általános megoldása

$$T_k(t) = C_2^k e^{\lambda_k t}, \tag{2.131}$$

ahol C_2^k tetszőleges állandó. Tehát a (2.129) és a (2.131) képletekkel azt kapjuk, hogy minden $k = 1, 2, \ldots$ esetén tetszőleges C_k állandó mellett az

$$u_k(x,t) = X_k(x)T_k(t) = C_k e^{-k^2\pi^2 t}\sin(k\pi x) \tag{2.132}$$

függvények olyan függvények, amelyek megoldásai a (2.120) egyenletnek, és kielégítik a (2.121) mindkét (homogén) peremfeltételét. Tegyük fel, hogy a $\sum_{k=1}^{\infty} u_k(x,t)$ függvénysor egyenletesen konvergens! Ekkor az

$$u(x,t) = \sum_{k=1}^{\infty} u_k(x,t) \tag{2.133}$$

függvény is rendelkezik ezekkel a tulajdonságokkal. Válasszuk meg a tetszőleges C_k állandókat úgy, hogy a (2.121) kezdeti feltétele is teljesüljön erre az u függvényre! A (2.133) és a (2.132) képletek alapján

$$u(x,0) = \sum_{k=1}^{\infty} C_k \sin(k\pi x). \tag{2.134}$$

Ugyanakkor a $u_0(x)$ függvény Fourier-sora

$$u_0(x) = \sum_{k=1}^{\infty} u_0^k \sin(k\pi x) \tag{2.135}$$

alakú, ahol

$$u_0^k = 2\int_0^1 u_0(s)\sin(k\pi s)\,\mathrm{d}s. \tag{2.136}$$

A (2.134) és a (2.135) képletek összevetéséből tehát

$$C_k = u_0^k. \tag{2.137}$$

Összegezve: az

$$u(x,t) = \sum_{k=1}^{\infty} u_0^k e^{-k^2\pi^2 t}\sin k\pi x \tag{2.138}$$

függvénysorral definiált $u(x,t)$ függvény - a függvénysor egyenletes konvergenciája esetén - megoldása a (2.120)-(2.121) feladatnak.

2.6. Megjegyzés Hogyan biztosítható a (2.138) függvénysor egyenletes konvergenciája? Megmutatható ([13]), hogy ha u_0 folytonosan differenciálható, valamint $u_0(0) = u_0(1) = 0$, akkor az egyenletes konvergencia érvényes.[18]

2.7. Megjegyzés A fenti megoldás segítségével a

$$\frac{\partial U(x,t)}{\partial t} - \frac{\partial^2 U(x,t)}{\partial x^2} = 0, \quad (x,t) \in \Omega_{t^\star} \tag{2.139}$$

$$U(x,0) = \mathrm{U}_0(x), \quad x \in (0,1); \quad U(0,t) = \alpha, \quad u(1,t) = \beta, \quad t \in [0,t^\star] \tag{2.140}$$

(ahol α és β adott állandók) inhomogén feladat megoldása is könnyen előállítható hasonló végtelen függvénysor alakjában. Jelölje ugyanis

$$\tilde{u}(x,t) = \alpha(1-x) + \beta x, \quad x \in [0,1] \tag{2.141}$$

ismert függvényt. Ekkor az $u = U - \tilde{u}$ függvény megoldása lesz a (2.120)-(2.121) feladatnak, ahol $u_0(x) = \mathrm{U}_0(x) - (\alpha(1-x) + \beta x)$. Ezért ennek a függvénynek a (2.136) képlet szerinti Fourier-együtthatóival előállított (2.138) függvénysora meghatározza az u függvényt, és ennek ismeretében az ismeretlen függvényt kiszámolhatjuk az $U = u + \tilde{u}$ összefüggésből.

Az előzőekben tehát megmutattuk, hogy a (2.79)-(2.80) feladat megoldása a (2.113) Poisson-képlet segítségével, míg a (2.120)-(2.121) feladat megoldása pedig a (2.138) függvénysor segítségével állítható elő. Ugyanakkor ezek a megoldások többnyire csak formálisnak tekinthetők, a gyakorlatban nem, vagy csak nagyon speciális esetben alkalmazhatók egy konkrét feladat megoldására. (Mivel a (2.113) képletben egy végtelen tartományon vett improprius integrál, a (2.138) képletben pedig egy végtelen függvénysor összegének meghatározása szükséges.) További probléma, hogy mindkét megoldást csak a tárgyalt speciális feladatosztályra tudjuk alkalmazni, azaz például $\kappa = \kappa(x)$ nem állandó hővezetési együttható esetén nem működnek.

Mindez azt motiválja, hogy ezen feladatok megoldására valamilyen közelítő eljárást, azaz numerikus módszert célszerű alkalmazni. Könyvünk tárgya a közönséges differenciálegyenletek (illetve a közönséges differenciálegyenlet-rendszerek) numerikus megoldása. Ezért célunk, hogy a fenti parciális differenciálegyenletekből ilyen feladatot konstruáljunk.

Nemkorlátos térbeli tartomány esetén a Fourier-transzformáció segítségével a (2.91)-(2.92) közönséges differenciálegyenlet elsőrendű Cauchy-feladatát nyerjük az ismeretlen megoldásfüggvény Fourier-transzformáltjára, amelynek megoldása formálisan közvetlenül előállítható. Ugyanakkor a kezdeti érték meghatározásához egy nehezen realizálható

[18] Valójában u_0 simaságára elégséges a $[0,1]$ intervallumon való folytonosság, valamint a $(0,1)$ intervallumon való szakoszonkénti folytonosan differenciálhatóság is.

integrál meghatározása szükséges, valamint egy inverz Fourier-transzformáció numerikus realizálása. Könyvünkben ezt az esetet nem részletezzük, az érdeklődőknek javasoljuk a [10] könyvet.

A továbbiakban a korlátos térbeli tartomány esetével foglalkozunk részletesebben.

2.2.3. A korlátos térbeli tartományon kitűzött hővezetési egyenlet és szemidiszkretizációjának tulajdonságai

Tekintsük a továbbiakban $\Omega_{t^\star}$ pontjaiban a

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad (x,t) \in \Omega_{t^\star} \tag{2.142}$$

egyenletet a $\Gamma_{t^\star}$ parabolikus peremen megadott

$$u(x,0) = u_0(x), \quad x \in (0,1); \quad u(0,t) = \alpha, \quad u(1,t) = \beta, \quad t \in [0,t^\star] \tag{2.143}$$

feltételekkel. Mint láttuk, ennek a feladatnak létezik megoldása, amely a 2.7. megjegyzés szerint előállítható.

Ugyanakkor felmerülhet a kérdés: nem létezik-e a (2.142)-(2.143) feladatnak ettől eltérő más megoldása is? Milyen tulajdonságai vannak a megoldásnak? Érzékeny-e a megoldás a bemenő adatokra? A következőkben ezzel foglalkozunk.

A korlátos térbeli tartományon kitűzött hővezetési egyenlet korrekt kitűzése

Vizsgálatunkat egy alapvető fontosságú lemmával kezdjük.

2.8. Lemma *Tegyük fel, hogy a $w(x,t)$ függvény az $\Omega_{t^\star}$ halmazon kielégíti a (2.142) egyenletet, és folytonos az $\overline{\Omega}_{t^\star}$ halmazon. Ekkor teljesül rá az ún.* parabolikus maximum-minimum elv*: a függvény a legnagyobb és legkisebb értékét felveszi a $\Gamma_{t^\star}$ parabolikus peremen.*

Bizonyítás. Tehát belátandók a

$$\max_{\overline{\Omega}_{t^\star}} w(x,t) = \max_{\Gamma_{t^\star}} w(x,t), \quad \min_{\overline{\Omega}_{t^\star}} w(x,t) = \min_{\Gamma_{t^\star}} w(x,t) \tag{2.144}$$

állítások. Jelölje $M = \max_{\Gamma_{t^\star}} w(x,t)$. Indirekt módon tegyük fel, hogy létezik olyan $(x_0,t_0) \in \overline{\Omega}_{t^\star} \setminus \Gamma_{t^\star}$ pont, amelyre

$$w(x_0,t_0) = \mu > M. \tag{2.145}$$

Vezessük be a

$$v(x,t) = w(x,t) + \frac{\mu - M}{2t^\star}(t_0 - t) \tag{2.146}$$

függvényt! Ha $(x,t) \in \Gamma_{t^\star}$, akkor

$$v(x,t) \leq M + \frac{\mu - M}{2} = \frac{\mu + M}{2} < \mu.$$

Mivel $v(x_0,t_0) = w(x_0,t_0) = \mu$, ezért az $\overline{\Omega}_{t^\star}$ zárt halmazon folytonos függvény a maximumát nem veheti fel a $\Gamma_{t^\star}$ parabolikus peremen. Ezért, ha $\max_{\overline{\Omega}_{t^\star}} v(x,t) = v(x_1,t_1)$, akkor $(x_1,t_1) \notin \Gamma_{t^\star}$, azaz $x_1 \in (0,1)$ és $t_1 \in (0,t^\star]$. Mivel a v függvény simasága megegyezik w függvény simaságával, ezért ebben az (x_1,t_1) pontban létezik t szerinti első, illetve x szerinti második deriváltja is. Vizsgáljuk meg ezek előjelét!

- Ha $t_1 = t^\star$, akkor a $v(x_1,t)$ függvény a t_1 pontban nem csökken, hiszen ebben a pontban maximum van. Ezért a v függvény t-szerinti parciális deriváltja nemnegatív az (x_1,t_1) pontban.

- Ha $t_1 < t^\star$, akkor $v(x_1,t)$ függvénynek a $t = t_1$ pontban lokális maximuma van, azaz a v függvény t-szerinti parciális deriváltja az (x_1,t_1) pontban nulla.

- Mindkét fenti esetben a $v(x,t_1)$ függvénynek az $x = x_1$ pont belső pontja, és itt lokális maximuma van. Ezért tehát a v függvény x-szerinti 2. parciális deriváltja az (x_1,t_1) pontban nempozitív.

Így tehát összefoglalóan

$$\frac{\partial v}{\partial t}(x_1,t_1) \geq 0, \quad \frac{\partial^2 v}{\partial x^2}(x_1,t_1) \leq 0,$$

azaz

$$\frac{\partial v}{\partial t}(x_1,t_1) - \frac{\partial^2 v}{\partial x^2}(x_1,t_1) \geq 0. \tag{2.147}$$

Másrészt, a v függvény a (2.146) szerinti definiciója alapján

$$\frac{\partial v}{\partial t}(x,t) - \frac{\partial^2 v}{\partial x^2}(x,t) = \frac{\partial w}{\partial t}(x,t) - \frac{\partial^2 w}{\partial x^2}(x,t) - \frac{\mu - M}{2t^\star} < 0 \tag{2.148}$$

minden (x,t) pontra, amely nem a parabolikus peremen helyezkedik el. Ez viszont ellentmond a (2.147) tulajdonságnak, hiszen $(x_1,t_1) \notin \Gamma_{t^\star}$. Tehát a (2.145) indirekt feltevésünk nem érvényes, és ezzel beláttuk a (2.144) állításunk első részét. A második (minimumra vonatkozó) állítás közvetlen adódik az maximumra vonatkozó állítás $-w(x,t)$ függvényre történő alkalmazásával. □

A lemma állítása alapján tehát a (2.142) egyenletet kielégítő, az $\overline{\Omega}_{t^\star}$ halmazon folytonos $w(x,t)$ függvényre alsó és felső becslés adható a parabolikus peremen felvett értékei segítségével. Nevezetesen, érvényes a

$$\min_{\Gamma_{t^\star}} w(x,t) \leq w(x,t) \leq \max_{\Gamma_{t^\star}} w(x,t), \quad (x,t) \in \overline{\Omega}_{t^\star} \tag{2.149}$$

becslés.

A 2.8. lemmának több fontos következménye is van. Ugyanis

2.9. Következmény A vizsgált (2.142)-(2.143) feladatnak *csak egy megoldása van.* Tegyük fel, hogy az $u_1(x,t)$ és $u_2(x,t)$ függvények egyaránt megoldásai a feladatnak, és $u_1 \neq u_2$. Ekkor a $w(x,t) = u_1(x,t) - u_2(x,t)$ nem az azonosan nulla függvény kielégíti a (2.142) egyenletet, és emellett a $\Gamma_{t^\star}$ parabolikus peremen azonosan nulla. Ezért a lemma értelmében

$$0 = \min_{\Gamma_{t^\star}} w(x,t) \leq w(x,t) \leq \max_{\Gamma_{t^\star}} w(x,t) = 0$$

minden $(x,t) \in \overline{\Omega}_{t^\star}$ pontban, azaz $w(x,t) = 0$ az $\overline{\Omega}_{t^\star}$ halmazon, ami ellentmondás, így az állításunk igaz.

2.10. Következmény A vizsgált (2.142)-(2.143) feladatra érvényes a *nemnegativitás*: a nemnegatív bemenő adatokhoz tartozó megoldások is nemnegatívak, azaz $u_0(x) \geq 0$, $\alpha, \beta \geq 0$ esetén $u(x,t) \geq 0$. Ez az állítás közvetlenül nyerhető a 2.8.. lemmából a $\min_{\Gamma_{t^\star}} u(x,t) \geq 0$ egyenlőtlenség alapján.

2.11. Következmény A vizsgált (2.142)-(2.143) feladatra érvényes a *monotonitás*: a nem kisebb bemenő adatokhoz tartozó megoldások is nem kisebbek. Legyenek ugyanis az $i = 1, 2$ indexekre az $u_i(x,t)$ függvények a (2.142) egyenlet megoldásai az

$$u_i(x,0) = u_0^i(x), \quad x \in (0,1); \quad u_i(0,t) = \alpha^i, \quad u_i(1,t) = \beta^i$$

kiegészítő feltételekkel, és tegyük fel, hogy $u_0^1(x) \geq u_0^2(x)$, $\alpha_1 \geq \alpha_2$ és $\beta_1 \geq \beta_2$. Ekkor a $w(x,t) = u_1(x,t) - u_2(x,t)$ függvény egyrészt szintén kielégíti a (2.142) egyenletet, másrészt pedig $w(x,0) \geq 0$, $w(0,t) \geq 0$ és $w(1,t) \geq 0$. Ezért állításunk a nemnegativitásra vonatkozó 2.10. megjegyzés közvetlen következménye.

2.12. Következmény A (2.142)-(2.143) feladat *stabil kitűzésű a maximumnormában*: a bemenő adatok (azaz u_0, α és β) kis megváltoztatása a megoldást is csak kicsit változtatja meg a maximumnormában. Ugyanis, ha (2.143) helyett a

$$\tilde{u}(x,0) = \tilde{u}_0(x), \quad x \in (0,1); \quad \tilde{u}(0,t) = \tilde{\alpha}, \quad \tilde{u}(1,t) = \tilde{\beta}, \quad t \in [0, t^\star]$$

feltételekkel oldjuk meg a (2.142) egyenletet, akkor a 2.8. lemma alapján

$$\min_{\Gamma_{t^\star}}(u(x,t) - \tilde{u}(x,t)) \leq u(x,t) - \tilde{u}(x,t) \leq \max_{\Gamma_{t^\star}}(u(x,t) - \tilde{u}(x,t)),$$

azaz

$$\max_{\Gamma_{t^\star}}(-(u(x,t) - \tilde{u}(x,t))) \leq u(x,t) - \tilde{u}(x,t) \leq \max_{\Gamma_{t^\star}}(u(x,t) - \tilde{u}(x,t)).$$

Ebből már közvetlenül adódik a kívánt

$$\|u - \tilde{u}\|_{C(\overline{\Omega})} \leq \|u - \tilde{u}\|_{C(\Gamma_{t^\star})} \tag{2.150}$$

becslés.

Tehát a (2.142)-(2.143) feladatnak

- a 2.7. megjegyzés alapján létezik megoldása (egzisztencia);
- a 2.9. megjegyzés alapján ez a megoldás egyetlen (unicitás);
- a 2.12. megjegyzés alapján ez a megoldás a maximumnormában folytonosan függ a bemenő adatoktól (stabilitás).

Ezért ez a feladat korrekt kitűzésű.

A hővezetési egyenlet szemidiszkretizációja

Az előzőekben beláttuk, hogy a (2.142)-(2.143) feladat korrekt kitűzésű, azaz létezik egyetlen, a bemenő adatoktól folytonosan függő megoldása. A megoldás előállítására ugyan megadtunk egy eljárást, a 2.7. megjegyzésben,de ez az előállítás a gyakorlatban nem hatékony, és általában "pontosan" nem is határozható meg, továbbá, hogy egy általánosabb feladatosztályon nem is alkalmazható.

Ezért numerikus módszert választunk, amelynek lényege hogy két lépésben a folytonos feladatot diszkrét feladatok sorozatára vezetjük vissza. Ez a két lépés az alábbiakból áll:

- Első lépésben térbeli diszkretizációt hajtunk végre, amelynek eredményeként időben folytonos közönséges differenciálegyenlet-rendszerek Cauchy-feladatainak sorozatát kapjuk.
- Második lépésben a fenti Cauchy-feladatokat időben diszkretizáljuk, és numerikusan megoldva előállítjuk a térben és időben egyaránt diszkretizált megoldást.

Ebben a szakaszban az első lépéssel foglalkozunk részletesebben, hiszen a további fejezetek kizárólag majd a második lépésről szólnak. Ez az első lépés az ún. szemidiszkretizáció, amely a (2.142)-(2.143) feladat esetén a következő lépések sorozatát jelenti.

- Kijelölünk a $[0,1]$ intervallumon egy $\omega_h = \{x_j = jh,\ h = 1/(N+1),\ j = 0,1,\ldots,N+1\}$ rácshálót, ahol N egy adott természetes szám.
- Mindegyik x_j csomóponthoz hozzárendelünk egy $y_j(t)$ függvényt, amelyről azt szeretnénk, hogy a (2.142)-(2.143) feladat megoldásának jó közelítése legyen az x_j pontban, azaz $y_j(t) \approx u(x_j,t)$.
- A vizsgált feladatból kiindulva megadjuk azt a szabályt, amely alapján meghatározhatók a fenti elvárásnak eleget tevő $y_j(t)$ függvények.

Ezt a módszert az *egyenesek módszerének (method of lines, MOL)* szokásos nevezni. A (2.143) peremfeltételei alapján az $N+1$ darab ismeretlen függvény közül az $x_0 = 0$ és az $x_{N+1} = 1$ pontokhoz tartozó függvényeket pontosan ismerjük, azaz $t \in [0, t^\star]$ esetén

$$y_0(t) = u(0,t) = \alpha, \quad y_{N+1}(t) = u(1,t) = \beta. \tag{2.151}$$

Másrészt, ugyancsak a (2.143)-beli kezdeti feltétele alapján minden rácspontban ismerjük a $t = 0$ értékhez tartozó pontos értéket, azaz

$$y_j(0) = u_0(x_j), \quad j = 1, 2, \dots N. \tag{2.152}$$

Írjuk fel a (2.142) egyenletet az $x = x_j$ pontban:

$$\frac{\partial u}{\partial t}(x_j, t) - \frac{\partial^2 u}{\partial x^2}(x_j, t) = 0, \quad t \in [0, t^\star], \quad j = 1, 2, \dots, N. \tag{2.153}$$

Ezért, ha $y_j(t)$ jó közelítése az $u(x_j, t)$ függvénynek, akkor a jól ismert véges differenciás közelítésekkel

$$\begin{aligned} \frac{\partial u}{\partial t}(x_j, t) &\approx y_j'(t) \\ \frac{\partial^2 u}{\partial x^2}(x_j, t) &\approx \frac{y_{j+1}(t) - 2y_j(t) + y_{j-1}(t)}{h^2} \end{aligned}$$

approximációkat eredményezi.

Ezek alapján tehát az az elvárásunk, hogy az $y_j(t)$ ismeretlen függvények megoldásai legyenek az

$$\begin{aligned} y_j'(t) - \frac{y_{j+1}(t) - 2y_j(t) + y_{j-1}(t)}{h^2} &= 0, \quad t \in [0, t^\star], \\ y_j(0) &= u_0(x_j) \end{aligned} \tag{2.154}$$

(ahol $j = 1, 2, \dots, N$, $y_0(t)$ és $y_{N+1}(t)$ adottak) közönséges differenciálegyenlet-rendszer Cauchy feladatának. Ekkor az $\mathbf{y}_h : [0, t^\star] \to \mathbb{R}^N$, y_j koordináta-függvényű ismeretlenre a (2.154) feladat felírható

$$\begin{aligned} \mathbf{y}_h'(t) - \mathbf{A}_h \mathbf{y}_h(t) &= \mathbf{f}_h(t), \quad t \in [0, t^\star], \\ \mathbf{y}_h(0) &= \mathbf{u}_{0,h} \end{aligned} \tag{2.155}$$

alakban, ahol

$$\begin{aligned} \mathbf{A}_h = \frac{1}{h^2} \operatorname{tridiag}[1, -2, 1] \in \mathbb{R}^{N \times N}, \quad \mathbf{f}_h(t) = \frac{1}{h^2}[\alpha, 0, \dots, 0, \beta]^T \in \mathbb{R}^N, \\ \mathbf{u}_{0,h} = [u_0(x_1), u_0(x_2), \dots, u_0(x_N)]^T \in \mathbb{R}^N. \end{aligned} \tag{2.156}$$

Alkalmazva a szokásos

$$\exp(t\mathbf{A}_h) = \mathbf{I} + t\mathbf{A}_h + \frac{t^2}{2!}\mathbf{A}_h^2 + \cdots + \frac{t^n}{n!}\mathbf{A}_h^n + \ldots \tag{2.157}$$

azonosságot, a (2.155) Cauchy-feladat megoldása felírható az

$$\mathbf{y}_h(t) = \exp(t\mathbf{A}_h)\mathbf{u}_{0,h} + \int_0^t \exp((t-s)\mathbf{A}_h)\mathbf{f}_h(s)\,\mathrm{d}s \tag{2.158}$$

alakban.[19]

A szemidiszkretizációs mátrixsereg tulajdonságai

A (2.120)-(2.121) feladat felírható

$$\frac{\partial u}{\partial t}(x,t) - (\mathcal{A}u)(x,t) = 0, \quad u(x,0) = u_0(x) \tag{2.159}$$

alakban, ahol $\mathcal{A}$ egy olyan operátor, amely az olyan, megfelelően sima $w(x,t)$ függvényeken van értelmezve, amelyek az $x = 0$ és az $x = 1$ pontokban nulla értéket vesznek fel, és

$$(\mathcal{A}w)(x,t) = \frac{\partial^2 w}{\partial x^2}(x,t). \tag{2.160}$$

Összevetve a fenti (2.159) feladatot a szemidiszkretizáció (2.155) alakjával, természetes elvárás, hogy az $\mathbf{A}_h$ mátrixsereg kis h esetén jól közelítse az $\mathcal{A}$ operátort.

A továbbiakban $\mathbf{A}_h$ mátrix ezen approximációs viselkedését, illetve néhány további tulajdonságát vizsgáljuk.

Először a spektrum approximációjával foglalkozunk. Arra keressük a választ, hogy az $\mathbf{A}_h$ mátrix sajátértékei és sajátvektorai vajon jól közelítik-e az $\mathcal{A}$ operátor sajátértékeit és sajátvektorait. A változók szétválasztásának módszerével már korábban megmutattuk, hogy az $\mathcal{A}$ operátor sajátértékei és sajátvektorai

$$\lambda_k(\mathcal{A}) = \lambda_k = -k^2\pi^2, \quad u_k^{\mathcal{A}}(x) = u_k(x) = \sin(k\pi x), \quad k = 1, 2, \ldots \tag{2.161}$$

alakúak.

Határozzuk meg az $\mathbf{A}_h \in \mathbb{R}^{N\times N}$ mátrix sajátértékeit és sajátvektorait! Legyenek v_j ($j = 0, 1, \ldots, N+1$) olyan számok, amelyekre $v_0 = v_{N+1} = 0$, és valamely $\lambda^{(h)} \in \mathbb{R}$ valós szám mellett teljesülnek a

$$\frac{v_{j-1} - 2v_j + v_{j+1}}{h^2} = \lambda^{(h)} v_j, \quad j = 1, 2, \ldots, N$$

[19] Egy vektorértékű függvény integrálján azt a vektort értjük, amelynek koordinátái a koordinátafüggvények integráljai.

egyenlőségek. Ekkor nyilvánvalóan a $\mathbf{v}^{(h)} = [v_1, v_2, \ldots, v_N]^T \in \mathbb{R}^N$ vektor sajátvektora az $\mathbf{A}_h$ mátrixnak a $\lambda^{(h)}$ sajátértékkel. A fenti feltételek ekvivalensek a

$$v_{j-1} - 2(1 + 0.5\lambda^{(h)}h^2)v_j + v_{j+1} = 0, \quad j = 1, 2, \ldots, N \tag{2.162}$$

lineáris algebrai egyenletrendszer megoldásával, ahol $v_0 = v_{N+1} = 0$ adottak. Keressük a (2.162) megoldását a $v_j = \sin(\gamma x_j) = \sin(\gamma jh)$ $(j = 0, 1, \ldots, N + 1)$ alakban, ahol γ egyelőre ismeretlen valós szám. Helyettesítsük be ezt az alakot az egyenletrendszerbe, és használjuk fel a nyilvánvaló

$$v_{j-1} + v_{j+1} = \sin(\gamma(j-1)h) + \sin(\gamma(j+1)h) = 2\sin(\gamma jh)\cos(\gamma h)$$

összefüggést! Ekkor a (2.162) feladatból a

$$2\left[\cos(\gamma h) - (1 + 0.5\lambda^{(h)}h^2)\right] v_j = 0, \quad j = 1, 2, \ldots, N \tag{2.163}$$

feladatot kapjuk. Mivel mindegyik v_j nem lehet nulla, ezért tehát $\cos(\gamma h) - (1 + 0.5\lambda^{(h)}h^2) = 0$, azaz

$$\lambda^{(h)} = \frac{-1 + \cos(\gamma h)}{0.5h^2} = -\frac{2}{h^2}\left(1 - \cos(\gamma h)\right) = -\frac{4}{h^2}\sin^2\frac{\gamma h}{2}. \tag{2.164}$$

A tetszőleges γ paramétert úgy kell megválasztani, hogy a $v_0 = v_{N+1} = 0$ feltételek is teljesüljenek v_j fenti előállításában. A $v_0 = \sin(0\gamma) = 0$ egyenlőség minden γ esetén igaz. Mivel $v_{N+1} = \sin(\gamma x_{N+1}) = \sin(\gamma 1) = \sin\gamma = 0$, ezért tehát a megfelelő paraméterválasztás $\gamma = \gamma_k = k\pi$, ahol $k = 1, 2, \ldots$. Így az $\mathbf{A}_h$ $N \times N$-es mátrix sajátértékei a következők:

$$\lambda_k^{(h)} = -\frac{4}{h^2}\sin^2\frac{k\pi h}{2} = -\frac{4}{h^2}\sin^2\frac{\pi x_k}{2}, \quad k = 1, 2, \ldots, N. \tag{2.165}$$

Az $\mathbf{A}_h$ mátrix $\lambda_k^{(h)}$ sajátértékéhez tartozó sajátvektora az a $\mathbf{v}_k^{(h)} \in \mathbb{R}^N$ vektor, amelynek j-edik koordinátája

$$\left(\mathbf{v}_k^{(h)}\right)_j = \sin(k\pi jh) = \sin(k\pi x_j), \quad j = 1, 2, \ldots, N. \tag{2.166}$$

A (2.165) összefüggés alapján beláthatóak az $\mathbf{A}_h$ mátrix sajátértékeinek és sajátvektorainak néhány fontos tulajdonsága.

1. Tetszőleges N esetén $\lambda_k^{(h)} < 0$ és értékei k-ban szigorúan monoton csökkennek. (Ez közvetlenül leolvasható a képletből.)

2. Mivel az előző monotonitási tulajdonság alapján $\lambda_{max}^{(h)} = \lambda_1^{(h)}$, ezért a mátrix maximális sajátértékének elhelyezkedésére becslés adható. Nevezetesen, bevezetve a

$\varsigma = \pi h/2$ jelölést, $\lambda_{max}^{(h)} = -\pi^2(\sin\varsigma/\varsigma)^2$ alakot ölti, ahol a $h \geq 0.5$ természetes feltételezés következtében $\varsigma \in (0, \pi/4]$. Innen közvetlenül nyerjük a

$$-\pi^2 \leq \lambda_{max}^{(h)} \leq -\pi^2 \left(\sin\varsigma/\varsigma\right)^2_{\varsigma=\pi/4} = -8 \tag{2.167}$$

becslést.[20]

3. Az előző tulajdonság alapján $\lambda_{max}^{(h)} \in (-\pi^2, -8)$, ezért tetszőleges $h > 0$ esetén az $\mathbf{A}_h$ mátrix minden sajátértéke a negatív valós tengelyen helyezkedik el, és a nulla nem sajátértéke. Ezért az $\mathbf{A}_h$ mátrixok mindegyike reguláris.

4. Vizsgáljuk meg a sajátvektor approximácós tulajdonságát! Mivel az $x = x_j$ pontban az $\mathcal{A}$ operátor sajátfüggvénye $u_k(x_j) = \sin(k\pi x_j)$ (ld. (2.161)), ezt összevetve az $\mathbf{A}_h$ mátrix $\mathbf{v}_k^{(h)}$ k-adik sajátvektorának j-edik koordinátájával (ld. (2.166)), jól látható a sajátvektorok approximációja.

5. Mivel

$$\lim_{h\to 0} \lambda_k^{(h)} = \lim_{h\to 0} -\frac{4}{h^2}\sin^2\frac{\pi k h}{2} = -\lim_{s\to 0}\frac{\sin^2 \pi k s}{s^2} = -k^2\pi^2 = \lambda_k, \tag{2.168}$$

ezért tehát $h \to 0$ esetén $\mathbf{A}_h$ mátrix k-adik sajátértéke tart az $\mathcal{A}$ differenciáloperátor k-adik sajátértékéhez. (Tehát $\lambda_{max}^{(h)}$ a $(-\pi^2, -8)$ intervallumon jobbról tart az $\mathcal{A}$ operátor $-\pi^2$ maximális sajátértékéhez.)

6. Az $\mathbf{A}_h$ mátrix negatív definit. Ez a tulajdonság közvetlenül adódik abból az ismert lineáris algebrai állításból, hogy egy szimmetrikus valós mátrix pontosan akkor pozitív definit, amikor minden sajátértéke pozitív. (Pl. [16].)[21]

7. Vizsgáljuk meg a (2.155) állandó együtthatós lineáris közönséges differenciálegyenlet-rendszer merevségét! Az $\mathbf{A}_h$ mindegyik sajátértéke valós és negatív. Így a 2.22. definíció alapján a rendszer merevségi mutatója

$$S(h) = \frac{-\lambda_N^{(h)}}{-\lambda_1^{(h)}} \tag{2.169}$$

[20] Felhasználtuk, hogy a $h(\varsigma) = \sin\varsigma/\varsigma$ függvény a $(0, \pi/4]$ intervallumon szigorúan monoton csökken, tehát maximumát a $\varsigma = \pi/4$ pontban veszi fel. Emellett, $\lim_{\varsigma\to 0} h(\varsigma) = 1$.

[21] Ez a jól ismert tétel közvetlenül belátható. Tetszőleges $\mathbf{B}$ szimmetrikus mátrix sajátértékei valósak, valamint a sajátvektorai ortogonálisak, ezért a sajátvektoraiból álló $\mathbf{Q}$ mátrix ortogonális (azaz $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$), és $\mathbf{B} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$. Ezért tetszőleges $\mathbf{v}$ vektor esetén $(\mathbf{B}\mathbf{v}, \mathbf{v}) = (\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\mathbf{v}, \mathbf{v}) = (\mathbf{\Lambda}\mathbf{Q}^T\mathbf{v}, \mathbf{Q}^T\mathbf{v}) = (\mathbf{\Lambda}\mathbf{y}, \mathbf{y})$ az $\mathbf{y} = \mathbf{Q}^T\mathbf{v}$ jelöléssel. Ez utóbbi pedig az állításunkat igazolja.

Figyelembe véve a sajátértékek (2.165) értékeit, ekkor

$$S(h) = \frac{\sin^2(\pi N h/2)}{\sin^2(\pi h/2)} = \left(\frac{\sin(1-h)\pi/2}{\sin(\pi h/2)}\right)^2 = \left(\frac{\cos(\pi h/2)}{\sin(\pi h/2)}\right)^2 = \operatorname{ctg}^2(\pi h/2)\,. \tag{2.170}$$

Ezért tehát

$$S(h) = \mathcal{O}(1/h^2) \gg 1.^{22}$$

Ez azt jelenti, hogy a (2.155) merev rendszert alkot, és így a numerikus megoldása, azaz a teljes diszkretizáció előállítása, speciális eljárást igényel.

8. Mivel a szimmetrikus $\mathbf{A}_h \in \mathbb{R}^{N\times N}$ mátrix negatív definit, így a $\mathbf{v}_k^{(h)} \in \mathbb{R}^N$ $(k = 1, 2, \ldots, N)$ a (2.166) alakú sajátvektor-rendszere ortogonális bázist alkot $\mathbb{R}^N$-ben. Ezért a skaláris szorzatukra

$$(\mathbf{v}_j^{(h)}, \mathbf{v}_k^{(h)}) = \begin{cases} 0 & \text{amikor } j \neq k; \\ \|\mathbf{v}_k^{(h)}\|_2^2, & \text{amikor } j = k, \end{cases} \tag{2.171}$$

és

$$\|\mathbf{v}_k^{(h)}\|_2^2 = \sum_{j=1}^{N} (\mathbf{v}_k^{(h)})_j^2 = \sum_{j=1}^{N} \sin^2(jk\pi h).$$

Határozzuk meg $\|\mathbf{v}_k^{(h)}\|_2^2$ értékét, azaz adjunk zárt képletet a $\sum_{j=1}^{N} \sin^2(jk\pi h)$ összegre! Megmutatható[23], hogy

$$\sum_{j=0}^{N} \cos(jx) = \frac{\sin\frac{(N+1)x}{2}\cos\frac{Nx}{2}}{\sin\frac{x}{2}}. \tag{2.172}$$

---

[22] Ez a jól ismert és a L'Hospital szabállyal könnyen belátható $\lim_{x\to 0} \dfrac{\operatorname{ctg} x}{1/x} = 1$ határértékből következik.

[23] A képlet belátásához alkalmazzuk a $\sum_{j=0}^{N} \left(e^{ix}\right)^j = \sum_{j=0}^{N} e^{ixj} = \dfrac{e^{i(N+1)x}-1}{e^{ix}-1} = \dfrac{(\cos(N+1)x - 1) + i\sin(N+1)x}{(\cos x - 1) + i\sin x}$ összefüggést. Mivel $e^{ixj} = \cos(jx) + i\sin(jx)$, ezért tehát $\sum_{j=0}^{N} \cos(jx) = Re\dfrac{(\cos(N+1)x - 1) + i\sin(N+1)x}{(\cos x - 1) + i\sin x}$. Néhány elemi trigonometrikus azonosság felhasználásával az utolsó kifejezés a (2.172) összefüggést jelenti. (Pfeil Tamás személyes közlése alapján.)

Mivel $\sin^2(jx) = (1-\cos(2jx))/2$, ezért felhasználva a (2.172) képletet, érvényes a

$$\sum_{j=1}^{N} \sin^2(jx) = \sum_{j=1}^{N} \frac{1-\cos(2jx)}{2} = \frac{N}{2} - \frac{1}{2}\sum_{j=1}^{N} \cos(2jx) = \\ \frac{N}{2} - \frac{1}{2}\left[\sum_{j=0}^{N} \cos(2jx) - 1\right] = \frac{N+1}{2} - \frac{\sin(N+1)x\cos(Nx)}{2\sin x} \tag{2.173}$$

összefüggés.

Helyettesítsük be a (2.173) képletbe az $x = k\pi h$ értéket! Ekkor

$$\sum_{j=1}^{N} \sin^2(jk\pi h) = \frac{N+1}{2} - \frac{\sin(N+1)k\pi h\cos(Nk\pi h)}{2\sin(k\pi h)}. \tag{2.174}$$

Mivel $(N+1)h = 1$, ezért a (2.174) képlet alapján

$$\sum_{j=1}^{N} \sin^2(jk\pi h) = \frac{N+1}{2} - \frac{\sin(k\pi)\cos(Nk\pi h)}{2\sin(k\pi h)}.$$

Ezért a $\sin(k\pi) = 0$ következtében

$$\sum_{j=1}^{N} \sin^2(jk\pi h) = \frac{N+1}{2}. \tag{2.175}$$

A (2.175) összefüggés alapján tehát

$$\|\mathbf{v}_k^{(h)}\|_2^2 = \frac{N+1}{2}. \tag{2.176}$$

9. Az $\exp(t\mathbf{A}_h)$ mátrix elemei tetszőleges $t \geq 0$ esetén nemnegatívak, azaz $\exp(t\mathbf{A}_h) \geq 0$.

2.13. Megjegyzés Az utolsó tulajdonság igazolásához bizonyítsunk be egy általánosabb lemmát.(Ld. pl. [2].)

2.14. Lemma *Legyen* $\boldsymbol{B}$ *egy tetszőleges mátrix. Ekkor* $\exp(t\boldsymbol{B}) \geq 0$ *minden* $t \geq 0$ *esetén, amikor a* $\boldsymbol{B}$ *mátrix főátlóján kívüli elemei nemnegatívak, azaz* $\operatorname{offdiag} \boldsymbol{B} \geq 0$.

Bizonyítás. A *szükségesség* az $\exp(t\mathbf{B})$ mátrix (2.157) alakú előállításából közvetlenül látszik, ugyanis kis t értékek esetén az $\exp(t\mathbf{B}) \approx \mathbf{I} + t\mathbf{B} + \mathcal{O}(t^2)$, így az exponenciális

mátrix diagonálison kívüli elemeinek előjelét a $\mathbf{B}$ mátrix ezen elemeinek előjele határozza meg.

Az *elégségesség* belátásához vegyük észre, hogy valamely $\mathbf{C}$ és $\mathbf{D}$ mátrixok esetén $\exp t(\mathbf{C}+\mathbf{D}) = \exp(t\mathbf{C})\exp(t\mathbf{D})$ pontosan akkor, amikor a $\mathbf{C}$ és $\mathbf{D}$ mátrixok kommutálnak, azaz $\mathbf{C}\cdot\mathbf{D} = \mathbf{D}\cdot\mathbf{C}$.[24] Mivel az $\mathbf{I}\in\mathbb{R}^{N\times N}$ egységmátrix és az $\mathbf{A}_h$ mátrix nyilvánvalóan kommutálnak, ezért tetszőleges $c\in\mathbb{R}$ szám esetén

$$\exp(t\mathbf{B}) = \exp(t(c\mathbf{I}+\mathbf{B}) - tc\mathbf{I}) = \exp(t(c\mathbf{I}+\mathbf{B}))\cdot\exp(-tc\mathbf{I}). \tag{2.177}$$

Az $\exp(-tc\mathbf{I})$ egy olyan diagonális mátrix, amelynek főátlójában az $exp(-tc)$ számok állnak, azaz $\exp(-tc\mathbf{I}) = \exp(-tc)\mathbf{I}$. Ezért tehát tetszőleges c esetén $\exp(-tc\mathbf{I})\geq 0$. Válasszuk meg a c számot úgy, hogy $\mathbf{B}+c\mathbf{I}\geq 0$ legyen. (Azaz legyen $c\geq\max_{j=1,2,\dots,N}|\mathbf{B}_{ii}|$.) Ekkor $\exp(t(c\mathbf{I}+\mathbf{B}))\geq 0$, hiszen a (2.157) típusú sorban minden elem nemnegatív. Ezért tehát (2.177) jobb oldalán lévő szorzatban mindkét mátrix nemnegatív, és ez a lemma állítását bizonyítja. □

Mivel (2.156) alapján offdiag $\mathbf{A}_h\geq 0$, ezért a 2.14.. lemma alapján tetszőleges $t\geq 0$ esetén érvényes az

$$\exp(t\mathbf{A}_h)\geq 0 \tag{2.178}$$

nemnegativitási tulajdonság.

A szemidiszkretizáció approximációja

Ebben a szakaszban először formálisan, majd precízen megmutatjuk, hogy a szemidiszkretizációval előállított megoldás a rácspontokban valóban jól közelíti a hővezetési egyenlet megoldását.

Az általánosság megszorítása nélkül feltesszük, hogy $\alpha=\beta=0$, azaz a homogén peremfeltételű (2.120)-(2.121) feladat $u(x,t)$ megoldására mutatjuk meg, hogy az $x_i\in\omega_h$ pontokban az $u(x_i,t)$ függvényt a megfelelő szemidiszkrét feladat megoldásának i-edik koordináta-függvénye jól közelíti.

Emlékeztetőül: korábban megmutattuk (ld. (2.138)), hogy a (2.120)-(2.121) feladat megoldása felírható

$$u(x,t) = \sum_{k=1}^{\infty} u_0^k e^{\lambda_k t} u_k(x) \tag{2.179}$$

alakban, ahol $u_k(x)=\sin k\pi x$, $\lambda_k=-k^2\pi^2$ és

$$u_0^k = 2\int_0^1 u_0(s)\sin(k\pi s)\,\mathrm{d}s. \tag{2.180}$$

[24] Ez az egyes exponenciálisok (2.157) alakú előállításából, majd a sorok összeszorzásával közvetlenül leolvasható.

A homogén peremfeltétel következtében $\mathbf{f}_h(t) = 0$, azaz a szemidiszkrét feladat (2.155) alapján

$$\begin{aligned} \mathbf{y}_h'(t) - \mathbf{A}_h \mathbf{y}_h(t) &= \mathbf{0}, \quad t \in [0, t^\star], \\ \mathbf{y}_h(0) &= \mathbf{u}_{0,h} \end{aligned} \tag{2.181}$$

alakú, ahol a (2.156) jelöléseket használtuk. Bár ezen feladat megoldása felírható az

$$\mathbf{y}_h(t) = \exp(t\mathbf{A}_h)\mathbf{u}_{0,h} \tag{2.182}$$

alakban, ennek kiszámolása a gyakorlatban nem lehetséges.

Állítsuk elő a szemidiszkrét megoldást más, a gyakorlatban is realizálható alakban! Tetszőleges rögzített $t \in [0, t^\star]$ pontban $\mathbf{y}_h(t) \in \mathbb{R}^N$. Mivel az $\mathbf{A}_h$ mátrix $\mathbf{v}_k^{(h)}$ $(k = 1, 2, \ldots, N)$ sajátvektorai bázist alkotnak $\mathbb{R}^N$ térben, ezért

$$\mathbf{y}_h(t) = \sum_{k=1}^{N} c_k^{(h)}(t)\mathbf{v}_k^{(h)}. \tag{2.183}$$

ahol $c_k^{(h)}(t)$ ismeretlen függvények. Nyilvánvalóan ekkor $\mathbf{y}_h'(t) = \sum_{k=1}^{N} c_k^{(h)\prime}(t)\mathbf{v}_k^{(h)}$. Behelyettesítve a fenti alakokat a (2.181) egyenletbe, az ismeretlen együtthatófüggvényre a tetszőleges $t \in [0, t^\star]$ esetén a következő összefüggést kapjuk:[25]

$$\begin{aligned} \mathbf{0} =& \mathbf{y}_h'(t) - \mathbf{A}_h \mathbf{y}_h(t) = \sum_{k=1}^{N} c_k'(t)\mathbf{v}_k - \mathbf{A}_h \left(\sum_{k=1}^{N} c_k(t)\mathbf{v}_k \right) = \\ & \sum_{k=1}^{N} c_k'(t)\mathbf{v}_k - \sum_{k=1}^{N} c_k(t) \left(\mathbf{A}_h \mathbf{v}_k \right) = \sum_{k=1}^{N} c_k'(t)\mathbf{v}_k - \\ & \sum_{k=1}^{N} c_k(t) \left(\lambda_k^{(h)} \mathbf{v}_k \right) = \sum_{k=1}^{N} \left(c_k'(t) - \lambda_k^{(h)} c_k(t) \right) \mathbf{v}_k. \end{aligned} \tag{2.184}$$

Ezért tehát minden $k = 1, 2, \ldots, N$ esetén $c_k'(t) - \lambda_k^{(h)} c_k(t) = 0$, azaz

$$c_k(t) = e^{\lambda_k^{(h)} t} c_k(0), \quad t \in [0, t^\star]. \tag{2.185}$$

Határozzuk meg a $c_k(0)$ értékeit! A (2.183) és a (2.181) feladat kezdeti feltétele alapján

$$\mathbf{u}_{0,h} = \mathbf{y}_h(0) = \sum_{j=1}^{N} c_j(0)\mathbf{v}_j.$$

[25] A továbbiakban a könnyebb írásmód kedvéért elhagyjuk a (h) felsőindex jelölést a $\mathbf{v}_k^{(h)}$ vektorra és a $c_k^{(h)}(t)$ függvényre.

Ezt skalárisan megszorozva a $\mathbf{v}_k$ vektorral és felhasználva a bázis ortogonalitását és a (2.176) összefüggést, a

$$(\mathbf{u}_{0,h}, \mathbf{v}_k) = \sum_{j=1}^{N} c_j(0)(\mathbf{v}_j, \mathbf{v}_k) = c_k(0)(\mathbf{v}_k, \mathbf{v}_k) = \frac{N+1}{2} c_k(0)$$

összefüggést kapjuk. Innen tehát

$$c_k(0) = \frac{2}{N+1}(\mathbf{u}_{0,h}, \mathbf{v}_k) = 2h(\mathbf{u}_{0,h}, \mathbf{v}_k). \tag{2.186}$$

Ezt behelyettesítve a (2.185) képletbe, a

$$c_k(t) = e^{\lambda_k^{(h)} t} 2h(\mathbf{u}_{0,h}, \mathbf{v}_k), \quad t \in [0, t^\star] \tag{2.187}$$

összefüggést nyerjük. A (2.183) és a (2.187) képletek együttesen a szemidiszkrét megoldást eredményezik:

$$\mathbf{y}_h(t) = \sum_{k=1}^{N} 2h(\mathbf{u}_{0,h}, \mathbf{v}_k) e^{\lambda_k^{(h)} t} \mathbf{v}_k. \tag{2.188}$$

Célunk megmutatni, hogy megfelelően kis h esetén a (2.188) szerinti $\mathbf{y}_h(t)$ vektorfüggvény j-edik koordinátafüggvénye jól közelíti a (2.179) megoldását az $x = x_j$ pontban, azaz az

$$\left(\mathbf{y}_h(t)\right)_j = \sum_{k=1}^{N} 2h(\mathbf{u}_{0,h}, \mathbf{v}_k) e^{\lambda_k^{(h)} t} \left(\mathbf{v}_k\right)_j \tag{2.189}$$

függvény jól közelíti a

$$u(x_j, t) = \sum_{k=1}^{\infty} u_0^k e^{\lambda_k t} u_k(x_j) \tag{2.190}$$

függvényt.

Mivel $\left(\mathbf{v}_k\right)_j = u_k(x_j) = \sin(k\pi jh)$, ezért a határátmenet formális elvégzésével látható ez a tulajdonság, ugyanis

$$\begin{aligned} &\lim_{h\to 0} \mathbf{y}_h(t) = \lim_{h\to 0} \sum_{k=1}^{N} 2h(\mathbf{u}_{0,h}, \mathbf{v}_k) e^{\lambda_k^{(h)} t} \mathbf{v}_k = \\ &\sum_{k=1}^{\infty} \underbrace{\left(\lim_{h\to 0} 2h(\mathbf{u}_{0,h}, \mathbf{v}_k)\right)}_{S_1} \underbrace{\left(\lim_{h\to 0} e^{\lambda_k^{(h)} t}\right)}_{S_2} \underbrace{\left(\lim_{h\to 0} \mathbf{v}_k\right)}_{S_3}. \end{aligned} \tag{2.191}$$

Számítsuk ki a szummában lévő határértékeket!

- Mivel $(\mathbf{u}_{0,h}, \mathbf{v}_k) = \sum_{j=1}^{N} u_0(x_j)\sin(k\pi x_j)$, ezért $2h(\mathbf{u}_{0,h}, \mathbf{v}_k)$ egy Riemann-féle közelítő összege a h lépésközű, ekvidisztáns felosztáson a $2\int_0^1 u_0(s)\sin(k\pi s)\,\mathrm{d}s$ integrálnak, azaz a (2.180) képlet alapján $S_1 = u_0^k$.

- Az $\mathbf{A}_h$ mátrix sajátértékeinek előzőekben megmutatott konvergenciája miatt $\lim_{h\to 0} e^{\lambda_k^{(h)} t} = e^{\lambda_k t}$, azaz $S_2 = e^{\lambda_k t}$.

- A konstrukcióból következően $\lim_{h\to 0} \mathbf{v}_k = u_k(x)$, azaz $S_3 = u_k(x)$.

Ezért (2.191) alapján $\lim_{h\to 0} \mathbf{y}_h(t) = \sum_{k=1}^{\infty} u_0^k e^{\lambda_k t} u_k(x)$, ami a (2.179) képlet alapján a kívánt konvergenciát jelenti.

A továbbiakban egy szigorú bizonyítást adunk a konvergenciára, és egyben becslést is adunk a konvergencia rendjére is! Tekintsük a következő elsőrendű, inhomogén jobb oldalú elsőrendű közönséges differenciálegyenlet-rendszert:

$$\begin{aligned} \mathbf{w}_h'(t) - \mathbf{A}_h \mathbf{w}_h(t) &= \mathbf{g}_h(t), \quad t \in [0, t^\star], \\ \mathbf{w}_h(0) &= \mathbf{u}_{0,h}, \end{aligned} \tag{2.192}$$

és állítsuk elő a megoldását a homogén esetre alkalmazott módszerrel! Mivel minden rögzített $t \in [0, t^\star]$ esetén $\mathbf{g}_h(t) \in \mathbb{R}^N$, ezért előállítható az

$$\mathbf{g}_h(t) = \sum_{k=1}^{N} g_k(t)\mathbf{v}_k \tag{2.193}$$

alakban. Ezt a kifejezést skalárisan megszorozva $\mathbf{v}_j$ vektorral, az $\mathbf{u}_{0,h}$ felbontásához hasonlóan az ismeretlen együtthatófüggvényekre az

$$g_k(t) = 2h(\mathbf{g}_h(t), \mathbf{v}_k)$$

összefüggést kapjuk. (V.ö. a (2.186) képlettel.) Behelyettesítve a $\mathbf{w}_h(t)$ (2.183) és $\mathbf{g}_h(t)$ (2.193) alakját a (2.192) egyenletbe, (2.184) analógiájára tetszőleges $t \in [0, t^\star]$ esetén a

$$\sum_{k=1}^{N} \left(c_k'(t) - \lambda_k^{(h)} c_k(t) - g_k(t) \right) \mathbf{v}_k = 0 \tag{2.194}$$

egyenletet nyerjük. Így minden $k = 1, 2, \ldots, N$ esetén a $c_k'(t) - \lambda_k^{(h)} c_k(t) = g_k(t)$ egyenletet nyerjük, amelynek megoldása a

$$c_k(t) = e^{\lambda_k^{(h)} t} c_k(0) + \int_0^t e^{\lambda_k^{(h)}(t-s)} g_k(s)\,\mathrm{d}s, \quad t \in [0, t^\star] \tag{2.195}$$

függvény. Tehát a (2.192) Cauchy-feladat megoldása

$$\mathbf{w}_h(t) = \sum_{k=1}^{N} c_k(t)\mathbf{v}_k, \tag{2.196}$$

ahol $c_k(t)$ a (2.195) szerinti együttható. (A $c_k(0)$ együtthatók a kezdeti függvény $\mathbf{v}_k$ szerinti felbontásából, (2.186) alapján számíthatók ki.)

Jelölje tetszőleges $t \in [0, t^\star]$ esetén $\mathbf{u}_h(t)$ a homogén peremfeltételű (2.120)-(2.121) feladat $u(x,t)$ pontos megoldásának rácspontokban felvett értékét, azaz $\mathbf{u}_h(t) = [u(x_1,t), \dots, u(x_N,t)]^T \in \mathbb{R}^N$. Célunk a (2.181) homogén egyenlet $\mathbf{y}_h(t)$ megoldásának és az $\mathbf{u}_h(t)$ különbségének, azaz az $\mathbf{e}_h(t) = \mathbf{y}_h(t) - \mathbf{u}_h(t)$ ún. globális hibafüggvénynek a becslése.

Jelölje

$$\mathbf{l}_h(t) = -\mathbf{u}_h'(t) + \mathbf{A}_h \mathbf{u}_h(t)$$

$\mathbb{R}^N$-be képező függvényt. (Ez az ún. lokális hibafüggvény.) Figyelembe véve a (2.192) egyenletet, ezen összefüggés alapján az $\mathbf{e}_h(t)$ globális hibára a

$$\begin{aligned} \mathbf{e}_h'(t) - \mathbf{A}_h \mathbf{e}_h(t) &= \mathbf{l}_h(t), \quad t \in [0, t^\star], \\ \mathbf{e}_h(0) &= \mathbf{0}, \end{aligned} \tag{2.197}$$

egyenletet nyerjük. Ez egy (2.192) alakú közönséges differenciálegyenlet-rendszer, amelynek megoldása (2.196)-(2.195) alakú. Mivel $\mathbf{e}_h(0) = \mathbf{0}$, ezért a megoldásban mindegyik $c_k(0) = 0$, és így a (2.197) feladat megoldása

$$\mathbf{e}_h(t) = \sum_{k=1}^{N} \left(\int_0^t e^{\lambda_k^{(h)}(t-s)} l_k(s)\, \mathrm{d}s \right) \mathbf{v}_k, \quad t \in [0, t^\star], \tag{2.198}$$

ahol $l_k(t) = 2h(\mathbf{l}_h(t), \mathbf{v}_k)$.

Tehát a $\|\cdot\|_\infty$ maximumnormában

$$
\begin{aligned}
\|\mathbf{e}_h(t)\|_\infty = \left\| \sum_{k=1}^{N} \left(\int_0^t e^{\lambda_k^{(h)}(t-s)} 2h(\mathbf{l}_h(s), \mathbf{v}_k)\, \mathrm{d}s \right) \mathbf{v}_k \right\|_\infty \leq \\
\sum_{k=1}^{N} \|\mathbf{v}_k\|_\infty \int_0^t e^{\lambda_k^{(h)}(t-s)} 2h |(\mathbf{l}_h(s), \mathbf{v}_k)|\, \mathrm{d}s \leq \\
2h \sum_{k=1}^{N} \|\mathbf{v}_k\|_\infty^2 \int_0^t e^{\lambda_k^{(h)}(t-s)} \|\mathbf{l}_h(s)\|_\infty\, \mathrm{d}s \leq \\
2h \sum_{k=1}^{N} \|\mathbf{v}_k\|_\infty^2 \int_0^t e^{\lambda_k^{(h)} s} \|\mathbf{l}_h(t-s)\|_\infty\, \mathrm{d}s \qquad (2.199) \\
\leq 2h \max_{[0,t]} \|\mathbf{l}_h(s)\|_\infty \sum_{k=1}^{N} \|\mathbf{v}_k\|_\infty^2 \int_0^t e^{\lambda_{max}^{(h)} s}\, \mathrm{d}s = \\
2h \max_{[0,t]} \|\mathbf{l}_h(s)\|_\infty \frac{e^{\lambda_{max}^{(h)} t} - 1}{\lambda_{max}^{(h)}} \sum_{k=1}^{N} \|\mathbf{v}_k\|_\infty^2 \\
\leq 2Nh \max_{[0,t]} \|\mathbf{l}_h(s)\|_\infty \frac{1 - e^{\lambda_{max}^{(h)} t}}{-\lambda_{max}^{(h)}} \max_{k=1,\ldots,N} \|\mathbf{v}_k\|_\infty^2.
\end{aligned}
$$

Mivel $\lambda_{max}^{(h)} \in (-\pi^2, -8)$, ezért $\left(1 - e^{\lambda_{max}^{(h)} t}\right) / -\lambda_{max}^{(h)} \leq 1/8$. Felhasználva továbbá a nyilvánvaló $Nh < 1$ és $\|\mathbf{v}_k\|_\infty \leq 1$ egyenlőtlenségeket, (2.199) következtében érvényes az

$$
\|\mathbf{e}_h(t)\|_\infty \leq \frac{1}{4} \cdot \max_{[0,t^\star]} \|\mathbf{l}_h(s)\|_\infty, \quad t \in [0, t^\star] \qquad (2.200)
$$

becslés.

Tekintsük a $\mathbf{l}_h(t)$ függvényt, és adjunk becslést a $\max_{[0,t^\star]} \|\mathbf{l}_h(s)\|_\infty$ kifejezésre! Emlékeztetünk a (2.153) összefüggésre: tetszőleges $x_j \in \omega_h$ pontban érvényes a

$$
\frac{\partial u}{\partial t}(x_j, t) - \frac{\partial^2 u}{\partial x^2}(x_j, t) = 0, \quad t \in [0, t^\star], \quad j = 1, 2, \ldots, N \qquad (2.201)
$$

egyenlőség. Másrészt, a lokális approximációs hiba definíciója alapján

$$
\left(\mathbf{l}_h(t)\right)_j = -\left(\mathbf{u}_h'(t)\right)_j + \left(\mathbf{A}_h \mathbf{u}_h(t)\right)_j. \qquad (2.202)
$$

Felhasználva a (2.201) egyenlőséget,

$$
\left(\mathbf{u}_h'(t)\right)_j = \frac{\partial u}{\partial t}(x_j, t) = \frac{\partial^2 u}{\partial x^2}(x_j, t). \qquad (2.203)
$$

Ugyanakkor az $\mathbf{A}_h$ mátrix definíciója következtében

$$\left(\mathbf{A}_h \mathbf{u}_h(t)\right)_j = \frac{u(x_{j+1},t) - 2u(x_j,t) + u(x_{j-1},t)}{h^2}. \tag{2.204}$$

Ezért a (2.203) és a (2.204) összefüggések felhasználásával, a (2.202) kifejezés felírható

$$\left(\mathbf{l}_h(t)\right)_j = -\frac{\partial^2 u}{\partial x^2}(x_j,t) + \frac{u(x_{j+1},t) - 2u(x_j,t) + u(x_{j-1},t)}{h^2} \tag{2.205}$$

alakban. A második derivált véges differenciás approximációjának hibabecslése alapján (pl. [5])

$$\left| -\frac{\partial^2 u}{\partial x^2}(x_j,t) + \frac{u(x_{j+1},t) - 2u(x_j,t) + u(x_{j-1},t)}{h^2} \right| \leq \frac{M_4}{12} h^2, \tag{2.206}$$

ahol $M_4 = \max\limits_{\Omega_{t^\star}} \left| \dfrac{\partial^4 u}{\partial x^4}(x,t) \right|$. Mindezek alapján érvényes a

$$\max_{[0,t^\star]} \|\mathbf{l}_h(s)\|_\infty \leq \frac{M_4}{12} h^2$$

egyenlőtlenség, azaz (2.200) alapján

$$\|\mathbf{e}_h(t)\|_\infty \leq \frac{M_4}{48} h^2, \quad t \in [0, t^\star], \tag{2.207}$$

azaz $\|\mathbf{e}_h(t)\|_\infty = \mathcal{O}(h^2)$, és így a szemidiszkretizáció megoldása másodrendben approximálja az eredeti feladat megoldását.

Disszipativitás

A korábbiakban megmutattuk, hogy a hővezetési egyenlet és annak szemidiszkretizációja egyaránt rendelkezik a nemnegativitás tulajdonsággal: mindkét feladat esetén nemnegatív bemenő adatok esetén a megoldás is nemnegatív. A továbbiakban egy további kvalitatív tulajdonságot, az ún. *disszipativitást* vizsgáljuk meg ugyanezen szempontból.[26]

Tekintsük ismételten a

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad x \in (0,1), \quad t > 0 \tag{2.208}$$

egyenletet a parabolikus peremen megadott

$$u(x,0) = u_0(x), \quad x \in (0,1) \tag{2.209}$$

[26] Ezt a részt elsősorban az érdeklődő Olvasónak szánjuk.

kezdeti, és az

$$u(0,t) = \alpha, \quad u(1,t) = \beta, \quad t > 0 \tag{2.210}$$

peremfeltételekkel. (Mint láttuk, ennek a feladatnak létezik megoldása, amely a 2.7. megjegyzés szerint előállítható.)

Általában, egy időfüggő feladatot disszipatívnak nevezünk, ha az azonos egyenletet és peremfeltételt kielégítő, de különböző kezdeti állapotból induló megoldások (négyzetes integrál értelemben vett) távolsága időben csökken.

2.15. Lemma *A fenti (2.208)-(2.210) feladat disszipatív, azaz a (2.208) egyenlet bármely két olyan $u(x,t)$ és $v(x,t)$ megoldására, amelyek kielégítik a (2.210) peremfeltételeket, a*

$$w(t) = \frac{1}{2}\int_0^1 (u(x,t) - v(x,t))^2 \; dx \tag{2.211}$$

ún. energiafüggvény monoton csökkenő.

Bizonyítás. Tekintsük az $\eta(x,t) = u(x,t) - v(x,t)$ függvényt! Ekkor

$$\frac{\partial \eta}{\partial t}(x,t) = \frac{\partial^2 \eta}{\partial x^2}(x,t), \quad x \in (0,1), \quad t > 0,$$

továbbá, $\eta(0,t) = \eta(1,t) = 0$ minden $t > 0$ esetén. Deriváljuk a (2.211) szerinti w függvényt! Ekkor az integrál és a derivált felcserélhetősége mellett érvényes a

$$\begin{aligned} \frac{\mathrm{d}w}{\mathrm{d}t}(t) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 \eta^2(x,t)\,\mathrm{d}x = \frac{1}{2}\int_0^1 2\eta\frac{\partial \eta}{\partial t}\,\mathrm{d}x = \int_0^1 \eta\frac{\partial \eta}{\partial t}\,\mathrm{d}x = \\ \int_0^1 \eta\frac{\partial^2 \eta}{\partial x^2}\,\mathrm{d}x = \underbrace{\left[\eta\frac{\partial \eta}{\partial x}\right]_{x=0}^{x=1}}_{=0} - \int_0^1 \left(\frac{\partial \eta}{\partial x}\right)^2\,\mathrm{d}x \le 0. \end{aligned} \tag{2.212}$$

becslés, amely az állításunkat bizonyítja. □

Most térjünk át a szemidiszkrét feladat hasonló tulajdonságának vizsgálatára!

Tekintsük tehát az

$$\mathbf{y}_h'(t) - \mathbf{A}_h\mathbf{y}_h(t) = \mathbf{f}_h(t), \quad t \in [0,\infty) \tag{2.213}$$

$$\mathbf{y}_h(0) = \mathbf{u}_{0,h} \tag{2.214}$$

szemidiszkrét feladatot, ahol $\mathbf{A}_h$ és $\mathbf{f}_h(t)$ (2.156) szerint adottak. Legyenek $\mathbf{v}_h(t)$ és $\mathbf{w}_h(t)$ $\mathbb{R}^N$-beli vektorfüggvények a (2.213) egyenlet két megoldása, és jelölje

$$\mathbf{w}_h(t) = \frac{1}{2}\sum_{j=1}^{N} \left(\mathbf{v}_h^j(t) - \mathbf{w}_h^j(t)\right)^2 \,\mathrm{d}x \tag{2.215}$$

a diszkrét energia függvényt, ahol $\mathbf{v}_h^j(t)$ és $\mathbf{w}_h^j(t)$ a $\mathbf{v}(t)$ és $\mathbf{w}(t)$ függvények j-dik koordinátafüggvényei. A következő állítás a 2.15.. lemma diszkrét megfelelője.

2.16. Lemma *A fenti (2.213) -(2.214) feladat disszipatív, azaz a (2.215) szerinti* $\boldsymbol{w}_h(t)$ *függvény monoton csökkenő.*

Bizonyítás. Jelölje $\boldsymbol{\eta}_h(t) = \mathbf{v}_h(t) - \mathbf{w}_h(t)$ és $(\cdot,\cdot)$ az $\mathbb{R}^N$-beli szokásos skaláris szorzást! Ekkor $\mathbf{w}_h(t) = 0.5(\boldsymbol{\eta}_h, \boldsymbol{\eta}_h)(t)$. Mivel (2.213) alapján $\boldsymbol{\eta}_h'(t) = \mathbf{A}_h\boldsymbol{\eta}_h(t)$, ezért

$$\frac{\mathrm{d}\boldsymbol{\eta}_h}{\mathrm{d}t}(t) = \mathbf{0}.5 \cdot 2\left(\boldsymbol{\eta}_h', \boldsymbol{\eta}_h\right)(t) = \left(\mathbf{A}_h\boldsymbol{\eta}_h, \boldsymbol{\eta}_h\right)(t). \tag{2.216}$$

Mivel az $\mathbf{A}_h$ mátrix negatív definit (lásd a 2.2.3 fejezetben az $\mathbf{A}_h$ mátrix 6. tulajdonságát), ezért $\left(\mathbf{A}_h\boldsymbol{\eta}_h, \boldsymbol{\eta}_h\right)(t) \leq 0$, ami az állításunkat igazolja. □

3. fejezet

Bevezetés a kezdetiérték-feladatok egylépéses numerikus módszereibe

3.1. Bevezetés az egylépéses módszerekbe

Az előző részben felsorolt tételek a megoldás létezésére és annak egyértelműségére vonatkoztak. Általános esetben a megoldást nem tudjuk előállítani, ugyanis a közönséges differenciálegyenletek kezdetiérték-feladatainak megoldásai csak nagyon speciális f függvények esetén adhatók meg képletek segítségével. Ehelyett *numerikus megoldást* állítunk elő, ami azt jelenti, hogy az értelmezési tartományának egyes pontjaiban az ismeretlen megoldásfüggvény értékeit *véges számú algoritmikus lépéssel közelítőleg* határozzuk meg. A fejezet további részeiben ilyen eljárásokat ismertetünk. Ebben a részben az olyan típusú eljárásokkal foglalkozunk, ahol valamely rögzített időpontbeli közelítést *az azt közvetlenül megelőző egy időpontbeli közelítés felhasználásával* határozunk meg.
Tehát a továbbiakban a célunk a

$$\frac{du}{dt} = f(t,u), \quad t \in [0,T], \tag{3.1}$$

$$u(0) = u_0 \tag{3.2}$$

feladat közelítő megoldása, ahol $T > 0$ olyan szám, amely mellett a (3.1)–(3.2) feladatnak létezik egyértelmű, megfelelően sima megoldása a $[0,T]$ intervallumon. Tehát feladatunk az $u(t)$ függvény közelítő meghatározása a $[0,T]$ intervallum véges számú pontjában[1]. Ebben a részben olyan módszereket tárgyalunk, amelyekben egy új időpontbeli közelítő érték kiszámításához csak egy, a közvetlen azt megelőző diszkrét időpontbeli közelítő értéket használjuk fel. Az ilyen módszereket nevezzük *egylépéses módszereknek*.

[1] Interpoláció segítségével ezután a $[0,T]$ intervallum tetszőleges pontjában kiszámítható a közelítés.

3.1.1. Taylor-sorba fejtéses módszer

Ez az egyik legrégebben ismert módszer. A definíció alapján a (3.1) egyenlet $u(t)$ megoldására érvényes az

$$u'(t) = f(t, u(t)), \quad t \in [0, T] \tag{3.3}$$

azonosság. Tegyük fel, hogy az f függvény analitikus, és így tetszőleges rendű parciális deriváltjai léteznek a Q_T halmazon. Ekkor az $u(t)$ megoldásfüggvény is analitikus, és ezért akárhányszor differenciálható [4, 17]. A láncszabály alkalmazásával rendre deriválva a (3.3) azonosságot a $t^\star \in [0, T]$ pontban, a következő egyenlőségeket nyerjük:

$$\begin{aligned} u'(t^\star) =& f(t^\star, u(t^\star)), \\ u''(t^\star) =& \partial_1 f(t^\star, u(t^\star)) + \partial_2 f(t^\star, u(t^\star))\ u'(t^\star), \\ u'''(t^\star) =& \partial_{11} f(t^\star, u(t^\star)) + 2\partial_{12} f(t^\star, u(t^\star))\ u'(t^\star) + \\ & \partial_{22} f(t^\star, u(t^\star))\ (u'(t^\star))^2 + \partial_2 f(t^\star, u(t^\star))\ u''(t^\star). \end{aligned} \tag{3.4}$$

Vegyük észre, hogy $u(t^\star)$ ismeretében mindegyik derivált pontosan kiszámítható. (Megjegyezzük, hogy tetszőleges, magasabb rendű derivált hasonló módon kiszámítható, csak a képletek egyre bonyolultabbá válnak.) Tegyük fel, hogy $t > t^\star$ olyan, amelyre $[t^\star, t] \subset [0, T]$. Mivel az $u(t)$ megoldásfüggvény analitikus, ezért Taylor-sora előállítja a $t^\star$ pont valamely környezetében. Tehát a

$$\sum_{k=0}^{n} \frac{u^{(k)}(t^\star)}{k!}(t - t^\star)^k \tag{3.5}$$

Taylor polinom $n \to \infty$ esetén konvergál az $u(t)$ megoldáshoz, ha t megfelelően közel van a $t^\star$ ponthoz. Ezért a konvergencia-tartományon belül a megoldás előállítható az

$$u(t) = \sum_{k=0}^{\infty} \frac{u^{(k)}(t^\star)}{k!}(t - t^\star)^k \tag{3.6}$$

egyenlőséggel. A megoldás a (3.6) képlet szerinti Taylor-soros előállítása a gyakorlati számítások során kivitelezhetetlen: feltételezi, hogy a $t^\star$ pontban az f függvény *végtelen sok* parciális deriváltját ismerjük, továbbá, hogy egy rögzített t pontban a jobb oldali *végtelen numerikus sort* pontosan tudjuk összegezni.

Az $u(t)$ pontos érték kiszámítása tehát a (3.6) képlet szerint nem valósítható meg. Ezért a továbbiakban ennek *közelítését* igyekszünk meghatározni. Kézenfekvő ötlet, hogy a Taylor-sor *véges szeletét* tekintsük közelítésnek, azaz

$$u(t) \simeq \sum_{k=0}^{p} \frac{u^{(k)}(t^\star)}{k!}(t - t^\star)^k =: T_{p,u}(t), \tag{3.7}$$

és ekkor az elhagyott rész (azaz a hiba) $\mathcal{O}((t-t^\star)^{p+1})$ nagyságrendű. Definíció alapján, $T_{p,u}(t)$ az $u(t)$ függvény $t^\star$ pontbeli p-ed fokú Taylor polinomja.

A (3.7) és a (3.4) összefüggések segítségével az alábbi közelítő eljárások definiálhatók.

a) Taylor-módszer

Válasszuk a $t^\star = 0$ pontot, ahol a kezdeti feltételünk adott.[2] Ekkor $u(t^\star) = u(0)$ ismert a kezdeti feltételből, és (3.4) alapján a deriváltak *pontosan* kiszámíthatók. Tehát a (3.7) közelítés alapján

$$u(t) \simeq \sum_{k=0}^{p} \frac{u^{(k)}(0)}{k!} t^k, \tag{3.8}$$

ahol $u^{(k)}(0)$ értékei (3.4) alapján kiszámolhatók. Természetesen nagy t esetén ez az approximáció eléggé pontatlan, igazából általános u függvény esetén csak a $t = 0$ pont *egy viszonylag kis környezetében* ad jó közelítést.

b) Lokális Taylor-módszer

Tekintsük az alábbi közelítő algoritmust.

1. A $[0, T]$ intervallumon a $t_0, t_1, \ldots t_N$ intervallumbeli pontok megadásával kijelölünk egy $\omega_h := \{0 = t_0 < t_1 < \ldots < t_{N-1} < t_N = T\}$ ún. *rácshálót*, amelynek *lépésközeit* $h_i = t_{i+1} - t_i$, (ahol $i = 0, 1, \ldots N-1$,) míg *finomságát* $h = \max_i h_i$-vel jelöljük. (A továbbiakban ezekben a pontokban határozzuk meg a közelítéseket, és az $u(t_i)$ közelítését y_i-vel, míg $u^{(k)}(t_i)$ közelítését $y_i^{(k)}$-val jelöljük, ahol $k = 0, 1, \ldots, p$.[3]
2. $y_0^{(k)}$ értékei a szükséges $k = 0, 1, \ldots, p$ értékekre a (3.4) összefüggések segítségével, $t^\star = 0$ behelyettesítéssel *pontosan* kiszámíthatók.
3. Az
$$y_1 = \sum_{k=0}^{p} \frac{y_0^{(k)}}{k!} h_0^k, \tag{3.9}$$
képlettel meghatározzuk $u(t_1)$ közelítését.
4. Az $i = 1, 2, \ldots, N-1$ értékekre y_i ismeretében a (3.4) összefüggések segítségével a $t^\star = t_i$ és $u(t^\star) = u(t_i) \approx y_i$ behelyettesítéssel *közelítőleg* meghatározzuk $y_i^{(k)}$ értékeit a $k = 0, 1, \ldots, p$ értékekre.

[2] Az előző szakaszban leírtakból következik, hogy értelmes a megoldásfüggvény $t = 0$ pontbeli deriváltjairól beszélni.

[3] Szokásosan a nulladik deriválási rend ($k = 0$) a függvényértéket jelenti.

5. Az

$$y_{i+1} = \sum_{k=0}^{p} \frac{y_i^{(k)}}{k!} h_i^k, \tag{3.10}$$

képlettel meghatározzuk az $u(t_{i+1})$ közelítését.

Írjuk ki a lokális Taylor-módszer algoritmusát (3.10) alapján a $p = 0, 1, 2$ esetekre!

- Ha $p = 0$, akkor $y_i = y_0$ minden i értékre, így ez az eset a gyakorlat szempontjából érdektelen.
- Legyen $p = 1$. Ekkor

$$y_{i+1} = y_i + y_i' h_i = y_i + h_i f(t_i, y_i), \quad i = 0, 1, \dots N-1, \tag{3.11}$$

ahol $y_0 = u_0$ adott.

- Legyen $p = 2$. Ekkor a számítási algoritmus

$$y_{i+1} = y_i + h_i y_i' + \frac{h_i^2}{2} y_i'' = y_i + h_i f(t_i, y_i) + \frac{h_i^2}{2} \left(\partial_1 f(t_i, y_i) + \partial_2 f(t_i, y_i) f(t_i, y_i) \right), \tag{3.12}$$

ahol $i = 0, 1, \dots N-1$ és y_0 adott.

Hasonlítsuk össze a fenti módszereket!

1. Mindkét módszer esetén a p-ed fokú Taylor-polinomot használjuk, ezért a módszer szükségessé teszi a $p-1$-ed rendig bezárólag valamennyi parciális derivált meghatározását. Ezek száma $p(p-1)/2$, és mindegyikben szükséges a függvények kiértékelése is. Ez már viszonylag kis p értékek mellett is rendkívül munkaigényes, ezért a gyakorlatban p értékét kicsinek szokás választani.[4] Ezért a Taylor-módszer gyakorlatban elérhető pontossága korlátozott.

2. A Taylor-módszer ugyan p növelésével egyre pontosabban közelíti a megoldást, valamint tetszőleges pontban közvetlenül kiszámítható a közelítés, de csak olyan t értékekre, amelyek a Taylor-sor konvergencia-sugaránál kisebbek. Ez a módszer egyik legnagyobb hátránya: a konvergencia-sugár általában nem ismert, vagy a konvergencia-sugár kisebb T-nél, így a teljes $[0, T]$ intervallumon nem lehetséges a közelítő megoldás előállítása.

[4] Az utóbbi években elterjedt szimbolikus számításos programok ugyan lehetőséget adnak az automatikus deriválásra, de a probléma még továbbra is fennáll.

3. A Taylor-módszer előnye, hogy ha csak egy rögzített $t = \hat{t}$ pontban vagyunk kíváncsiak a közelítésre, és ez a pont a konvergenciasugáron belül helyezkedik el, akkor közvetlenül, egy lépésben meghatározható a közelítés. A lokális Taylor-módszer kiküszöböli a fenti hiányosságokat, hiszen h a szükséges módon, akármilyen kicsinek választható. (Ugyanakkor, ennél a módszernél n darab feladat megoldása szükséges, ahol $h_0 + h_1 + \ldots + h_{n-1} = \hat{t}$, mivel csak a teljes $[0, \hat{t}]$ időintervallumon tudjuk előállítani a közelítést.)

4. A Taylor-módszer alkalmazása esetén a pontos és a közelítő megoldás eltérésére a Taylor-polinom hibatagjával becslést tudunk adni. A lokális Taylor-módszer esetén viszont a módszer hibája közvetlenül nem látszik, ugyanis az eltérés két részből adódik:

   a) minden lépésnél a Taylor-módszerhez hasonlóan a függvény n-ed fokú Taylor-polinommal történő approximációjából,

   b) a Taylor-polinom együtthatóit (azaz a megoldásfüggvény deriváltjait) csak közelítőleg tudjuk meghatározni. (Ráadásul, az itt elkövetett hiba a lépések során felhalmozódhat.)

5. Vegyük észre, hogy a fenti Taylor-módszerek felépítéséhez nem szükséges a megoldás analitikussága. Elegendő csak a megoldás $p+1$-szeres folytonos differenciálhatósága, azaz elegendő az $f \in C^p(Q_T)$ simasági feltétel.

3.1. Példa *Tekintsük az*

$$\begin{aligned} u'(t) &= -u(t) + t + 1, \quad t \in [0,1], \\ u(0) &= 1 \end{aligned} \tag{3.13}$$

feladatot, amelynek pontos megoldása $u(t) = \exp(-t) + t$.

Ebben a feladatban $f(t,u) = -u + t + 1$, ezért

$$\begin{aligned} u'(t) &= -u(t) + t + 1, \\ u''(t) &= -u'(t) + 1 = u(t) - t, \\ un'''(t) &= -u(t) + t, \end{aligned} \tag{3.14}$$

tehát $u(0) = 1, \quad u'(0) = 0, \quad u''(0) = 1, \quad u'''(0) = -1$. A globális Taylor-módszer esetén a közelítő polinomok:

$$\begin{aligned} T_{1,u}(t) &= 1, \\ T_{2,u}(t) &= 1 + t^2/2, \\ T_{3,u}(t) &= 1 + t^2/2 - t^3/6. \end{aligned} \tag{3.15}$$

t_i	a pontos megoldás	LT1	LT2	T1	T2	T3
0.1	1.0048	1.0000	1.0050	1.0000	1.0050	1.0048
0.2	1.0187	1.0100	1.0190	1.0000	1.0200	1.0187
0.3	1.0408	1.0290	1.0412	1.0000	1.0450	1.0405
0.4	1.0703	1.0561	1.0708	1.0000	1.0800	1.0693
0.5	1.1065	1.0905	1.1071	1.0000	1.1250	1.1042
0.6	1.1488	1.1314	1.1494	1.0000	1.1800	1.1440
0.7	1.1966	1.1783	1.1972	1.0000	1.2450	1.1878
0.8	1.2493	1.2305	1.2500	1.0000	1.3200	1.2347
0.9	1.3066	1.2874	1.3072	1.0000	1.4050	1.2835
1.0	1.3679	1.3487	1.3685	1.0000	1.5000	1.3333

3.1. táblázat. A lokális Taylor-módszer és a Taylor-módszer összehasonlítása a $h = 0.1$ lépésközű rácshálón

Ezért a $t = 1$ pontban $T_{1,u}(1) = 1$, $T_{2,u}(1) = 1.5$, $T_{3,u}(1) = 1.333$). (Könnyen kiszámítható, hogy $T_{4,u}(1) = 1.375$ és $T_{5,u}(1) = 1.3666$.) Mint látható, ezek az értékek csak viszonylag magas n esetén közelítik megfelelően az $u(1) = 1.367879$ értéket.

Alkalmazzuk a (3.10) lokális Taylor-módszert, figyelembe véve a (3.14) deriváltakat. Az elsőrendű módszer algoritmusa

$$y_{i+1} = y_i + h_i(-y_i + t_i + 1), \quad i = 0, 1, \ldots, N-1, \tag{3.16}$$

míg a másodrendű módszer algoritmusa

$$y_{i+1} = y_i + h_i(-y_i + t_i + 1) + \frac{h_i^2}{2}(y_i - t_i), \quad i = 0, 1, \ldots, N-1,$$

ahol $h_1 + h_2 + \ldots + h_N = T$.

Számításainkat a $h_i = h = 0.1$ egyenközű (ún. ekvidisztáns) rácshálón végezzük el. A 3.1 táblázatban a $[0, 1]$ intervallum rácspontjaiban hasonlítjuk össze a lokális és globális Taylor-módszereket. (LT1 és LT2 a lokális első- illetve másodrendű Taylor-módszert, míg T1, T2 és T3 a első-, másod- és harmadrendű Taylor-módszereket jelöli.)

A vizsgált módszerekkel nyert numerikus megoldás és az eredeti feladat megoldásának eltérése a rácsháló pontjaiban meghatározza az ún. hibavektort, és ennek maximumnormáját hasonlítjuk össze a 3.2 táblázatban a különböző, egyre finomodó rácshálókon. Jól látható, hogy h csökkenése esetén a lokális Taylor-módszer hibája csökken, viszont a Taylor-módszer által nyert eredmény változatlan marad. (Ez utóbbi természetes következménye annak, hogy a módszer független a lépésköz megválasztásától.)

A lokális Taylor-módszer egylépéses módszer, hiszen a t_i pontbeli értékek határozzák meg a t_{i+1}-beli közelítést. A hibaanalízise meglehetősen bonyolult. Mint azt a fenti példa

a h lépésköz	LT1	LT2	T1	T2	T3
0.1	$1.92e-02$	$6.62e-04$	0.3679	0.1321	0.0345
0.01	$1.80e-03$	$6.12e-06$	0.3679	0.1321	0.0345
0.001	$1.85e-04$	$6.14e-08$	0.3679	0.1321	0.0345
0.0001	$1.84e-05$	$6.13e-10$	0.3679	0.1321	0.0345

3.2. táblázat. A lokális Taylor-módszer és a Taylor-módszer hibája h lépésközű rácshálón a maximum normában

is jól mutatja, az y_{i+1} közelítésnek az $u(t_{i+1})$ pontos értéktől való eltérését több tényező okozza.

- Az ún. *lokális csonkolási hiba*, ami a Taylor-sor Taylor-polinommal való helyettesítéséből ered, feltételezve, hogy a t_i pontbeli értéket pontosan ismerjük. Ennek a $[t_i, t_i + h_i]$ intervallum hossza szerinti rendjét, vagyis az $u(t) - T_{n,u}(t)$ eltérés h_i szerinti rendjét *lokális hibarendnek* nevezzük. (Megfelelően sima függvények esetén ez a rend $\mathcal{O}(h_i^{p+1})$.)

- Minden lépésben (kivéve az elsőt) a sorbafejtésben nem a pontos t_i-beli értékek, hanem azoknak közelítései szerepelnek, és ezek az eltérések a lépések során felhalmozódhatnak.

- Minden számításnál *kerekítési hibák* is fellépnek, amelyek jelentősen torzíthatják a közelítést. (Ez természetes velejárója a számítógépek korlátozott pontosságának, és nagysága függ a gépi pontosságtól. (A módszerek vizsgálata során mi ezen hibával nem foglalkozunk.)

- Amikor a $[0, t^\star]$ intervallumon megoldjuk a feladatunkat, akkor a $t^\star$ pontbeli közelítés első két hibaforrásból eredő hibáját *globális hibának* nevezzük. Intuitív módon azt mondjuk, hogy a módszer konvergens a $t = t^\star$ pontban, amikor a maximális lépésköz h nullához tartása esetén ez a globális hiba is nullához tart. A globális hiba nullához tartásának rendjét a módszer *konvergenciarendjének* nevezzük. Ez a rend független a kerekítési hibáktól. Mivel a $t = t^\star$ pontbeli közelítés meghatározásához kb. n lépést kell tennünk, ahol $nh = t^\star$, ezért $\mathcal{O}(h^{p+1})$ lokális csonkolási hiba mellett a globális konvergencia várható rendje $\mathcal{O}(h^p)$. (A 3.2 táblázat LT1 és LT2 módszereihez tartozó eredmények ezt alátámasztják: az LT1 elsőrendben, míg LT2 másodrendben konvergens a $t^\star = 1$ pontban.)

A Taylor-módszer viselkedése az $u'(t) = 1 - t\sqrt[3]{u(t)}$ differenciálegyenletre jól látható a
http://math.fullerton.edu/mathews/a2001/Animations/OrdinaryDE/Taylor/Taylor.html
linken lévő animáción.

3.1.2. Néhány nevezetes egylépéses módszer

Az előző részben láttuk, hogy numerikus szempontból a lokális Taylor-módszer különösen $p = 1$ esetén előnyös: a (3.11) képlethez nem kell meghatározni az f függvény parciális deriváltjait, és a lépésközök csökkentésével a rácspontokban az ismeretlen függvény jól közelíthető. Ebben a részben az a célunk, hogy újabb, hasonló tulajdonságokkal rendelkező egylépéses módszereket definiáljunk.

Az LT1 módszert az ismeretlen $u(t)$ megoldásfüggvénynek a $[t_i, t_{i+1}]$ intervallumon $T_{1,u}(t)$ elsőrendű Taylor-polinommal való approximációjából nyertük.[5] Ekkor az elkövetett hiba (a lokális csonkolási hiba)

$$|u(t_{i+1}) - T_{1,u}(t_{i+1})| = \mathcal{O}(h_i^2), \quad i = 0, 1, \ldots, N-1, \tag{3.17}$$

azaz másodrendben pontos az approximáció. Adjunk meg $T_{1,u}(t)$ helyett olyan más, $P_1(t)$ elsőfokú polinomot, amely mellett a (3.17) becslés továbbra is érvényben marad, azaz

$$|u(t_{i+1}) - P_1(t_{i+1})| = \mathcal{O}(h_i^2). \tag{3.18}$$

Mivel $T_{1,u}(t)$ a megoldásgörbét a $(t_i, u(t_i))$ pontbeli érintő egyenes, ezért olyan $P_1(t)$ polinomot keresünk, amely szintén átmegy ezen a ponton, de irányát – mivel a megoldásgörbe rácspontbeli értékeiből akarjuk meghatározni – az $u(t)$ függvény t_i és t_{i+1} pontbeli érintőinek iránya határozza meg. Ezért legyen $P_1(t) := u(t_i) + \alpha(t - t_i)$ $(t \in [t_i, t_{i+1}])$ alakú, ahol $\alpha = \alpha(u'(t_i), u'(t_{i+1}))$ egy adott függvény. (Például, az $\alpha = u'(t_i)$ megválasztással $P_1(t) = T_{1,u}(t)$ és ekkor természetesen érvényes (3.18).)
Lehet-e más alkalmas megválasztás is? Mivel

$$u(t_{i+1}) = u(t_i) + u'(t_i)h_i + \mathcal{O}(h_i^2), \tag{3.19}$$

ezért az

$$u(t_{i+1}) - P_1(t_{i+1}) = h_i(u'(t_i) - \alpha) + \mathcal{O}(h_i^2),$$

azaz (3.18) pontosan akkor teljesül, amikor az

$$\alpha - u'(t_i) = \mathcal{O}(h_i) \tag{3.20}$$

becslés érvényes.

3.2. Tétel *Tetszőleges $\theta \in \mathbb{R}$ esetén az*

$$\alpha = (1 - \theta)u'(t_i) + \theta u'(t_{i+1}) \tag{3.21}$$

megválasztású α függvény esetén a (3.20) becslés érvényes.

[5] Mindegyik $[t_i, t_{i+1}]$ intervallumon más polinomot határozunk meg, de a polinomok ezen i-től való függését a jelöléseinkben nem hangsúlyozzuk.

Bizonyítás. Alkalmazzuk a (3.19) felbontást az $u'(t)$ függvényre!

$$u'(t_{i+1}) = u'(t_i) + u''(t_i)h_i + \mathcal{O}(h_i^2), \tag{3.22}$$

és behelyettesítve a (3.22) összefüggést a (3.21) képletbe, a

$$\alpha - u'(t_i) = \theta u''(t_i)h_i + \mathcal{O}(h_i^2) \tag{3.23}$$

összefüggést nyerjük, ami az állításunkat bizonyítja. □

3.3. Következmény A fenti $P_1(t)$ polinom az

$$y_{i+1} = y_i + \alpha h_i \tag{3.24}$$

egylépéses numerikus módszert határozza meg, ahol a (3.21) és a (3.1) összefüggések alapján

$$\alpha = (1-\theta)f(t_i, y_i) + \theta f(t_{i+1}, y_{i+1}). \tag{3.25}$$

3.4. Definíció *A (3.24)-(3.25) numerikus módszert θ-módszernek nevezzük.*

3.5. Megjegyzés A θ-módszer esetén is jellemző, hogy y_i valamilyen közelítése az $u(t_i)$ pontos értéknek, és az eltérés – a Taylor-módszernél is említettekkel megegyezően – alapvetően a következők miatt van:

a) minden lépésnél az $u(t)$ megoldásfüggvényt az elsőfokú $P_1(t)$ polinommal approximáljuk,

b) a $P_1(t)$ polinomban az α együtthatót (azaz a megoldásfüggvény deriváltjait) csak közelítőleg tudjuk meghatározni.

Mivel az α irányt a megoldásfüggvény t_i és t_{i+1} pontbeli érintőinek iránya határozza meg, ezért általában úgy választjuk meg, hogy ezen két érték közé essék. Ezért a θ paramétert csak a $[0,1]$ intervallumból szokás megválasztani. A továbbiakban három, speciálisan megválasztott $\theta \in [0,1]$ értékhez tartozó numerikus módszert vizsgálunk meg.

3.1.3. Az explicit Euler-módszer

Tekintsük a θ-módszert a $\theta = 0$ megválasztással! Ekkor (3.24) és (3.25) a következő numerikus módszert generálják:

$$y_{i+1} = y_i + h_i f(t_i, y_i), \quad i = 0, 1, \ldots, N-1. \tag{3.26}$$

Mivel y_i az ismeretlen $u(t)$ függvény t_i pontbeli közelítése, ezért értelemszerűen

$$y_0 = u(0) = u_0, \tag{3.27}$$

vagyis a (3.26) iterációban az $i = 0$ értékhez tartozó y_0 adott érték.

3.6. Definíció *A (3.26)–(3.27) képletekkel definiált egylépéses módszert* explicit Euler-módszernek *nevezzük.*

Mivel a $\theta = 0$ esetén $\alpha = u'(t_i)$, ezért ebben az esetben a módszert definiáló P_1 polinom megegyezik az elsőrendű Taylor-polinommal. Tehát az explicit Euler-módszer azonos a (3.11) képlettel definiált elsőfokú közelítéses lokális Taylor-módszerrel.

3.7. Megjegyzés A (3.26)–(3.27) módszert azért nevezzük explicitnek, mert a t_i pontbeli érték ismeretében közvetlenül, egy egyszerű függvénybehelyettesítéssel kiszámítható a t_{i+1} pontbeli közelítés.

Az explicit Euler-módszer szemléletes működése az alábbi animáción látható:
../animaciok/ee.gif

A 3.5. megjegyzésben felsoroltuk a módszer hibájának, azaz a pontos és a közelítő megoldás eltérésének forrásait. Alapvető kérdés, hogy egy rögzített $t^\star \in [0, T]$ pontban a rácsháló finomításával hogyan viselkedik a közelítő és pontos érték különbsége, és vajon, az elvárásainknak megfelelően, a minden határon túl való finomítás esetén nullához tart-e ?

A továbbiakban ezt a kérdést az explicit Euler-módszer esetén vizsgáljuk meg. (Az előzőekben leírtaknak megfelelően feltesszük, hogy az f függvény a második változójában lipschitzes, és a megoldás kellően sima.)

Először az ekvidisztáns rácshálók sorozatán vizsgáljuk a kérdést, azaz legyenek

$$\omega_h := \{t_i = ih;\ i = 0, 1, \ldots, N;\ h = T/N\}$$

adott rácshálók és tegyük fel, hogy a rögzített $t^\star \in [0, T]$ egy olyan pont, amely mindegyik rácshálónak pontja, azaz $t^\star \in \omega_h$ minden vizsgált $h > 0$ esetén. Egy rögzített ω_h rácshálón jelölje n azt az indexet, amelyre $nh = t^\star$. Ekkor n függ h megválasztásától és $h \to 0$ esetén n a végtelenbe tart. Jelölje

$$e_i = y_i - u(t_i),\ i = 0, 1, \ldots N \tag{3.28}$$

a t_i pontbeli hibát. Mi a továbbiakban h csökkenése mellett az e_n viselkedésére vagyunk kíváncsiak, pontosabban arra, hogy hogyan viselkedik a $t^\star$ pontbeli pontos és közelítő megoldás különbsége $h \to 0$ esetén.[6] A (3.26) explicit Euler-módszer képletébe behelyettesítve a (3.28) definícióból következő $y_i = e_i + u(t_i)$ kifejezést, érvényes a következő egyenlőség:

$$\begin{aligned} e_{i+1} - e_i = &- (u(t_{i+1}) - u(t_i)) + hf(t_i, e_i + u(t_i)) = \\ &[hf(t_i, u(t_i)) - (u(t_{i+1}) - u(t_i))] + \\ &h\left[f(t_i, e_i + u(t_i)) - f(t_i, u(t_i))\right]. \end{aligned} \tag{3.29}$$

[6] A továbbiakban látszik, hogy a $t^\star \in \omega_h$ minden $h > 0$ esetén feltétel enyhíthető: elegendő megkövetelni, hogy a t_n rácspontok tartsanak $t^\star$ ponthoz, azaz $\lim_{h\to 0}(t^\star - t_n) = 0$.

Így, bevezetve a

$$\begin{aligned} g_i &= hf(t_i, u(t_i)) - (u(t_{i+1}) - u(t_i)), \\ \psi_i &= f(t_i, e_i + u(t_i)) - f(t_i, u(t_i)) \end{aligned} \tag{3.30}$$

jelöléseket az

$$e_{i+1} - e_i = g_i + h\psi_i \tag{3.31}$$

ún. *hibaegyenletet* kapjuk.

3.8. Megjegyzés Vizsgáljuk meg a (3.30) definícióban szereplő két tagot! A g_i tag azt mutatja, hogy a pontos megoldás az explicit Euler-módszer (3.26) $hf(t_i, y_i) - (y_{i+1} - y_i) = 0$ alakban felírt képletét milyen pontosan elégíti ki. Ez a kifejezés azt a hibát tartalmazza, ami az $u(t)$ megoldásfüggvénynek a $[t_i, t_{i+1}]$ intervallumon történő, elsőfokú Taylor-polinommal való approximációjából ered. A ψ_i tag azt jellemzi, hogy a módszer egy lépése során mekkora hiba keletkezik abból, hogy az y_{i+1} érték kiszámolására szolgáló képletben a pontos $u(t_i)$ érték helyett annak y_i közelítésével számolunk.

Az f függvény lipschitzessége következtében

$$|\psi_i| = |f(t_i, e_i + u(t_i)) - f(t_i, u(t_i))| \le L|(e_i + u(t_i)) - u(t_i)| = L|e_i|. \tag{3.32}$$

Így a (3.31) és a (3.32) összefüggések alapján

$$|e_{i+1}| \le |e_i| + |g_i| + h|\psi_i| \le (1 + hL)|e_i| + |g_i| \tag{3.33}$$

minden $i = 0, 1, \ldots, n-1$ értékre. Ezért

$$\begin{aligned} |e_n| &\le (1 + hL)|e_{n-1}| + |g_{n-1}| \le \\ &\quad (1 + hL)\left[(1 + hL)|e_{n-2}| + |g_{n-2}|\right] + |g_{n-1}| = \\ &\quad (1 + hL)^2|e_{n-2}| + \left[(1 + hL)|g_{n-2}| + |g_{n-1}|\right] \\ &\quad \le (1 + hL)^n|e_0| + \sum_{i=0}^{n-1}(1 + hL)^i|g_{n-1-i}| \\ &\quad < (1 + hL)^n \left[|e_0| + \sum_{i=0}^{n-1}|g_{n-1-i}|\right]. \end{aligned} \tag{3.34}$$

(Az utolsó lépésben az $(1 + hL)^i < (1 + hL)^n$, $i = 0, 1, \ldots n-1$ egyenlőtlenséget alkalmaztuk.) Mivel tetszőleges pozitív x esetén $1 + x < \exp(x)$, ezért az $nh = t^\star$ egyenlőség miatt $(1 + hL)^n < \exp(nhL) = \exp(Lt^\star)$. Így (3.34) alapján

$$|e_n| \le \exp(Lt^\star)\left[|e_0| + \sum_{i=0}^{n-1}|g_{n-1-i}|\right]. \tag{3.35}$$

Adjunk becslést a $|g_i|$ kifejezésre! Könnyen láthatóan

$$u(t_{i+1}) - u(t_i) = u(t_i + h) - u(t_i) = hu'(t_i) + \frac{1}{2}u''(\xi_i)h^2, \tag{3.36}$$

ahol $\xi_i \in (t_i, t_{i+1})$ egy adott pont. Mivel $f(t_i, u(t_i)) = u'(t_i)$, ezért g_i (3.30) szerinti definíciója következtében érvényes az

$$|g_i| \leq \frac{M_2}{2}h^2, \quad M_2 = \max_{[0,t^\star]} |u''(t)| \tag{3.37}$$

egyenlőtlenség. Ekkor a (3.35) és a (3.37) becslések alapján

$$|e_n| \leq \exp(Lt^\star) \left[|e_0| + hn\frac{M_2}{2}h \right] = \exp(Lt^\star) \left[|e_0| + \frac{t^\star M_2}{2}h \right]. \tag{3.38}$$

Mivel $e_0 = 0$, ezért

$$|e_n| \leq \exp(Lt^\star)\frac{t^\star M_2}{2}h. \tag{3.39}$$

Így $h \to 0$ esetén $e_n \to 0$, és emellett $e_n = \mathcal{O}(h)$.

3.9. Megjegyzés A numerikus megoldás konvergenciájának fenti bizonyítása viszonylag egyszerűen kiterjeszthető az alkalmasan megválasztott nemekvidisztáns rácshálókra is. Legyen most

$$\omega_{h_v} := \{0 = t_0 < t_1 < \ldots < t_{N-1} < t_N = T\}$$

finomodó, változó lépéshosszúságú rácshálók sorozata. Jelölje $h_i = t_{i+1} - t_i$, $i = 0, 1, , \ldots, N-1$ és $h = T/N$. Feltesszük, hogy a rácspontok számának növelésével a rácsháló mindenütt finomodik, azaz létezik olyan $0 < c < \infty$ állandó, amelyre minden N esetén

$$h_i \leq ch, \quad i = 1, 2, \ldots, N. \tag{3.40}$$

Továbbra is feltesszük, hogy a rögzített $t^\star \in [0, T]$ pont mindegyik rácshálónak az eleme. Egy ilyen rögzített rácshálón jelölje n azt a indexet, amelyre $h_0 + h_1 + \ldots + h_{n-1} = t^\star$. A

$$\begin{aligned} g_i &= h_i f(t_i, u(t_i)) - (u(t_{i+1}) - u(t_i)), \\ \psi_i &= f(t_i, e_i + u(t_i)) - f(t_i, u(t_i)) \end{aligned} \tag{3.41}$$

jelölésekkel a (3.33) becslés így írható át:

$$\begin{gathered} |e_{i+1}| \leq |e_i| + |g_i| + h_i|\psi_i| \leq |e_i| + |g_i| + h_i L|e_i| \leq \\ (1 + h_i L)|e_i| + |g_i| \leq \exp(h_i L)|e_i| + |g_i| \leq \exp(h_i L)\left[|e_i| + |g_i|\right]. \end{gathered} \tag{3.42}$$

Ekkor a (3.34) becslés a (3.42) figyelembevételével így alakul:

$$\begin{aligned}|e_n| \leq & \exp(h_{n-1}L)\left[|e_{n-1}| + |g_{n-1}|\right] \leq \\ & \exp(h_{n-1}L)\left[\exp(h_{n-2}L)\left(|e_{n-2}| + |g_{n-2}|\right) + |g_{n-1}|\right] = \\ & \exp((h_{n-1}+h_{n-2})L)\left(|e_{n-2}| + |g_{n-2}| + |g_{n-1}|\right) \leq \\ & \exp((h_{n-1}+h_{n-2}+\ldots+h_0)L)\left[|e_0| + \sum_{i=1}^{n}|g_{n-i}|\right].\end{aligned} \tag{3.43}$$

Mivel $h_{n-1} + h_{n-2} + \ldots + h_0 = t^\star$ és érvényes a

$$|g_i| \leq \frac{M_2}{2}h_i^2 \leq \frac{M_2 c^2}{2}h^2 \tag{3.44}$$

becslés, ezért (3.43) és (3.44) alapján

$$|e_n| \leq \exp(t^\star L)\left[|e_0| + hn\frac{M_2c^2}{2}h\right] = \exp(t^\star L)\left[|e_0| + t^\star\frac{M_2c^2}{2}h\right]. \tag{3.45}$$

A (3.45) becslés azt mutatja, hogy megfelelően finomodó rácshálókon $h \to 0$ esetén $e_n \to 0$, és emellett $e_n = \mathcal{O}(h)$.

3.10. Megjegyzés Látható, hogy az explicit Euler-módszer esetén a $\lim_{h\to 0} e_n = 0$ egyenlőséghez nem szükséges az $y_0 = u_0$ megválasztás, elegendő, ha $y_0 = u_0 + \mathcal{O}(h)$, mert ebben az esetben $e_0 = \mathcal{O}(h)$. (Emellett, továbbra is $e_n = \mathcal{O}(h)$.)

3.1.4. Az implicit Euler-módszer

Tekintsük a θ-módszert a $\theta = 1$ megválasztással! Ekkor (3.24) és (3.25) a következő numerikus módszert generálja:

$$y_{i+1} = y_i + h_i f(t_{i+1}, y_{i+1}), \quad i = 0, 1, \ldots, N-1, \tag{3.46}$$

ahol ismét $y_0 = u_0$.

3.11. Definíció *A (3.46)–(3.27) képletekkel definiált egylépéses módszert* implicit Euler-módszernek *nevezzük.*

3.12. Megjegyzés A (3.46) implicit Euler-módszert azért nevezzük implicitnek, mert az időben való előrehaladáshoz y_i ismeretében y_{i+1} értékét minden egyes időlépésben egy (tipikusan nemlineáris) egyenlet megoldásával tudjuk csak meghatározni.

Az implicit Euler-módszer e_i hibafüggvényére a hibaegyenlet a következő módon írható fel:

$$\begin{aligned} e_{i+1} - e_i = &- (u(t_{i+1}) - u(t_i)) + h_i f(t_{i+1}, u(t_{i+1}) + e_{i+1}) = \\ &[h_i f(t_{i+1}, u(t_{i+1})) - (u(t_{i+1}) - u(t_i))] + \\ &h_i \left[f(t_{i+1}, u(t_{i+1}) + e_{i+1}) - f(t_{i+1}, u(t_{i+1})) \right]. \end{aligned} \tag{3.47}$$

Így a

$$\begin{aligned} g_i &= h_i f(t_{i+1}, u(t_{i+1})) - (u(t_{i+1}) - u(t_i)), \\ \psi_i &= f(t_{i+1}, u(t_{i+1}) + e_{i+1}) - f(t_{i+1}, u(t_{i+1})) \end{aligned} \tag{3.48}$$

jelölésekkel ismételten a (3.31) alakú hibaegyenletet nyerjük.
Nyilvánvalóan

$$u(t_{i+1}) - u(t_i) = u(t_{i+1}) - u(t_{i+1} - h_i) = h_i u'(t_{i+1}) - \frac{1}{2} u''(\xi_i) h_i^2, \tag{3.49}$$

ahol $\xi_i \in (t_i, t_{i+1})$ egy adott pont. Másrészt, $f(t_{i+1}, u(t_{i+1})) = u'(t_{i+1})$. Ezért g_i (3.48) szerinti definíciója következtében érvényes a

$$|g_i| \leq \frac{M_2}{2} h_i^2 \tag{3.50}$$

egyenlőtlenség. Ugyanakkor az implicit Euler-módszer esetén ψ_i a t_{i+1} pontbeli közelítéstől és a pontos megoldástól egyaránt függ, ezért az explicit Euler-módszer vizsgálata ebben az esetben változatlan formában nem ismételhető meg. (A módszer e_n hibafüggvényének nullához tartásával később foglalkozunk.)

3.1.5. A trapéz-módszer

Tekintsük a θ-módszert a $\theta = 0.5$ megválasztással! Ekkor (3.24) és (3.25) a következő numerikus módszert generálja:

$$y_{i+1} - y_i = \frac{h_i}{2} \left[f(t_i, y_i) + f(t_{i+1}, y_{i+1}) \right], \quad i = 0, 1, \ldots, N-1, \tag{3.51}$$

ahol $y_0 = u_0$.

3.13. Definíció *A (3.51) egylépéses módszert* trapéz módszer *nevezzük.*

(Vegyük észre, hogy a trapéz-módszer is implicit.)

A trapéz-módszer hibafüggvényére az explicit és implicit Euler-módszerek hibaegyenleteinek kombinálásával könnyen nyerhető a (3.31) alakú hibaegyenletet, ahol most

$$\begin{aligned} g_i &= \frac{1}{2} h_i \left[f(t_i, u(t_i)) + f(t_{i+1}, u(t_{i+1})) \right] - (u(t_{i+1}) - u(t_i)), \\ \psi_i &= \frac{1}{2} \left[f(t_i, u(t_i) + e_i) - f(t_i, u(t_i)) \right] + \\ &\quad \frac{1}{2} \left[f(t_{i+1}, u(t_{i+1}) + e_{i+1}) - f(t_{i+1}, u(t_{i+1})) \right]. \end{aligned} \tag{3.52}$$

Adjunk becslést a (3.52) képletben szereplő g_i kifejezésre! A $t_{i+\frac{1}{2}} = t_i + 0.5h_i$ jelöléssel fejtsük sorba az $u(t_i) = u(t_{i+\frac{1}{2}} - h_i/2)$ és az $u(t_{i+1}) = u(t_{i+\frac{1}{2}} + h_i/2)$ kifejezéseket a $t = t_{i+\frac{1}{2}}$ pont körül. Ekkor

$$u(t_{i+1}) - u(t_i) = h_i u'(t_{i+1/2}) + \frac{h_i^3}{48}(u'''(\xi_i^1) + u'''(\xi_i^2)), \tag{3.53}$$

ahol $\xi_i^1, \xi_i^2 \in (t_i, t_{i+1})$ adott pontok. Másrészt

$$f(t_i, u(t_i)) + f(t_{i+1}, u(t_{i+1})) = u'(t_i) + u'(t_{i+1}). \tag{3.54}$$

Sorba fejtve a (3.54) jobb oldali függvényeit a $t = t_{i+\frac{1}{2}}$ pont körül, az

$$\frac{1}{2}\left[f(t_i, u(t_i)) + f(t_{i+1}, u(t_{i+1}))\right] = u'(t_{i+\frac{1}{2}}) + \frac{h_i^2}{16}(u'''(\xi_i^3) + u'''(\xi_i^4)). \tag{3.55}$$

egyenlőséget kapjuk. Ezért (3.53) és (3.55) alapján a g_i kifejezésre érvényes a

$$|g_i| \leq \frac{M_3}{6} h_i^3, \quad M_3 = \max_{[0,t^\star]} |u'''(t)| \tag{3.56}$$

egyenlőtlenség.

A 3.3 táblázatban a (3.13) tesztfeladat fenti három numerikus módszerrel való megoldását ismertetjük. A 3.4 táblázatban a hibákat a maximumnormában hasonlítjuk össze a különböző, egyre finomodó rácshálókon. Az eredményekből megállapítható, hogy rögzített rácshálón a numerikus megoldás a explicit Euler-módszer és a implicit Euler-módszer esetén nagyjából hasonló pontosságot ad, míg a trapéz-módszer pontosabb az előző két módszernél. A finomodó rácshálókon azt figyelhetjük meg, hogy a trapéz-módszer hibafüggvénye $\mathcal{O}(h^2)$, az explicit Euler-módszer és az implicit Euler-módszer hibafüggvénye viszont csak $\mathcal{O}(h)$ rendben tart nullához. (Az általános alakú θ-módszer esetén a hibafüggvény nullához tartását a későbbiekben bizonyítjuk be.)

Az Euler-módszer viselkedése az $u'(t) = 1 - t\sqrt[3]{u(t)}$ differenciálegyenletre jól látható a
http://math.fullerton.edu/mathews/a2001/Animations/Animations9.html
linken megtalálható animáción.

3.1.6. Az egylépéses módszerek konvergenciája

Ebben a szakaszban az

$$\omega_h := \{t_i = ih;\ i = 0, 1, \ldots, N;\ h = T/N\}$$

ekvidisztáns rácshálón (avagy azok sorozatán) megadjuk az egylépéses módszerek általános alakját, és definiáljuk a numerikus módszerek alapfogalmait. Tekintsük az

$$y_{i+1} = y_i + h\Phi(h, t_i, y_i, y_{i+1}) \tag{3.57}$$

t_i	a pontos érték	EE	IE	TM
0.1	1.0048	1.0000	1.0091	1.0048
0.2	1.0187	1.0100	1.0264	1.0186
0.3	1.0408	1.0290	1.0513	1.0406
0.4	1.0703	1.0561	1.0830	1.0701
0.5	1.1065	1.0905	1.1209	1.1063
0.6	1.1488	1.1314	1.1645	1.1485
0.7	1.1966	1.1783	1.2132	1.1963
0.8	1.2493	1.2305	1.2665	1.2490
0.9	1.3066	1.2874	1.3241	1.3063
1.0	1.3679	1.3487	1.3855	1.3676

3.3. táblázat. Az explicit Euler-módszer (EE), az implicit Euler-módszer (IE) és a trapéz módszer (TM) összehasonlítása a $h = 0.1$ lépésközű rácshálón

a h lépésköz	EE	IE	TM
0.1	$1.92e-02$	$1.92e-02$	$3.06e-04$
0.01	$1.84e-03$	$1.84e-03$	$3.06e-06$
0.001	$1.84e-04$	$1.84e-04$	$3.06e-08$
0.0001	$1.84e-05$	$1.84e-05$	$3.06e-10$
0.0001	$1.84e-06$	$1.84e-06$	$5.54e-12$

3.4. táblázat. Az explicit Euler-módszer (EE), az implicit Euler-módszer (IE) és a trapéz módszer (TM) hibája h lépésközű rácshálón a maximum normában

egylépéses módszert, ahol Φ a numerikus módszert meghatározó adott függvény. A továbbiakban azt a numerikus módszert, amelyet a (3.57) képlet realizál, Φ-*numerikus módszernek* (röviden: Φ-*módszernek*) nevezzük.

3.14. Megjegyzés Speciálisan megválasztott Φ függvények esetén a korábbi módszereink előállíthatók a (3.57) alakban. Például,

- $\Phi(h, t_i, y_i, y_{i+1}) = f(t_i, y_i)$ esetén az explicit Euler-módszert;
- $\Phi(h, t_i, y_i, y_{i+1}) = f(t_i + h, y_{i+1})$ esetén az implicit Euler-módszert;
- $\Phi(h, t_i, y_i, y_{i+1}) = 0.5\,[f(t_i, y_i) + f(t_i + h, y_{i+1})]$ esetén a trapéz módszert,
- $\Phi(h, t_i, y_i, y_{i+1}) = (1 - \theta) f(t_i, y_i) + \theta f(t_i + h, y_{i+1})$ esetén a θ-módszert

jelenti.

3.15. Definíció *Azokat a módszereket, amelyekre* $\Phi = \Phi(h, t_i, y_i)$ *(tehát a* Φ *függvény nem függ* y_{i+1}*-től)* explicit módszereknek *nevezzük. Amennyiben* $\Phi = \Phi(h, t_i, y_i, y_{i+1})$*, a módszert* implicitnek *nevezzük.*

Jelölje továbbra is $u(t)$ a (3.1)–(3.2) feladat megoldását, és legyen $t_\star \in (0, T]$ rögzített érték.

3.16. Definíció *Az* $e_{t_\star}(h) = y_n - u(t_\star), (nh = t_\star)$ *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*

$$\lim_{h \to 0} e_h(t_\star) = 0. \tag{3.58}$$

Ha a Φ *numerikus módszer konvergens a* $[0, T]$ *minden pontjában, akkor egyszerűen* konvergens módszernek *nevezzük. A (3.58) konvergenciájának rendjét a* Φ*-módszer* konvergenciarendjének *nevezzük.*

A Φ-módszer lokális viselkedését jól jellemzi az a közelítés, amelyet a módszerrel – a megoldásból indulva – egy lépés elvégzése után nyerünk, azaz az

$$\hat{y}_{i+1} = u(t_i) + h\Phi(h, t_i, u(t_i), \hat{y}_{i+1}) \tag{3.59}$$

egyenlettel definiált $\hat{y}_{i+1}$ közelítő érték.

3.17. Definíció *Az* $l_i(h) = \hat{y}_i - u(t_i)$ *függvényt (ahol* $t_i = ih \in \omega_h$*) a (3.57) alakú* Φ*-numerikus módszer* t_i *pontbeli* lokális diszkretizációs hibafüggvényének *nevezzük.*

Az implicit Φ-módszerek esetén a lokális diszkretizációs hibát általában nem könnyű meghatározni, mivel ahhoz ismerni kell $\hat{y}_{i+1}$ értékét a (3.59) képletből. Ennek megkerülésére vezessük be a

$$g_i(h) = -u(t_{i+1}) + u(t_i) + h\Phi(h, t_i, u(t_i), u(t_{i+1})) \tag{3.60}$$

függvényt, amelynek rendje (a kiinduló Cauchy-feladat egyenletének felhasználásával) sorfejtéssel a megoldás ismerete nélkül is meghatározható.

3.18. Definíció *Az $g_i(h)$ függvényt a (3.57) alakú Φ-numerikus módszer $t_i \in \omega_h$ pontbeli* képlethibájának *(más szóval,* lokális approximációs hibafüggvényének*) nevezzük. Azt mondjuk, hogy a Φ-numerikus módszer* r-ed rendben konzisztens*, ha*

$$g_i(h) = \mathcal{O}(h^{r+1}) \tag{3.61}$$

valamely $r > 0$ állandóval minden $t_i \in \omega_h$ rácspontban.

3.19. Megjegyzés Gyakran (és ebben a könyvben is több helyen) $g_i(h)$ helyett a

$$d_i(h) = \frac{u(t_{i+1}) - u(t_i)}{h} - h\Phi(h, t_i, u(t_i), u(t_{i+1})) \tag{3.62}$$

függvényt vizsgáljuk, amelyre nyilvánvalóan fennáll a $d_i(h) = g_i(h)/h$ összefüggés. Ez azt jelenti, hogy a Φ-numerikus módszer pontosan akkor *r-ed rendben konzisztens*, amikor

$$d_i(h) = \mathcal{O}(h^r). \tag{3.63}$$

Nyilvánvalóan a képlethiba és annak rendje azt mutatja meg, hogy a pontos megoldás milyen pontossággal elégíti ki a Φ-módszert.

3.20. Megjegyzés A (3.37), (3.50) és a (3.56) becslések alapján látható, hogy az explicit és implicit Euler-módszerek elsőrendűek, míg a trapéz-módszer másodrendű. Egyszerű számolással ellenőrizhető, hogy a θ-módszer csak a $\theta = 0.5$ esetén (azaz amikor a trapéz-módszert jelenti) másodrendű, egyébként elsőrendű.

A továbbiakban feltesszük, hogy a (3.57) alakú Φ numerikus módszerben a Φ függvény a harmadik és negyedik változójában egyaránt lipschitzes, azaz léteznek olyan $L_3 \geq 0$ és $L_4 \geq 0$ állandók, amelyek mellett tetszőleges s_1, s_2, p_1 és p_2 számok esetén

$$|\Phi(h, t_i, s_1, p_1) - \Phi(h, t_i, s_2, p_2)| \leq L_3|s_1 - s_2| + L_4|p_1 - p_2| \tag{3.64}$$

tetszőleges $t_i \in \omega_h$ és $h > 0$ esetén. Ha a Φ függvény nem függ y_i-től (avagy y_{i+1}-től), akkor $L_3 = 0$ (avagy $L_4 = 0$).

3.21. Megjegyzés A 3.14. megjegyzés alapján könnyen megmutatható, hogy az f függvény második változója szerinti lipschitzessége esetén tetszőleges θ esetén a θ-módszerre (és így az explicit és implicit Euler-módszerekre valamint a trapéz-módszerra egyaránt) alkalmasan megválasztott L_3 és L_4 állandókkal érvényes a (3.64) egyenlőtlenség.

A továbbiakban megvizsgáljuk a (3.64) tulajdonságú, r-ed rendben konzisztens Φ-módszerek konvergenciáját. Az egyszerűség kedvéért vezessük be az alábbi jelöléseket:

$$e_{t_i}(h) = e_i, \quad g_i(h) = g_i, \quad l_i(h) = l_i. \tag{3.65}$$

A globális és a lokális diszkretizációs hibafüggvények definíciója alapján fennáll az

$$\begin{aligned} e_{i+1} =& y_{i+1} - u(t_{i+1}) = (y_{i+1} - \hat{y}_{i+1}) + (\hat{y}_{i+1} - u(t_{i+1})) = \\ & (y_{i+1} - \hat{y}_{i+1}) + l_i \end{aligned} \tag{3.66}$$

összefüggés. Ezért a továbbiakban az

$$|e_{i+1}| \leq |y_{i+1} - \hat{y}_{i+1}| + |l_i| \tag{3.67}$$

egyenlőtlenség jobb oldalán lévő két tagra adunk felső becslést.

A lokális diszkretizációs hibafüggvényre az alábbi összefüggés érvényes:

$$\begin{aligned} l_{i+1} =& \hat{y}_{i+1} - u(t_{i+1}) = u(t_i) + h\Phi(h, t_i, u(t_i), \hat{y}_{i+1})) - u(t_{i+1}) = \\ & - u(t_{i+1}) + u(t_i) + h\Phi(h, t_i, u(t_i), u(t_{i+1})) + \\ & h\left[\Phi(h, t_i, u(t_i), \hat{y}_{i+1})) - \Phi(h, t_i, u(t_i), u(t_{i+1}))\right] \\ & = g_i + h\left[\Phi(h, t_i, u(t_i), \hat{y}_{i+1})) - \Phi(h, t_i, u(t_i), u(t_{i+1}))\right]. \end{aligned} \tag{3.68}$$

Ezért, felhasználva a (3.64) feltételt,

$$|l_{i+1}| \leq |g_i| + hL_4|\hat{y}_{i+1} - u(t_{i+1})| = |g_i| + hL_4|l_{i+1}| \tag{3.69}$$

Így, (3.69) alapján, kellően kis h esetén érvényes a

$$|l_{i+1}| \leq \frac{1}{1 - hL_4}|g_i| \tag{3.70}$$

egyenlőtlenség.

3.22. Megjegyzés A (3.70) egyenlőtlenség azt is mutatja, hogy a lokális diszkretizációs hiba rendje nem lehet kisebb, mint a módszer konzisztenciarendje. (Tehát egy r-ed rendben konzisztens módszer esetén a lokális diszkretizációs hiba is legalább r-ed rendben tart nullához.)

Térjünk át a (3.67) jobb oldali első tagjának becslésére.

$$
\begin{aligned}
&|y_{i+1}-\hat{y}_{i+1}| = \\
&\quad |(y_i + h\Phi(h,t_i,y_i,y_{i+1})) - (u(t_i) + h\Phi(h,t_i,u(t_i),\hat{y}_{i+1}))| \leq \\
&\quad |e_i| + h|\Phi(h,t_i,y_i,y_{i+1}) - \Phi(h,t_i,u(t_i),\hat{y}_{i+1})| \leq \\
&\quad |e_i| + hL_3|y_i - u(t_i)| + hL_4|y_{i+1} - \hat{y}_{i+1}| = \\
&\quad (1+hL_3)|e_i| + hL_4|y_{i+1} - \hat{y}_{i+1}|.
\end{aligned}
\tag{3.71}
$$

Így (3.71) alapján, kellően kis h esetén érvényes az

$$
|y_{i+1} - \hat{y}_{i+1}| \leq \frac{1+hL_3}{1-hL_4}|e_i| \tag{3.72}
$$

egyenlőtlenség.

A (3.70) és a (3.72) felhasználásával a (3.67) egyenlőtlenség átírható az

$$
|e_{i+1}| \leq \frac{1+hL_3}{1-hL_4}|e_i| + \frac{1}{1-hL_4}|g_i| \tag{3.73}
$$

alakra. Vezessük be a

$$
\mu = \mu(h) = \frac{1+hL_3}{1-hL_4}, \quad \chi = \chi(h) = \frac{1}{1-hL_4} \tag{3.74}
$$

jelöléseket.

3.23. Megjegyzés A (3.74) jelölések mellett

$$
\mu = 1 + h\frac{L_3+L_4}{1-hL_4} \tag{3.75}
$$

és így $\mu = 1 + \mathcal{O}(h)$. Ezért választhatók olyan h_0, μ_0 és λ_0 állandók[7], amelyek mellett

$$
\mu = \mu(h) \leq 1 + \mu_0 h, \quad \chi = \chi(h) \leq \chi_0, \quad \forall h \in (0, h_0). \tag{3.76}
$$

A (3.74) jelöléssel (3.73) felírható az

$$
|e_{i+1}| \leq \mu|e_i| + \chi|g_i| \tag{3.77}
$$

alakban. Rekurzív módon alkalmazva a (3.77) relációt, a következő egyenlőtlenséget nyerjük:

$$
\begin{aligned}
|e_n| &\leq \mu|e_{n-1}| + \chi|g_{n-1}| \leq \mu\left[\mu|e_{n-2}| + \chi|g_{n-2}|\right] + \chi|g_{n-1}| = \mu^2|e_{n-2}| \\
&+ \chi\left[\mu|g_{n-2}| + |g_{n-1}|\right] \leq \ldots \leq \mu^n|e_0| + \chi\sum_{i=0}^{n-1}\mu^i|g_{n-1-i}| \\
&\leq \mu^n\left[|e_0| + \chi\sum_{i=0}^{n-1}|g_{n-1-i}|\right].
\end{aligned}
\tag{3.78}
$$

[7] Például, a $h_0 = \frac{1}{2L_4}, \mu_0 = 2(L_3+L_4)$ és $\chi_0 = 2$ egy alkalmas megválasztás.

A (3.76) összefüggés alapján, minden $h \in (0, h_0)$ esetén $\chi \leq \chi_0$ és

$$\mu^n \leq (1 + \mu_0 h)^n \leq \exp(\mu_0 h n) = \exp(\mu_0 t_\star). \tag{3.79}$$

Ezért (3.78) alapján érvényes az

$$|e_n| \leq \exp(\mu_0 t_\star) \left[|e_0| + \chi_0 \sum_{i=0}^{n-1} |g_{n-1-i}| \right] \tag{3.80}$$

becslés. Mivel a Φ-módszer p-ed rendben konzisztens, ezért a (3.58) definíció alapján, megfelelően kis h esetén valamely $c_0 \geq 0$ állandóval fennáll a $|g_i| \leq c_0 h^{r+1}$ egyenlőtlenség. Ezért kis h esetén

$$\sum_{i=0}^{n-1} |g_{n-1-i}| \leq n c_0 h^{r+1} = c_0 t_\star h^r. \tag{3.81}$$

Összevetve a (3.80) és a (3.81) formulákat, az

$$|e_n| \leq \exp(\mu_0 t_\star) \left[|e_0| + c_1 h^r \right] \tag{3.82}$$

becslést nyerjük, ahol $c_1 = \chi_0 c_0 t_\star$=állandó. Mivel $e_0 = 0$, ezért a (3.82) alapján beláttuk a következő állítást.

3.24. Tétel *Tegyük fel, hogy a (3.57) képlettel definiált Φ numerikus módszer*

- *r-ed rendben konzisztens;*
- *a módszert definiáló Φ függvény folytonos, és érvényes rá a (3.64) Lipschitz-feltétel.*

Ekkor a Φ-módszer r-ed rendben konvergens a $[0, T]$ intervallumon.

3.25. Következmény A 3.20. és a 3.21. megjegyzések alapján a θ-módszer $\theta = 0.5$ esetén másodrendben, egyébként pedig elsőrendben konvergens. Ezért tehát az explicit és az implicit Euler-módszer elsőrendben, a trapéz-módszer pedig másodrendben konvergens.

3.2. Az egylépéses módszerek általános alakjának vizsgálata

Az előzőekben a konvergenciát a (3.77) összefüggésből vezettük le, mégpedig a benne szereplő tagok két tulajdonságából:

- a módszer konzisztens, azaz $g_i = \mathcal{O}(h^{r+1})$ valamely r pozitív számmal, emellett a $\chi(h)$ függvény korlátos;

- a $\mu(h)$ függvényre a (3.76) nagyságrendi becslés teljesül. Ez a tulajdonság azt fejezi ki, hogy az egyik időrétegről a következő időrétegre való áttérésnél a lépésszámok növelésével (azaz h csökkenésével) a hiba csak korlátosan növekedhet.

Ebben a szakaszban a második tulajdonságot pontosítjuk, és a segítségével kapcsolatot teremtünk a konzisztencia és a konvergencia között.

3.2.1. Egylépéses módszerek operátoros alakban

Tekintsük ismét a (3.1)–(3.2) Cauchy-feladatot, azaz a

$$\frac{du}{dt} = f(t,u), \quad t \in [0,T], \tag{3.83}$$

$$u(0) = u_0 \tag{3.84}$$

kezdetiérték-feladatot. Feltesszük, hogy $f : \mathbb{R}^2 \to \mathbb{R}$ (azaz, az egyszerűség kedvéért a skaláris Cauchy-feladatot vizsgáljuk), és a feladatnak létezik egyetlen $u(t)$ megoldása a $[0,T]$ intervallumon. Írjuk fel ezt a feladatot *operátoregyenlet* formájában! Ehhez vezessük be a következő $\mathcal{L}$ operátort. Legyen $\mathcal{L}$ egy olyan operátor, amely valamely szabály szerint valamely adott, a $[0,T]$ intervallumon értelmezett függvényhez egy másik, szintén a $[0,T]$ intervallumon értelmezett függvényt rendeli hozzá. Konkrétan, egy $w(t) \in C^1[0,T]$ függvényhez rendelje hozzá az

$$(\mathcal{L}w)(t) = \begin{cases} w'(t) - f(t,w(t)), & \text{ha } t \in (0,T], \\ w(0), & \text{ha } t = 0 \end{cases}$$

függvényt. Az $\mathcal{L}$ operátort a feladat operátorának nevezzük. Mivel a (3.1)–(3.2) Cauchy-feladatra érvényes az egzisztencia és unicitás, ezért az $\mathcal{L}$ operátor injektív.[8] Ezért valamely adott $\beta(t)$ függvény esetén az $\mathcal{L}u = \beta(t)$ feladatnak létezik egyértelmű megoldása. Legyen

$$\varphi(t) = \begin{cases} 0, & \text{ha } t \in (0,T], \\ u_0, & \text{ha } t = 0 \end{cases}$$

függvény. Ekkor az

$$\mathcal{L}u = \varphi \tag{3.85}$$

[8] Az $\mathcal{L}$ operátor injektivitásához azt kell megmutatni, hogy ha $w_1(t)$ és $w_2(t)$ olyan C^1-beli függvények, amelyekre $(\mathcal{L}w_1)(t) = (\mathcal{L}w_2)(t)$, akkor $w_1(t) = w_2(t)$. A feltétel alapján $w_1'(t) - f(t,w_1(t)) = w_2'(t) - f(t,w_2(t))$ a $(0,T]$ intervallumon, és $w_1(0) = w_2(0)$. Jelölje $g(t) = w_1'(t) - f(t,w_1(t)) = w_2'(t) - f(t,w_2(t))$ közös értéket! Ekkor tehát $g(t)$ egy adott, folytonos függvény. Nyilvánvalóan a $w' = f(t,w) + g(t)$, $w(0) = \textit{adott}$ kezdetiérték-feladatnak létezik egyértelmű megoldása, hiszen a jobb oldalon szereplő függvény a második változójában ugyanazon állandóval lipschitzes, mint az f függvény. Ezért ennek a feladatnak is létezik egyértelmű megoldása. Mivel g definíciójából $w_1'(t) = f(t,w_1(t)) + g(t)$ és $w_2'(t) = f(t,w_2(t)) + g(t)$, valamint $w_1(0) = w_2(0)$, ezért a kezdetiérték-feladat megoldásának előzőekben belátott egyértelműségének következtében $w_1(t) = w_2(t)$.

operátoregyenletnek, amely éppen a (3.1)–(3.2) Cauchy-feladatot jelenti, szintén létezik egyértelmű megoldása. A (3.85) egyenlet felírható az operátor értelmezési tartományának alkalmas leszűkítésével más alakban is. Nevezetesen, legyen $\mathcal{L}$ az az operátor, amely értelmezve van a

$$\mathrm{dom}(\mathcal{L}) = \{w \in C^1[0,T] : w(0) = u_0\}$$

halmazon, és a hozzárendelés legyen a

$$(\mathcal{L}w)(t) = w'(t) - f(t, w(t)), \quad t \in (0,T] \tag{3.86}$$

szabály szerinti. Ekkor $\mathcal{L}$ egy olyan operátor, amely a $[0,T]$-n értelmezett, megfelelően sima, és a $t = 0$ pontban adott értékű w függvényhez egy másik, a $(0,T]$ intervallumon értelmezett $\mathcal{L}w$ függvényt feleltet meg. Ekkor az $\mathcal{L}$ operátort két adat definiálja: az u_0 skalár és az f kétváltozós függvény,[9] és a (3.83)–(3.84) Cauchy-feladat felírható az alábbi operátoregyenlet alakjában:

$$\mathcal{L}u = 0, \tag{3.87}$$

azaz keressük azt az $u(t)$ függvényt a $\mathrm{dom}(L)$ halmazból, amelyhez az $\mathcal{L}$ operátor a $(0,T]$ intervallumon értelmezett, azonosan nulla függvényt rendeli hozzá. (Mivel feltételeztük, hogy a (3.83)–(3.84) Cauchy-feladatnak létezik egyértelmű megoldása, ezért a (3.87) egyenlet is egyértelműen megoldható.) Fogalmazzuk át a feladatunkat: keressük azon $u(t)$ függvényt, amely a $t = 0$ pontban adott, és ismerjük az L képét (amely nulla), azaz feladatunk $u(0)$ és $\mathcal{L}u$ ismeretében az u ismeretlen függvény meghatározása a $t \in (0,T]$ pontokban.

A numerikus megoldás meghatározásához tekintsük az

$$\omega_h := \{0 = t_0 < t_1 < \ldots < t_{N-1} < t_N = T\}$$

finomodó, változó lépéshosszúságú rácsháló sorozatát. Jelölje $h_i = t_{i+1} - t_i$, $i = 0, 1, , \ldots, N-1$ és $h = T/N$. Feltesszük, hogy a rácspontok számának növelésével a rácsháló mindenütt finomodik, azaz létezik olyan $0 < c < \infty$ állandó, amelyre minden N esetén

$$h_i \leq ch, \quad i = 1, 2, \ldots, N. \tag{3.88}$$

Továbbra is feltesszük, hogy a rögzített $t^\star \in [0,T]$ pont mindegyik rácsháló az eleme. Egy ilyen rögzített rácsháló jelölje n azt a indexet, amelyre $h_0 + h_1 + \ldots + h_{n-1} = t^\star$. Jelölje $\omega_h^0 = \omega_h \backslash \{t_0 = 0\}$ rácsot, $\mathbb{F}(\omega_h)$ és $\mathbb{F}(\omega_h^0)$ az ω_h illetve az ω_h^0 rácsokon értelmezett rácsfüggvények vektorterét.

[9] Az $\mathcal{L}$ operátor ezen bevezetése valójában a következőt jelenti. Legyen $c \in \mathbb{R}$ egy tetszőleges adott szám, $g : \mathbb{R}^2 \to \mathbb{R}$ egy adott függvény. Jelölje $\mathcal{L}_{c,g}$ azt az operátort, amelyre $\mathrm{dom}(\mathcal{L}_{c,g}) = \{w \in C^1[0,T] : w(0) = c\}$, és $(\mathcal{L}_{c,g}w)(t) = w'(t) - g(t, w(t)), \quad t \in (0,T]$. Jelölje $\mathcal{L}$ a $c = u_0$ és $g = f$ megválasztású $\mathcal{L}_{c,g}$ operátort, azaz $\mathcal{L} = \mathcal{L}_{u_0,f}$.

Legyen $\mathcal{N}_h$ egy olyan operátor, amely az ω_h pontjaiban értelmezett rácsfüggvényekhez egy másik, szintén az ω_h pontjaiban értelmezett rácsfüggvényt rendel hozzá, azaz $\mathcal{N}_h : \mathbb{F}(\omega_h) \to \mathbb{F}(\omega_h)$ adott leképezés. Konkrétan, egy $w_h \in \mathbb{F}(\omega_h)$ rácsfüggvényhez rendelje hozzá az

$$(\mathcal{N}_h w_h)(t_n) = \begin{cases} \Phi(h_n, t_{n-1}, w_h(t_{n-1}), w_h(t_n)), & \text{ha } t_n \in \omega_h^0, \\ w_h(0), & \text{ha } t_n = 0 \end{cases} \tag{3.89}$$

rácsfüggvényt, ahol $\Phi(h_n, t_{n-1}, w_h(t_{n-1}), w_h(t_n))$ az adott numerikus módszert leíró függvény. A módszer realizálhatóságához szükséges feltennünk, hogy rögzített első három változó mellett a Φ függvény a negyedik változójában invertálható, azaz a $g(s) := \Phi(h^\star, t^\star, w^\star, s)$ függvény injektív. Az $\mathcal{N}_h$ operátort a diszkretizációs módszer operátorának nevezzük. Vegyük észre, hogy az $\mathcal{N}_h : \mathbb{F}(\omega_h) \to \mathbb{F}(\omega_h)$ operátor injektív.[10] Ezért tetszőleges rögzített $\beta_h \in \mathbb{F}(\omega_h)$ rácsfüggvény esetén az $\mathcal{N} w_h = \beta_h$ feladatnak létezik egyértelmű megoldása. Bevezetve a

$$\varphi_h(t) = \begin{cases} 0, & \text{ha } t \in \omega_h^0, \\ u_0, & \text{ha } t = 0 \end{cases}$$

rácsfüggvényt, akkor az

$$\mathcal{N}_h y_h = \varphi_h \tag{3.90}$$

operátoregyenlet éppen a (3.1)–(3.2) Cauchy-problémához tartozó diszkretizált feladatot jelenti.

3.1. Példa *Határozzuk meg az explicit Euler-módszerhez tartozó $\mathcal{N}_h$ operátort!*

Mint azt már láttuk, az explicit Euler-módszer esetén az ω_h rácsháló pontjaiban a közelítéseket a (3.26) képlet alapján definiáljuk, ezért az explicit Euler-módszer operátora

$$(\mathcal{N}_h w_h)(t_n) = \begin{cases} \frac{w_h(t_n) - w_h(t_{n-1})}{h_n} - f(t_{n-1}, w_h(t_{n-1})), & \text{ha } t_n \in \omega_h^0, \\ w_h(0), & \text{ha } t_n = 0 \end{cases}$$

alakú.

A diszkretizált feladat ekvivalens módon felírható más alakban is. Nevezetesen, az $\mathcal{N}_h$ operátor értelmezési tartományának alkalmas megválasztásával az egylépéses numerikus módszerek felírhatók

$$\mathcal{N}_h y_h = 0 \tag{3.91}$$

[10] Ennek belátásához azt kell megmutatni, ha w_h^1 és w_h^2 két olyan $\mathbb{F}(\omega_h)$-beli rácsfüggvény, amelyekre $\mathcal{N}_h w_h^1 = \mathcal{N}_h w_h^2$, akkor $w_h^1 = w_h^2$. A feltételünk alapján, figyelembe véve az $\mathcal{N}_h$ operátor definícióját, $w_h^1(t_0 = 0) = w_h^2(t_0 = 0)$. Mivel $\Phi(h_1, t_0, w_h(t_0 = 0), w_h^1(t_1)) = \Phi(h_1, t_0, w_h(t_0 = 0), w_h^2(t_1))$, jelölje ezt a közös értéket g_1. Feltételeink szerint ekkor $\Phi(h_1, t_0, w_h(t_0 = 0), w_h^1(t_1)) = g_1$ és a $\Phi(h_1, t_0, w_h(t_0 = 0), w_h^2(t_1)) = g_1$ egyenlőségekből a $w_h^1(t_1)$ és a $w_h^2(t_1)$ ismeretlenek egyértelműen meghatározhatók, azaz $w_h^1(t_1) = w_h^2(t_1)$. Folytatva a gondolatmenetet, a $w_h^1 = w_h^2$ egyenlőség már közvetlenül következik.

alakban, ahol $y_h \in \mathbb{F}(\omega_h)$ az ismeretlen rácsfüggvény, $\mathcal{N}_h : \mathbb{F}(\omega_h) \to \mathbb{F}(\omega_h^0)$ típusú operátor, amelynek értelmezési tartománya a

$$\mathrm{dom}(\mathcal{N}_h) = \{w_h \in \mathbb{F}(\omega_h) : \quad w_h(t_0 = 0) = u_0\}$$

halmaz.

3.2. Megjegyzés Mivel $\mathcal{N}_h : \mathbb{F}(\omega_h) \to \mathbb{F}(\omega_h^0)$ típusú operátor, és az $\mathbb{F}(\omega_h)$ valamint az $\mathbb{F}(\omega_h^0)$ vektorterek különböző dimenziójúak ($dim(\mathbb{F}(\omega_h)) = N + 1$, $dim(\mathbb{F}(\omega_h^0)) = N$), ezért $\mathcal{N}_h$ a teljes $\mathbb{F}(\omega_h)$ téren nyilván nem injektív, azaz a (3.91) feladatnak nem létezik egyértelmű megoldása. Ezért szükséges $\mathbb{F}(\omega_h)$ egy olyan alkalmas $\mathrm{dom}(\mathcal{N}_h) \subset \mathbb{F}(\omega_h)$ halmazra való leszűkítése, amelyre $dim(\mathrm{dom}(\mathcal{N}_h)) = dim(\mathbb{F}(\omega_h^0))$, és ezen az $\mathcal{N}_h$ operátor invertálható. Ezt az $\mathbb{F}(\omega_h)$-beli rácsfüggvények $t = t_0 = 0$ pontbeli értékének rögzítésével érhetjük el, hiszen ezt az értéket a kezdetiérték-feladatból ismerjük.

Mint azt már láttuk, az explicit Euler-módszer esetén az ω_h rácsháló pontjaiban a közelítéseket a (3.26) képlet alapján definiáljuk, azaz az ismeretlen $y_h \in \mathbb{F}(\omega_h)$ rácsfüggvényt az

$$y_h(t_n) = y_h(t_{n-1}) + h_n f(t_{n-1}, y_h(t_{n-1})), \quad i = 0, 1, \ldots, N, \tag{3.92}$$

módon határozzuk meg, ahol $y_h(0) = u_0$. (Mivel $y_h(t_n)$ az ismeretlen $u(t)$ függvény t_n pontbeli közelítése, ezért ez a megválasztás természetes.) Tehát a (3.26) iterációban az $n = 0$ értékhez tartozó $y_h(0)$ adott érték.

Mindezek alapján az $\mathcal{N}_h$ operátort a következő módon határozzuk meg: valamely $w_h \in \mathbb{F}(\omega_h)$ rácsfüggvényhez rendelje hozzá az

$$(\mathcal{N}_h w_h)(t_n) = \frac{w_h(t_n) - w_h(t_{n-1})}{h_n} - f(t_{n-1}, w_h(t_{n-1})) \tag{3.93}$$

rácsfüggvényt, ahol $n = 1, 2, \ldots, N$ és $h_1 + h_2 + \ldots + h_n = T$. Nyilvánvalóan ekkor $\mathcal{N}_h w_h \in \mathbb{F}(\omega_h^0)$.

Az operátor értelmezési tartományát értelmezzük a

$$\mathrm{dom}(\mathcal{N}_h) = \{w_h \in \mathbb{F}(\omega_h), \quad w_h(t_0 = 0) = u_0\} \tag{3.94}$$

módon.

Könnyen látható, hogy az $\mathcal{N}_h$ operátor ilyen megválasztása esetén a (3.91) egyenlet éppen a (3.92) alakú explicit Euler-módszert jelenti, és az operátoregyenlet megoldása a következő feladatot jelenti: keressük azt az $y_h \in \mathbb{F}(\omega_h)$ rácsfüggvényt, amelynek ismerjük értékét a $t = t_0 = 0$ rácspontban, továbbá adott az $\mathcal{N}_h$-képe. (Esetünkben ezek $y_h(0) = u_0$ és $\mathcal{N}_h y_h = 0 \in \mathbb{F}(\omega_h^0)$.)

3.2.2. Egylépéses módszerek konzisztenciája

A numerikus módszerek tanulmányozásának lényege annak meghatározása, hogy a különböző approximációk (az $\mathcal{N}_h$ képzési szabálya és $y_h(0)$ megválasztása esetén) milyen hibák keletkeznek, és ezek a hibák hogyan viselkednek az algoritmus realizálása során. Ebben a szakaszban azt vizsgáljuk meg, hogy az $\mathcal{N}_h$ operátor milyen hibával approximálja az $\mathcal{L}_h$ operátort a pontos megoldáson. (A következő szakaszban megvizsgáljuk, hogy az algoritmus lépéseinek során ezek a hibák hogyan viselkednek a rácsháló finomításával.) Ez tehát azt jelenti, hogy a $t_n \in \omega_h$ rácspontokban az $(\mathcal{L}_h u)(t_n)$ értékeket hasonlítjuk össze az $u(t)$ megoldás rácspontbeli értékeiből képzett rácsfüggvény $\mathcal{N}_h$-képével. Pontosabban, bevezetve a $\mathcal{P}_h : C[0,T] \to \mathbb{F}(\omega_h)$ operátort az

$$(\mathcal{P}_h u)(t_n) = u(t_n) \tag{3.95}$$

egyenlőséggel, célunk a

$$d_n = |(\mathcal{L}u)(t_n) - (\mathcal{N}_h \mathcal{P}_h u)(t_n)|, \quad t_n \in \omega_h \tag{3.96}$$

kifejezés becslése.

Vegyük észre, hogy az operátorok definíciója következtében $t_0 = 0$ pontban $(\mathcal{L}u)(t_0) = u(t_0)$, és $(\mathcal{N}_h \mathcal{P}_h u)(t_0) = (\mathcal{P}_h u)(t_0) = u(t_0)$, azaz a $t_0 = 0$ pontban a hiba nulla. Mivel a $t_n \in \omega_h^0$ pontokban $(\mathcal{L}u)(t_n) = 0$, ezért az ezen pontokhoz tartozó ún. lokális approximációs hiba

$$d_n = |(\mathcal{N}_h \mathcal{P}_h u)(t_n)|, \quad t_n \in \omega_h^0. \tag{3.97}$$

Célunk a $d_n = d_n(h)$ h-tól való függésének vizsgálata.

3.3. Definíció *Az $\mathcal{N}_h$ operátorral leírt numerikus módszert konzisztensnek nevezzük az $\mathcal{L}$ operátorú feladattal az $F[0,T] \subset \mathrm{dom}(L)$ függvényosztályon, ha*

$$\lim_{h \to 0} d_n(h) = 0. \tag{3.98}$$

Azt mondjuk, hogy p-ed rendben konzisztens, amikor $d_n(h) = \mathcal{O}(h^p)$.

3.4. Példa *Mint ismeretes, az explicit Euler-módszer (3.94)-(3.93) alakú $\mathcal{N}_h$ operátora elsőrendben konzisztens a (3.86) alakú $\mathcal{L}$ operátorral az $F[0,T] = C^2[0,T]$ függvényosztályon.*

Ez az állítás a $\mathcal{P}_h u = u_h$ és az $u_n = u(t_n)$ jelölések alkalmazásával az alábbi módon nyerhető.

$$\begin{aligned}
&(\mathcal{N}_h \mathcal{P}_h u)(t_n) = (\mathcal{N}_h u_h)(t_n) = \frac{u_n - u_{n-1}}{h_n} - f(t_{n-1}, u_{n-1}) = \\
&\frac{(u_{n-1} + h_n u'(t_{n-1}) + \mathcal{O}(h_n^2) - u_{n-1}}{h_n} - f(t_{n-1}, u_{n-1}) = \\
&u'(t_{n-1}) - f(t_{n-1}, u_{n-1}) + \mathcal{O}(h_n) = \\
&u'(t_{n-1}) - f(t_{n-1}, u(t_{n-1})) + \mathcal{O}(h_n) = (\mathcal{L}u)(t_{n-1}) + \mathcal{O}(h_n).
\end{aligned} \tag{3.99}$$

Mivel $(\mathcal{L}u)(t_{n-1}) = 0$, ezért

$$d_n = |(\mathcal{N}_h \mathcal{P}_h u)(t_n)| = \mathcal{O}(h_n),$$

azaz a séma elsőrendben konzisztens. Emellett,

$$d_n = \frac{h_n}{2} u''(t_n) + \mathcal{O}(h_n) \tag{3.100}$$

minden n esetén.

3.2.3. Egylépéses módszerek konvergenciája és 0-stabilitása

Egy numerikus séma megadásával természetesen nem a konzisztencia, hanem a konvergencia a célunk. Ezért feladatunk, hogy megmutassuk: finomodó felosztássorozat esetén a numerikus megoldások sorozata és a pontos megoldás eltérése hogyan viselkedik.

Legyen

$$h = \max_{1 \le n \le N} h_n$$

és tegyük fel, hogy Nh korlátos $h \to 0$ esetén. (Tehát a $[0, T]$ intervallum mindegyik részén finomítunk.) Ezt a$h_i \le ch$ feltétellel, ahol c egy h-tól független állandó, biztosítani tudjuk. Legyen $t^\star \in (0, T]$ egy olyan rögzített pont, amelyre $t^\star \in \omega_h$ minden vizsgált $h > 0$ érték esetén.

3.5. Definíció *Az $\mathcal{N}_h$ operátorral leírt numerikus módszert konvergensnek nevezzük, ha minden $t^\star \in (0, T]$ esetén*

$$e_n(h) = y_h(t_n) - u(t^\star) \tag{3.101}$$

ún. globális hiba $h \to 0$ esetén nullához tart, ahol $t_n = t^\star$. Azt mondjuk, hogy q-ed rendben konvergens, amikor $e_n(h) = \mathcal{O}(h^q)$.

3.6. Megjegyzés Fontos kérdés, hogy konzisztencia és konvergencia esetén van-e kapcsolat a két rend között. Megmutatjuk, hogy általában a két rend egymással egyenlő.

Felmerül a kérdés: milyen kapcsolat van a konzisztencia és a konvergencia között? Következik-e a konvergencia a konzisztenciából? Megmutatjuk, hogy a válasz nemleges: önmagában a konzisztencia nem garantálja a numerikus módszer konvergenciáját, ehhez egy újabb tulajdonság, a stabilitás is szükséges.

3.7. Definíció *Az $\mathcal{N}_h$ operátorral leírt numerikus módszert 0-stabilnak (zéró-stabilnak) nevezzük, ha léteznek olyan h_0 és K pozitív állandók, hogy minden $h < h_0$ esetén két tetszőleges $\mathbb{F}(\omega_h)$-beli x_h és z_h rácsfüggvényre érvényes az*

$$|x_h(t_n) - z_h(t_n)| \le K\{|x_h(t_0) - z_h(t_0)| + \max_{1 \le j \le N} |\mathcal{N}_h x_h(t_j) - \mathcal{N}_h z_h(t_j)|\} \tag{3.102}$$

egyenlőtlenség tetszőleges $n = 1, 2, \ldots, N$ indexre.

A 0-stabilitás tehát azt fejezi ki, hogy ha két $\mathbb{F}(\omega_h)$-beli rácsfüggvény olyan, hogy a

- a $t = t_0$ pontban közel vannak egymáshoz,
- az $\mathcal{N}_h$-képük $\mathbb{F}(\omega_h^0)$-ben is közel vannak egymáshoz,

akkor maguk a függvények is pontonként közel vannak egymáshoz $\mathbb{F}(\omega_h)$-ben. Ez azt jelenti, hogy a numerikus módszert leíró adatokra (az $\mathcal{N}_h$ képzési szabályra és a rácsfüggvény $t_0 = 0$ pontbeli értékének megválasztására) nézve a numerikus módszer stabil: ezek kis megváltoztatása esetén a numerikus megoldás is csak korlátosan változik.

3.8. Megjegyzés Vegyük észre, hogy a 0-stabilitás fogalma a numerikus séma belső tulajdonsága, nincs kapcsolata azzal a folytonos feladattal, amelynek numerikus megoldására alkalmazni szeretnénk.

A 0-stabilitás nemcsak a numerikus módszer stabilitását, hanem az injektivitását is biztosítja rögzített kezdeti értékek esetén. Ezért érvényes az alábbi állítás.

3.9. Tétel *A*

$$\mathrm{dom}(\mathcal{N}_h) := \{w_h \in \mathbb{F}(\omega_h) : \ w_h(t_0) \ \textit{rögzített}\}$$

halmazon definiált 0-stabil $\mathcal{N}_h$ operátor invertálható.

Bizonyítás. Azt kell megmutatnunk, hogy az $\mathcal{N}_h$ operátor injektív $\mathrm{dom}(\mathcal{N}_h)$ halmazon, azaz ha két $x_h, z_h \in \mathrm{dom}(\mathcal{N}_h)$ rácsfüggvényre $x_h \neq z_h$, akkor $\mathcal{N}_h x_h \neq \mathcal{N}_h z_h$. Ehhez azt kell belátni, hogy ha $\mathcal{N}_h x_h = \mathcal{N}_h z_h$, akkor $x_h = z_h$. Ez viszont a 0-stabilitás definíciójából nyilvánvaló, hiszen $\mathcal{N}_h x_h = \mathcal{N}_h z_h$ következtében

$$\max_{1 \le j \le N} |\mathcal{N}_h x_h(t_j) - \mathcal{N}_h z_h(t_j)| = 0,$$

másrészt, mivel $x_h, z_h \in \mathrm{dom}(\mathcal{N}_h)$, ezért $x_h(t_0) - z_h(t_0) = 0$. Így a (3.102) egyenlőtlenség alapján $|x_h(t_n) - z_h(t_n)| = 0$ minden $n = 0, 1, \ldots, N$ esetén, amely a bizonyítandó $x_h = z_h$ egyenlőséget jelenti. □

3.10. Megjegyzés A 3.9. tétel tehát azt bizonyítja, hogy az $\mathcal{N}_h$ operátor a (3.89) képlet szerinti definíciójában a Φ függvény negyedik változója szerinti injektivitása a'priori érvényes.

3.11. Következmény Egy 0-stabil $\mathcal{N}_h$-módszer megfelelően kis h lépésközű rácshálókon mindig realizálható, emellett a diszkretizált feladatra érvényes az egzisztencia, unicitás és stabilitás, azaz a diszkrét feladat korrekt kitűzésű.

3.2.4. Egylépéses módszerek alaptétele és alkalmazása az explicit Euler-módszerre

A továbbiakban megvizsgáljuk a korábban feltett kérdést: van-e kapcsolat a konzisztencia, a konvergencia és a 0-stabilitás között? A következő tétel (amelyet gyakran a numerikus analízis alaptételének is szokás nevezni) választ ad erre a kérdésre.

3.12. Tétel *Tegyük fel, hogy*

- *a (3.85) operátoregyenletnek létezik egyértelmű megoldása;*
- *az $\mathcal{N}_h : \mathbb{F}(\omega_h) \to \mathbb{F}(\omega_h)$ operátorral leírt numerikus módszer konzisztens és 0-stabil.*

Ekkor

- *a (3.90) egyenleteknek létezik y_h egyértelmű megoldása a*

$$\mathrm{dom}(\mathcal{N}_h) := \{w_h \in \mathbb{F}(\omega_h) : \ w_h(t_0) = u_0\} \tag{3.103}$$

  halmazon;

- *a numerikus megoldások sorozata konvergens, és a konvergencia rendje megegyezik a konzisztencia rendjével.*

Bizonyítás. Jelölje $u(t)$ a (3.86) operátoregyenlet megoldását. (Feltételezésünk szerint a folytonos feladat korrekt kitűzésű, tehát a megfelelő $\mathrm{dom}(\mathcal{L})$-ben $u(t)$ létezik.) Jelölje u_h az u függvény projekcióját az ω_h rácsra, azaz $u_h = \mathcal{P}_h u$. Mivel $\mathcal{N}_h$ 0-stabil, ezért a 3.11. következmény miatt a (3.91) egyenletnek (az alkalmasan megválasztott $\mathrm{dom}(\mathcal{N}_h)$ halmazon) szintén létezik egyértelmű y_h megoldása.

Ekkor a globális hiba rácsfüggvénye $e_h = y_h - u_h \in \mathbb{F}(\omega_h)$, és azt kell megmutatnunk, hogy ω_h^0 pontjaiban $h \to 0$ esetén nullához tart a normája. Alkalmazzuk a 0-stabilitás (3.102) egyenlőtlenségét az $x_h = y_h$ és $z_h = u_h$ megválasztással! Ekkor

$$\begin{aligned} |e_h(t_n)| =& |y_h(t_n) - u_h(t_n)| \\ &\le K\{|\underbrace{y_h(t_0) - u_h(t_0)}_{=0}| + \max_{1\le j\le N} |\underbrace{(\mathcal{N}_h y_h)(t_j)}_{=0} - (\mathcal{N}_h u_h)(t_j)|\} \\ &\le K \max_{1\le j\le N} |\mathcal{N}_h u_h(t_j)| = K \max_{1\le j\le N} |d_h(t_j)|, \end{aligned} \tag{3.104}$$

ahol $d_h(t_j)$ a lokális approximációs hiba. Ezért, ha a módszer p-ed rendben konzisztens, akkor $|d_h(t_j)| = \mathcal{O}(h^p)$, és ezzel az állítást bizonyító $e_h(t_n) = \mathcal{O}(h^p)$ relációt nyerjük. □

A fenti alaptétel alapján tehát az explicit Euler-módszer konvergenciájához elegendő megmutatni a módszer 0-stabilitását. (Ugyanis a 3.4. példában láttuk, hogy a módszer első rendben konzisztens.)

3.13. Megjegyzés Az állításban a konvergencia akkor is érvényben marad, ha (3.103) helyett a

$$\mathrm{dom}(\mathcal{N}_h) := \{w_h \in \mathbb{F}(\omega_h),\ w_h(t_0) = u_0^h\} \tag{3.105}$$

halmazon definiáljuk az $\mathcal{N}_h$ operátort, ahol feltesszük, hogy $\lim_{h\to 0} u_0^h = u_0$. Emellett, ha $u_0^h - u_0 = \mathcal{O}(h^p)$, akkor a konvergencia rendje is megőrződik.

3.14. Tétel *Az explicit Euler-módszer 0-stabil.*

Bizonyítás. Azt kell megmutatnunk, hogy a (3.93) módon definiált $\mathcal{N}_h$ operátorra érvényes a (3.102) egyenlőtlenség.

Legyenek $x_h, z_h \in \mathbb{F}(\omega_h)$ két tetszőleges rácsfüggvény, és vezessük be a

$$\eta_h = x_h - z_h \in \mathbb{F}(\omega_h) \tag{3.106}$$

új rácsfüggvényt, illetve a

$$b_h = \max_{1\le j\le N} |\left(\mathcal{N}_h x_h\right)(t_j) - \left(\mathcal{N}_h z_h\right)(t_j)|$$

jelölést. Ekkor a fenti jelölésekkel tetszőleges $n = 1, 2, \ldots, N$ esetén

$$b_h \ge \left| \frac{\eta_h(t_n) - \eta_h(t_{n-1})}{h_n} - [f(t_{n-1}, x_h(t_{n-1})) - f(t_{n-1}, z_h(t_{n-1}))] \right|. \tag{3.107}$$

Felhasználva a valós számokra jól ismert $|a-b| \ge |a| - |b|$ egyenlőtlenséget, ekkor

$$b_h \ge \frac{|\eta_h(t_n)|}{h_n} - \left| \frac{\eta_h(t_{n-1})}{h_n} + [f(t_{n-1}, x_h(t_{n-1})) - f(t_{n-1}, z_h(t_{n-1}))] \right|. \tag{3.108}$$

Bevezetve az $\eta_n = \eta_h(t_n)$, $x_n = x_h(t_n)$ és $z_n = z_h(t_n)$ egyszerűsítő jelöléseket, érvényes a

$$b_h \ge \frac{|\eta_n|}{h_n} - \left| \frac{\eta_{n-1}}{h_n} + [f(t_{n-1}, x_{n-1}) - f(t_{n-1}, z_{n-1})] \right| \tag{3.109}$$

becslés. Mivel az f függvény a második változójában lipschitzes az L állandóval, ezért érvényes a

$$\begin{aligned} &\left| \frac{\eta_{n-1}}{h_n} + [f(t_{n-1}, x_{n-1}) - f(t_{n-1}, z_{n-1})] \right| \le \\ &\frac{|\eta_{n-1}|}{h_n} + L|\eta_{n-1}| = \left(\frac{1}{h_n} + L\right)|\eta_{n-1}| \end{aligned} \tag{3.110}$$

becslés. Ekkor a (3.109) és a (3.110) becslésekből a

$$b_h \ge \frac{|\eta_n|}{h_n} - \left(\frac{1}{h_n} + L\right)|\eta_{n-1}| \tag{3.111}$$

egyenlőtlenséget nyerjük. Átrendezve a (3.111) becslést, az

$$|\eta_n| \leq (1 + h_n L)|\eta_{n-1}| + b_h h_n \tag{3.112}$$

becslést nyerjük tetszőleges $n = 1, 2, \ldots, N$ értékekre. Felhasználva a (3.112) becslést egymás után az egyes tagokra, ekkor a következő egyenlőtlenséget kapjuk:

$$\begin{aligned}
&|\eta_n)| \leq (1 + h_n L)|\eta_{n-1}| + b_h h_n \leq \\
&\quad (1 + h_n L)(1 + h_{n-1} L)|\eta_{n-2}| + (1 + h_n L) b_h h_{n-1} + b_h h_n \leq \\
&\quad \ldots\ldots \leq \\
&\quad (1 + h_n L)(1 + h_{n-1} L) \ldots (1 + h_1 L))|\eta_0| + \\
&\quad b_h \sum_{j=1}^{n} h_j (1 + h_{j+1} L)(1 + h_{j+2} L) \ldots (1 + h_n L).
\end{aligned} \tag{3.113}$$

A további becsléshez használjuk a pozitív x értékekre jól ismert $1 + x < \exp(x) \equiv e^x$ egyenlőtlenséget! Ez alapján $1 + h_j L < e^{L h_j}$, és így

$$(1 + h_n L)(1 + h_{n-1} L) \ldots (1 + h_1 L) < e^{L(h_1 + h_2 + \ldots h_n)} = e^{L t_n}, \tag{3.114}$$

valamint minden $j = 0, 1, \ldots, n$ esetén

$$(1 + h_{j+1} L)(1 + h_{j+2} L) \ldots (1 + h_n L) < e^{L(h_{j+1} + h_{j+2} + \ldots h_n)} = e^{L(t_n - t_j)}. \tag{3.115}$$

Mivel $e^{L(t_n - t_j)} \leq e^{L(t_n - t)}$ minden $t \in [t_{j-1}, t_j]$ esetén, ezért integrálva ezt az egyenlőtlenséget a $[t_{j-1}, t_j]$ intervallumon, a

$$\int_{t_{j-1}}^{t_j} e^{L(t_n - t_j)} dt = h_j e^{L(t_n - t_j)} \leq \int_{t_{j-1}}^{t_j} e^{L(t_n - t)} dt$$

egyenlőtlenséget kapjuk. Ezt összegezve tehát érvényes a

$$\begin{aligned}
&\sum_{j=1}^{n} h_j e^{L(t_n - t_j)} \leq \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} e^{L(t_n - t)} dt = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} e^{L t_n} e^{-L t} dt = \\
&\quad \int_{t_0}^{t_n} e^{L t_n} e^{-L t} dt = e^{L t_n} \int_{t_0}^{t_n} e^{-L t} dt = e^{L t_n} \frac{1}{L} \left(1 - e^{-L t_n}\right) = \\
&\quad \frac{1}{L} \left(e^{L t_n - 1}\right).
\end{aligned} \tag{3.116}$$

Felhasználva a (3.114) és a (3.116) becsléseket a (3.113) egyenlőtlenségben, a

$$|\eta_n| \leq e^{L t_n} |\eta_0| + \frac{1}{L} \left(e^{L t_n - 1}\right) b_h \tag{3.117}$$

becslést nyerjük. Mivel $t_n \leq T$, ezért tehát érvényes a (3.102) 0-stabilitási becslés a

$$K = \max\{e^{LT}, \frac{1}{L}\left(e^{LT-1}\right)\} \tag{3.118}$$

konstanssal. □

3.15. Megjegyzés A (3.118) képlet felhasználásával közvetlen becslés adható az explicit Euler-módszer hibájára. Ugyanis (3.104) és a konzisztencia (3.100) becslése alapján

$$|e_h(t_n)| \leq K \max_{1\leq j\leq N} |d_h(t_j)| = \max\{e^{LT}, \frac{1}{L}\left(e^{LT-1}\right)\}\frac{M_2}{2}h, \tag{3.119}$$

ahol $M_2 = \max_{[0,T]} |u''(t)|$.

3.16. Példa *Tekintsük a*

$$u' = -\sin(t), \quad u(0) = 1$$

kezdetiérték-feladatot! Milyen lépésközű rácshálón kell megoldanunk az explicit Euler-módszerrel, hogy a $t^\star = 2$ pontban a megoldás $\varepsilon = 10^{-3}$ pontosságú legyen?

Az egyenletet deriválva $u''(t) = -\cos(t)$, azaz $M_2 = \max_{[0,2]} |-\cos(t)| = 1$. A példánkban a differenciálegyenlet jobb oldala $f(t,u) = -\sin(t)$, ezért $f(t,u_1) - f(t,u_2) = 0$, azaz f lipschitzes az $L = 0$ állandóval. Könnyen láthatóan $\lim_{L\to 0} \frac{e^{Lt^\star}-1}{L} = t^\star$, ezért a (3.119) becslés alapján

$$|e_h(t_n)| \leq \max\{e^0, t^\star\}\frac{1}{2}h = h.$$

Ezért ezen a feladaton a $h = \epsilon$ megválasztású rácshálón az explicit Euler-módszer eredménye garantálja az ϵ pontosságot.

Legyen tehát $h = 10^{-3}$! (Ekkor $t^\star = t_{2000}$, hiszen a lépésszám a $t^\star = 2$ pontig $t^\star/h = 2000$.) Ezen a rácshálón az explicit Euler-módszerrel kapott numerikus megoldás $y_{2000} = -0.41569...$ Mivel a feladat pontos megoldása $u(t) = \cos(t)$, ezért összevetve a $\cos(2) = -0.41614...$ pontos megoldással, a globális hibára az $e_{2000} = 0.4548 \cdot 10^{-3}$ értéket kapjuk.

3.17. Megjegyzés Az $\mathcal{N}_h$ operátorú numerikus módszer lokális approximációs hibájának (3.97) szerinti definíciója azt mutatja meg, hogy a megoldás rácshálón való vetülete milyen pontosan elégíti ki a differenciasémát. A módszer egy másik jellemzése lehet a következő: megvizsgáljuk, hogy a $t = t_{n-1}$ pontból a t_n pontba való áttérés során mekkora eltérés van a numerikus megoldás (y_n) és azon érték között, amelyet a numerikus megoldással akkor kapnánk, ha a t_{n-1} pontból nem a közelített y_{n-1} értékből indulnánk, hanem a pontos $u(t_{n-1})$ értéket használnánk. Például, az explicit Euler-módszer esetén ez az

$$l_n = u(t_n) - [u(t_{n-1}) + h_n f(t_{n-1}, u(t_{n-1}))] \tag{3.120}$$

értéket jelenti. Könnyen látható, hogy ekkor $l_n = u(t_n) - [u(t_{n-1}) + h_n u'(t_{n-1})] = h_n d_n$, azaz $l_n = \mathcal{O}(h_n^2)$, és általában, ha $l_n = \mathcal{O}(h_n^{p+1})$, akkor a módszer p-ed rendben konzisztens.

3.3. Abszolút stabilitás

Az előző szakaszban megmutattuk, hogy egy p-ed rendben konzisztens, 0-stabil módszer p-ed rendben konvergens is, azaz $|e_n| \approx C \cdot h^p$. Ugyanakkor az elméleti becslésből a C konstans nagyon nagy is lehet, ami alkalmazhatatlanná teheti ezt a becslést. A következőkben megvizsgáljuk, hogy az explicit Euler-módszer esetén mit jelent ez a becslés, illetve milyen numerikus probléma lép is fel valójában.

3.3.1. Az explicit Euler-módszer abszolút stabilitási tartománya

3.1. Példa *Tekintsük a*

$$u'(t) = -5tu^2(t) + \frac{5}{t} - \frac{1}{t^2}, \quad u(1) = 1$$

kezdetiérték-feladatot! Milyen lépésközű rácshálón kell megoldanunk az explicit Euler-módszerrel, hogy az $[1, 25]$ *intervallumon a megoldás* $\varepsilon = 10^{-3}$ *pontosságú legyen?*

Mivel a feladat pontos megoldása az $u(t) = 1/t$ függvény, ezért a Lipschitz-állandót a Jacobi-mátrix ezen megoldásbeli alakjából határozhatjuk meg. Feladatunkra $f(t, w) = -5tw^2 + 5/t - 1/t^2$, ezért $\partial_2 f(t, 1/t) = -10t \cdot 1/t = -10$. Tehát a Lipschitz-állandó értéke $L = 10$. Másrészt,

$$M_2 = \max_{[1,25]} |u''(t)| = \max_{[1,25]} |2/t^3| = 2.$$

Ezért az elméleti hibakorlát

$$|e_n| \leq \frac{2h}{2} \max\{\frac{e^{10 \cdot 25} - 1}{10}, e^{250}\} = e^{250} h.$$

Tehát a konvergencia érvényes ugyan (hiszen $h \to 0$ esetén a $[0, 25]$ mindegyik pontjában a globális hiba nullához tart), egy rögzített rácshálón a hiba nagyon nagy is lehet! Például, a feladat által megkövetelt $\varepsilon = 10^{-3}$ pontossággal ezen feltétel szerint a $h_0 = 10^{-3} \cdot e^{-250}$ melletti $h < h_0$ lépésközű rácshálókon tudnánk csak biztosítani. Mivel a számítógépen ebben az esetben $h_0 = 0$, ezért a fenti becsléssel elvileg nem tudjuk garantálni az előírt pontosságot, azaz nem választható meg a megfelelő rácsháló.

3.2. Megjegyzés Mi lehet az oka a fenti jelenségnek? Bár az eredeti feladat stabil (hiszen a Jacobi-mátrix a megoldáson negatív, tehát aszimptotikusan stabil is), a Lipschitz-állandó meghatározásánál ennek abszolút értékével számolunk, azaz a numerikus megoldás szempontjából pontosan úgy viselkedik, mintha a Jacobi-mátrix értéke 10 lenne, azaz instabil volna.

Felmerülhet a kérdés: a fenti becslés hogyan viszonyul a valós pontossághoz? A kérdés megválaszolásához tekintsük a következő példát.

h	$e_h(2)$ értéke
0.001000	$0.145252 \cdot 10^{77}$
0.000976	$0.588105 \cdot 10^{36}$
0.000950	$0.321089 \cdot 10^{-6}$
0.000800	$0.792298 \cdot 10^{-7}$
0.000400	$0.396033 \cdot 10^{-7}$

3.5. táblázat. Numerikus eredmények $\lambda = -2100$ esetén.

3.3. Példa *Oldjuk meg az explicit Euler-módszerrel az*

$$u'(t) = \lambda(u(t) - \cos(t)) - \sin(t); \quad u(0) = 1$$

kezdetiérték-feladatot!

A feladatban $f(t, w) = \lambda(w - \cos(t)) - \sin(t)$, tehát $\partial_2 f(t, w) = \lambda$. Így tetszőleges függvényen a Jacobi-mátrix értéke λ. Vegyük észre, hogy a feladat megoldása $u(t) = \cos(t)$, a λ megválasztásától függetlenül. Határozzuk meg az $\varepsilon = 10^{-3}$ pontosságú numerikus megoldást a $t^\star = 2$ pontban! Ebben a feladatban $M_2 = 1$ és $L = |\lambda|$. Ezért $|\lambda| \geq 1$ esetén a hibára a (3.119) becslés alapján az

$$|e_h(t^\star)| \leq 0.5 h e^{2|\lambda|} \tag{3.121}$$

becslést kapjuk. Legyen $\lambda = -10$. A (3.121) becslést figyelembevéve olyan rácshálót kellene megválasztanunk, amelynek finomsága a $h_0 = 2e^{-20}10^{-3}$ értéknél nem durvább, de ez a megválasztás nem reális. Mivel ez a feltétel egy elégséges feltételt jelent, ezért arra nem ad választ, hogy mi történik az ettől eltérő, nagyobb lépésközű rácshálón.

Próbálkozzunk a rácsháló más megválasztásával! Mivel ebben a feladatban a megoldás ugyanaz, mint a a 3.16. példában, ezért az ottani rácshálóval kísérletezzünk, azaz legyen a $h = 10^{-3}$ lépésközű ekvidisztáns rácsháló! Ebben az esetben a $t^\star = 2$ ponthoz tartozó közelítés $y_{2000} = -0.416163$ lesz, azaz a hibára az $e_h(t_{2000}) = |y_{2000} - \cos(2)| = 0.161 \cdot 10^{-4}$ adódik. Tehát ez a rácsháló-megválasztás jó a kívánt pontosság eléréséhez!

Legyen most $\lambda = -2100$! Ebben az esetben az elméleti hibabecslésben szereplő rácsméretre a $h_0 = 2e^{-4200}10^{-3}$ értéket kapjuk, ami a gépi nullát jelenti. Tehát ez a becslés nem alkalmazható a rácsháló megválasztására! Próbálkozzunk ismét a $h = 10^{-3}$ lépésközű ekvidisztáns rácshálóval! Eredményül ekkor az $y_{2000} = -0.2543 \cdot 10^{77}$ értéket kapjuk, amely nyilvánvalóan alkalmatlan a $\cos(2)$ közelítésére! Ezért próbálkozzunk kisebb h értékekkel! Eredményeink a 3.5 táblázatban láthatók.

Eredményeink azt mutatják, hogy valamilyen ismeretlen ok következtében a $h = 0.000976$ és a $h = 0.000950$ értékek között történik a változás, és az addig használhatatlan

numerikus eredmény jóvá válik. Az ok meghatározásához tekintsük a

$$u'(t) = \lambda u + g(t)$$

lineáris feladatot, azaz a 3.3. példát általános inhomogenitással. Ha az explicit Euler-módszert alkalmazzuk ezen feladat megoldására, akkor az

$$y_n = y_{n-1} + h(\lambda y_{n-1} + g(t_{n-1})) = (1 + \lambda h)y_{n-1} + hg(t_{n-1}) \tag{3.122}$$

formulát nyerjük. Mivel a lokális approximációs hibára a

$$d_n = \frac{u(t_n) - u(t_{n-1})}{h} - \lambda u(t_{n-1}) - g(t_{n-1})$$

összefüggés érvényes, ezért

$$u(t_n) = (1 + \lambda h)u(t_{n-1}) + hg(t_{n-1}) + hd_n. \tag{3.123}$$

A (3.122) és a (3.123) összefüggésekből a t_n pontbeli $e_h(t_n) = e_n$ globális hibára az

$$e_n = (1 + \lambda h)e_{n-1} - hd_n \tag{3.124}$$

egyenlőséget kapjuk.

A (3.124) képlet azt mutatja, hogy az $n-1$-dik pontról az n-dik pontra való áttérésnél a globális hiba két forrásból tevődik össze. Bár a második tag h csökkenésével csökken (ezt a konzisztencia biztosítja), az első tag az $(1 + \lambda h)$-szorosára változik. Ez azt jelenti, hogy csak akkor csökken, amikor $|1 + \lambda h| < 1$. Az első esetben, amikor $\lambda = -10$ volt, akkor $1 + \lambda h = 1 + (-10) \cdot 10^{-3} = 0.99$, tehát az egyenlőtlenségi feltétel teljesült. A második esetben, amikor $\lambda = -2100$ volt, akkor $1 + \lambda h = 1 + (-2100) \cdot 10^{-3} = -1.1$, tehát az egyenlőtlenségi feltétel nem teljesült, és az előző pontbeli globális hiba nagysága az új pontbeli hiba első összetevőjében megnőtt. Ez az összetevő tag tehát a 2000. lépés után $1.1^{2000} \approx 10^{82}$-szeresére növekszik.

Mivel az $|1 + (-2100)h| < 1$ egyenlőtlenség megoldása a $h = 2/2100 \approx 0.00095238..$, ez magyarázatot ad a 3.5 táblázat eredményeire.

A fenti példa azt mutatja, hogy a 0-stabilitás csupán azt mutatja, hogy $h \to 0$ esetén hogyan viselkedik a numerikus megoldás. Arra viszont nem ad választ, hogy egy rögzített rácshálón milyen tulajdonságú a numerikus megoldás! A numerikus megoldásokat azzal jellemezhetjük, hogyan viselkednek a

$$u'(t) = \lambda u(t), \quad t > 0 \tag{3.125}$$

tesztegyenleten. Alapvető elvárásunk, hogy a numerikus megoldás valamely rögzített rácshálón jól modellezze az $u(t) = e^{\lambda t}u(0)$ megoldást. Megjegyezzük, hogy a (3.125)

tesztegyenletben $\lambda \in \mathbb{C}$ adott szám[11]. Az általánosság megszorítása a továbbiakban feltesszük, hogy $u(0) = u_0 > 0$. Mint azt megmutattuk, a megoldás stabilitása $\mathrm{Re}(\lambda)$ előjelétől függ: pozitív értéke esetén instabil, míg nem-pozitív értéke esetén stabil a megoldás. Ezért a numerikus módszerünket olyan tesztfeladatra alkalmazzuk, amikor

$$\mathrm{Re}(\lambda) \leq 0. \tag{3.126}$$

Az $u(t) = \exp(\lambda t)u(0)$ függvényre a stabilitás következtében könnyen látható, hogy a megoldás monoton csökkenő, tehát tetszőleges $t_2 > t_1$ esetén $|u(t_2)| \leq |u(t_1)|$. A megoldás alakjából az is leolvasható, hogy minden $t > 0$ esetén $u(t) > 0$, és $|u(t)| \leq |u(0)| = u_0$. (Ha $\mathrm{Re}(\lambda) < 0$, akkor az aszimptotikus stabilitás következtében a megoldás szigorúan monoton csökkenő, és $t \to \infty$ esetén nullához tart.)

Alkalmazzuk az explicit Euler-módszert a feladat megoldására az ekvidisztáns ($h_n = h$) lépésközű ω_h rácshálón! Vizsgáljuk meg, hogy a folytonos megoldás felsorolt tulajdonságai milyen feltétel mellett őrződnek meg a rácshálón! Mivel

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

ezért a numerikus megoldás a rácsháló pontjaiban

$$y_n = (1 + h\lambda)^n u_0. \tag{3.127}$$

A stabilitás diszkrét megfelelőjét jelentő

$$|y_n| \leq |y_{n-1}| \tag{3.128}$$

feltétel akkor és csak akkor teljesül, amikor

$$|1 + \lambda h| \leq 1. \tag{3.129}$$

Ez azt jelenti, hogy a (3.125) tesztegyenletben szereplő λ adott érték esetén csak azon h-lépésközű rácshálókon teljesül a (3.128) numerikus stabilitási tulajdonság, amelyekre a (3.129) feltétel érvényes.

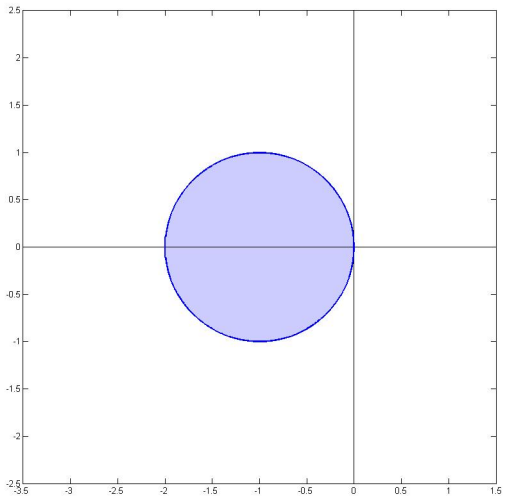
Mit jelent ez a feltétel? Vezessük be a $z = \lambda h \in \mathbb{C}$ jelölést, illetve az $R(z) = 1 + z$ függvényt! Ezen jelölésekkel az explicit Euler-módszer felírható

$$y_n = R(z)y_{n-1} \tag{3.130}$$

alakban, ahol az $R(z)$ függvényt az explicit Euler-módszer *stabilitási függvényének* nevezzük. A numerikus megoldás monoton csökkenésének

$$|R(z)| \leq 1 \tag{3.131}$$

a feltétele. Azon pontok halmazát, amelyekre a (3.131) feltétel teljesül, az explicit Euler-módszer *abszolút stabilitási tartományának* nevezzük.



3.1. ábra. Az explicit Euler-módszer abszolút stabilitási tartománya.

3.4. Megjegyzés Az explicit Euler-módszer stabilitási tartománya az $|1+z| \leq 1$ tulajdonságú komplex számok halmaza, azaz a 3.1 ábra szerinti, $\mathbb{C}^-$-beli kör. Tehát pl. valós λ esetén a feltétel $h\lambda \in [-2, 0]$, azaz $h \leq 2/(-\lambda)$. (Látható a probléma nagy $-\lambda$ esetén!) Vegyük észre, hogy a 3.3. példában, amikor a $\lambda = -10$ esetet vizsgáltuk, a $h = 10^{-3}$ megválasztás kielégíti ezt a feltételt, hiszen ezen λ esetén a korlát $h \leq 0.2$.

Fontos megjegyeznünk, hogy az abszolút stabilitási feltétel a stabilitásra, és nem a pontosság növelésére szolgál. Például, legyen a tesztegyenletben $\mathrm{Re}(\lambda) < 0$ és $u_0 = 10^{-15}$. Ekkor a pontos megoldás közel van az $u(t) = 0$ görbéhez, amely az $u_0 = 0$ megválasztáshoz tartozik. Ugyanakkor az $u_0 = 10^{-15}$ kezdeti érték megválasztást úgy is tekinthetjük, mint az $u_0 = 0$ pontos kezdeti érték perturbált (azaz, hibával terhelt) értékét. Ekkor, ha az explicit Euler-módszerben h értékét az abszolút stabilitási tartományból vesszük (azaz h-ra teljesül a (3.129) feltétel), akkor közel maradunk az $u(t) = 0$ megoldáshoz, és n növelésével a numerikus megoldás is, szigorúan monoton csökkenve nullához tart. Viszont nagyobb h esetén ez már nem teljesül: ha $|1 + \lambda h| > 1$, akkor $n \to \infty$ esetén $|1 + \lambda h|^n \to \infty$, azaz a numerikus megoldás nem marad korlátos.

Lineáris rendszerek esetén a stabilitási vizsgálat kissé bonyolultabb. Tekintsük a

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{A}\mathbf{u}(t), \quad t > 0 \tag{3.132}$$

[11] Megengedjük a komplex értéket is, hiszen, mint azt már láttuk, rendszerek esetén λ valamely mátrix sajátértéke, és így komplex mértékű is lehet.

lineáris rendszert, ahol $\mathbf{A} \in \mathbb{R}^{d\times d}$ adott, diagonalizálható mátrix. Ez azt jelenti, hogy létezik olyan $\mathbf{T}$ reguláris mátrix, amellyel a hasonlósági transzformáció után

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathrm{diag}[\lambda_1, \ldots, \lambda_d] = \Lambda, \tag{3.133}$$

ahol λ_i az $\mathbf{A}$ mátrix sajátértékei. Bevezetve $\mathbf{w}(t) = \mathbf{T}^{-1}\mathbf{u}(t)$ új változót, a (3.132) egyenlet átírható

$$\frac{d\mathbf{w}(t)}{dt} = \Lambda\mathbf{w}(t)$$

alakra. Mivel ez az egyenlet d darab skaláris tesztegyenletet jelent, ezért a stabilitásának feltétele, hogy mindegyik skaláris egyenlet stabil legyen, azaz teljesüljön a $\mathrm{Re}(\lambda_j) \leq 0$ feltétel minden $j = 1, 2, \ldots, d$ esetén.

Alkalmazzuk az explicit Euler-módszert a (3.132) feladatra! Ekkor az $\mathbf{u}(t_n) \in \mathbb{R}^d$ közelítését jelentő $\mathbf{y}_n \in \mathbb{R}^d$ vektort az

$$\mathbf{y}_n = \mathbf{y}_{n-1} + h\mathbf{A}\mathbf{y}_{n-1} \tag{3.134}$$

vektoriterációval állítjuk elő. Vezessük be az

$$\mathbf{w}_n = \mathbf{T}^{-1}\mathbf{y}_n \in \mathbb{R}^d \tag{3.135}$$

új vektort! Mivel $\mathbf{T}\Lambda\mathbf{T}^{-1} = \mathbf{A}$, ezért (3.134) felírható

$$\mathbf{y}_n = \mathbf{y}_{n-1} + h\mathbf{T}\Lambda\mathbf{T}^{-1}\mathbf{y}_{n-1},$$

azaz

$$\mathbf{T}^{-1}\mathbf{y}_n = \mathbf{T}^{-1}\mathbf{y}_{n-1} + h\Lambda\mathbf{T}^{-1}\mathbf{y}_{n-1}$$

alakban, ami az új vektorra áttérve a

$$\mathbf{w}_n = \mathbf{w}_{n-1} + h\Lambda\mathbf{w}_{n-1} \tag{3.136}$$

rekurziót jelenti. Ezért ha h értékét úgy választjuk meg, hogy $h\lambda_i$ értékek az explicit Euler-módszer stabilitási tartományában vannak, azaz

$$|1 + \lambda_i h| \leq 1, \quad i = 1, 2, \ldots, d, \tag{3.137}$$

akkor az explicit Euler-módszerrel előállított $\mathbf{w}_n$ vektorokra érvényes a

$$\|\mathbf{w}_n\| \leq \|\mathbf{w}_{n-1}\| \leq \cdots \leq \|\mathbf{w}_0\| \tag{3.138}$$

egyenlőtlenség. Ezért érvényes a

$$\begin{aligned} \|\mathbf{y}_n\| =& \|\mathbf{T}\mathbf{w}_n\| \leq \|\mathbf{T}\|\|\mathbf{w}_n\| \leq \|\mathbf{T}\|\|\mathbf{w}_0\| = \|\mathbf{T}\|\|\mathbf{T}^{-1}\mathbf{y}_0\| \\ \leq& \|\mathbf{T}\|\|\mathbf{T}^{-1}\|\|\mathbf{y}_0\| = \kappa(\mathbf{T})\|\mathbf{y}_0\| = \kappa(\mathbf{T})\|\mathbf{u}_0\| \end{aligned} \tag{3.139}$$

becslés, ahol $\kappa(\mathbf{T}) = \|\mathbf{T}\|\|\mathbf{T}^{-1}\|$ a $\mathbf{T}$ mátrix *kondíciószáma*[12]. Vegyük észre, hogy a (3.139) becslés azt mutatja, hogy rögzített lineáris rendszerre a (3.137) feltételt kielégítő h megválasztása esetén érvényes az $\|\mathbf{y}_n\| \leq$ állandó becslés, azaz a módszer numerikusan stabil. Mivel nyilvánvalóan $\kappa(T) \geq 1$, ezért általános esetben nem feltétlenül teljesül rá normában a (3.128) feltétel, azaz a kezdeti vektor normáját meghaladhatja egy későbbi közelítés normája.

3.5. Megjegyzés Ha $\mathbf{T}$ ortogonális mátrix, akkor inverze a transzponáltjával egyenlő, azaz $\mathbf{T}^{-1} = \mathbf{T}^T$. Ekkor egy tetszőleges $\mathbf{y} \in \mathbb{R}^d$ vektor és annak $\mathbf{w} = \mathbf{T}^{-1}\mathbf{y}$ transzformáltjainak euklideszi normái között az alábbi kapcsolat érvényes:

$$\begin{aligned} \|\mathbf{y}\|_2^2 &= <\mathbf{y},\mathbf{y}> = <\mathbf{Tw},\mathbf{Tw}> = <\mathbf{w},\mathbf{T}^T\mathbf{Tw}> \\ &= <\mathbf{w},\mathbf{T}^{-1}\mathbf{Tw}> = <\mathbf{w},\mathbf{w}> = \|\mathbf{w}\|_2^2, \end{aligned} \tag{3.140}$$

azaz az ortogonális transzformáció normatartó. Az ortogonális mátrixok egy további fontos tulajdonsága, hogy az euklideszi normabeli kondicionáltsági száma $\kappa_2(T) = 1$. (V.ö. [5], 61.oldal, 3.2.2 tétel.)[13]

A 3.5. megjegyzés következtében, ha $\mathbf{T}$ ortogonális mátrix, akkor a speciális euklideszi normában érvényes az $\|\mathbf{y}_n\|_2 \leq \|\mathbf{y}_0\|_2$ tulajdonság.[14]

3.6. Példa *Vizsgáljuk meg az explicit Euler-módszer alkalmazhatóságát a forrásmentes hővezetési egyenlet Δx egyenletes térbeli rácshálón való szemidiszkretizációjára, azaz a*

$$\boldsymbol{y}'(t) = \boldsymbol{A}_{\Delta x}\boldsymbol{y}(t) \tag{3.141}$$

egyenletre, ahol

$$\boldsymbol{A}_{\Delta x} = \frac{1}{(\Delta x)^2}\,\mathrm{tridiag}[1,-2,1] \in \mathbb{R}^{N\times N}$$

tridiagonális mátrix.

Mint ismeretes, egy valós $\mathbf{A}$ mátrix pontosan akkor diagonalizálható ortogonális mátrixszal, amikor szimmetrikus. (V.ö. [5], 28. oldal, 1.2.29 tétel.)[15] Mivel $\mathbf{A}_{\Delta x}$ ilyen

[12] A kondíciószám normafüggő, azaz függ az operátornorma megválasztásától. Ha hangsúlyozni akarjuk, hogy melyik normában értendő a kondíciószám, akkor ezt jelölésben is kifejezhetjük: a $\|\cdot\|_p$ operátornorma használata esetén az ebben normában értendő kondíciószámot κ_p-vel jelöljük.

[13] Ez a tulajdonság könnyen belátható az eddigi eredményeinkből is. Ugyanis a (3.138) és a (3.140) egyenlőtlenség alapján $\|\mathbf{y}_n\|_2 = \|\mathbf{w}_n\|_2 \leq \|\mathbf{w}_{n-1}\|_2 \leq \cdots \leq \|\mathbf{w}_0\|_2 = \|\mathbf{y}_0\|_2$, ezért az $\|\mathbf{y}_n\| \leq \kappa(\mathbf{T})\|\mathbf{y}_0\|$ egyenlőtlenség következtében $\kappa(\mathbf{T}) \leq 1$. Mivel az $1 = \|\mathbf{I}\|_2 = \|\ \mathbf{T}\mathbf{T}^{-1}\|_2 \leq \|\ \mathbf{T}\|_2\|\mathbf{T}^{-1}\|_2 = \kappa_2(\mathbf{T})$ becslés nyilvánvalóan igaz, ezért $\kappa_2(\mathbf{T}) = 1$.

[14] Ha egy numerikus módszer olyan, hogy az általa előállított numerikus megoldások sorozatára fennáll az $\|\mathbf{y}_n\| \leq \|\mathbf{y}_0\|$ tulajdonság, akkor a numerikus módszert ebben a normában *kontraktívnak* nevezzük.

[15] Ebben az esetben $\mathbf{T}$ transzformációs mátrixnak az $\mathbf{A}$ mátrix sajátvektoraiból összeállított mátrix választható, amelyek a $\|\cdot\|_2$ normában ortonormáltak.

tulajdonságú mátrix, ezért tehát diagonalizálható ortogonális mátrixszal, és így adott térbeli felosztás esetén a stabilitási tartományából választott h időbeli diszkretizációs lépésköz esetén a módszer stabil és ráadásul a $\|\cdot\|_2$ normában kontraktív. Mivel a mátrix mindegyik sajátértéke negatív, ezért a (3.137) feltétel a

$$h < \frac{2}{\lambda_{max}}$$

feltételt jelenti, ahol λ_{max} a maximális abszolút értékű sajátérték. Mivel az $\mathbf{A}_{\Delta x}$ mátrix sajátértékei

$$\lambda_k = -\frac{4}{(\Delta x)^2}\sin^2\frac{k\pi h}{2}, \quad k = 1, 2, \ldots, N, \tag{3.142}$$

(ahol $(N+1)\Delta x = 1$), (v.ö. (2.165) képlettel), ezért ebben a példában

$$\begin{aligned}\lambda_{max} = -\lambda_N &= \frac{4}{(\Delta x)^2}\sin^2\frac{(\Delta x)N\pi}{2} = \frac{4}{(\Delta x)^2}\sin^2\frac{(1-\Delta x)\pi}{2} \\ &= \frac{4}{(\Delta x)^2}\cos^2\frac{\pi\Delta x}{2}.\end{aligned} \tag{3.143}$$

Tehát rögzített térbeli rácsháló esetén (rögzített Δx mellett) az időbeli diszkretizáció lépésközére a

$$h < \frac{(\Delta x)^2}{2\cos^2\frac{\pi\Delta x}{2}} = \mathcal{O}((\Delta x)^2) \tag{3.144}$$

feltételt kapjuk. Ez $\Delta x \to 0$ esetén a $h \approx$ állandó $\cdot\,(\Delta x)^2$ feltétellel biztosítható, azaz az időbeli diszkretizációs lépésköz a térbeli lépésköz négyzetével arányos. (Mivel $\Delta x \to 0$ esetén $\cos^2(\pi\Delta x/2)$ alulról tart az 1 határértékhez, ezért tetszőleges Δx lépésközre az arányossági állandó értéke 0.5, tehát tetszőleges Δx esetén a korlát a $h/(\Delta x)^2 \le 0.5$ feltétellel biztosítható. Ez természetesen egy elégséges feltétel, és rögzített Δx esetén a korlát kicsit növelhető.)

Mint láttuk, az explicit Euler-módszer esetén az abszolút stabilitási feltétel túlságosan is megszorító lehet. Ugyanis, nagy abszolút értékű λ esetén csak nagyon kis h választható, amely mellett stabil marad a módszer. Ez ugyan nagy pontosságú numerikus megoldáshoz vezet, de nagy számítási munkával is jár. (Egy rögzített $t^\star$ időpont eléréséhez $t^\star/h$ számú rétegen kell a közelítést meghatározni, amely kis h esetén jelentős munkát eredményez.) Tehát, ha "be is érnénk" kisebb pontossággal, mert nem szükséges a numerikus megoldás túlságosan nagy pontossága, akkor sem választhatunk nagyobb h lépésközt, hiszen ezen megválasztás mellett az explicit Euler-módszer instabillá válna. Az explicit Euler-módszer ezen tulajdonsága az alkalmazhatósága szempontjából mindenféleképpen hátrányt jelent.

3.3.2. További módszerek abszolút stabilitási tartománya

A továbbiakban megvizsgáljuk, hogy más numerikus módszer esetén mi mondható el az abszolút stabilitási tartományról. Tudunk-e olyan módszereket megadni, amelyekre a lépésköz megválasztását csak a kívánt pontosság határozza meg? Elsőnek az implicit Euler-módszert tekintjük. (V.ö. a 3.1.4 szakasszal.)

Mint ismeretes, az implicit Euler-módszer esetén a numerikus közelítést az

$$y_n = y_{n-1} + h_n f(t_n, y_n), \quad n = 1, 2, \ldots, \tag{3.145}$$

rekurzióval határozzuk meg, ahol ismét $y_0 = u_0$ adott.

A következő levezetés az implicit Euler-módszer már korábban is megmutatott elsőrendű lokális approximációs tulajdonságát mutatja meg (a megfelelően sima megoldásokra):

$$\begin{aligned} d_n = &\frac{u(t_n) - u(t_{n-1})}{h_n} - f(t_n, u(t_n)) = \frac{u(t_n) - u(t_{n-1})}{h_n} - u'(t_n) = \\ &\frac{u(t_n) - (u(t_n) - h_n u'(t_n) + \mathcal{O}(h_n^2))}{h_n} - u'(t_n) = \mathcal{O}(h_n). \end{aligned} \tag{3.146}$$

Emlékeztetünk, hogy az alapvető különbség az explicit Euler-módszer és az implicit Euler-módszer között az, hogy míg az első módszer esetén y_{n-1} ismeretében *közvetlenül*, egy függvénykiértékeléssel meg tudjuk határozni y_n értékét, addig az implicit Euler-módszer esetén időrétegenként egy nemlineáris egyenletet (közönséges differenciálegyenletrendszerek esetén pedig egy nemlineáris algebrai egyenletrendszert) kell megoldani.

Vizsgáljuk meg az implicit Euler-módszer abszolút stabilitási tartományát! Alkalmazva a (3.145) képletet a (3.125) tesztegyenletre a h lépésközű ekvidisztáns rácshálón, a

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

egyenlőséget kapjuk, amely átrendezve az

$$y_n = \frac{1}{1 - h\lambda} y_{n-1} \tag{3.147}$$

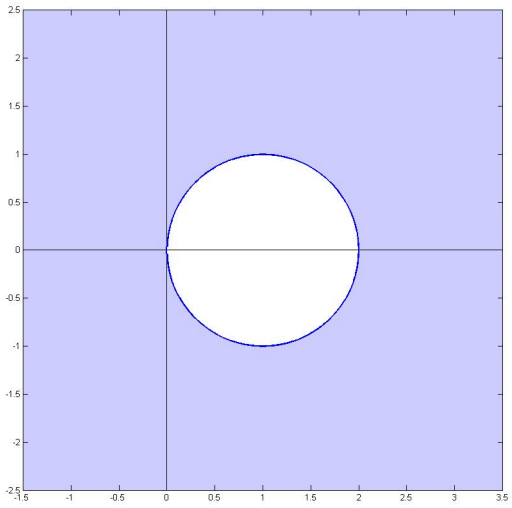
egylépéses iterációt eredményezi. A szokásos $z = \lambda h \in \mathbb{C}$ jelöléssel, az

$$R(z) = \frac{1}{1 - z} \tag{3.148}$$

stabilitási függvénnyel a módszer ismét a korábbi $y_n = R(z) y_{n-1}$ alakot ölti. Ezért az implicit Euler-módszer abszolút stabilitási tartománya az $|R(z)| \leq 1$ feltétel következtében a

$$\{z \in \mathbb{C} : \quad 1/|1 - z| \leq 1\} \tag{3.149}$$

ponthalmaz. Tehát implicit Euler-módszer stabilitási tartománya az $|1 - z| \geq 1$ tulajdonságú komplex számok halmaza, azaz a 3.2 ábra szerinti, tehát egy $\mathbb{C}^+$-beli kör komplementere.



3.2. ábra. Az implicit Euler-módszer abszolút stabilitási tartománya.

Ez azt jelenti, hogy valós $\operatorname{Re}\lambda \leq 0$ esetén nincs feltétel $h\lambda$ megválasztására, és ebben az esetben $h > 0$ tetszőlegesen választható meg. Ezért nagy $|\lambda|$ esetén sem szükséges túlságosan kis h értéket választanunk.

Mint már említettük, az implicit Euler-módszer realizálása során időrétegenként nemlineáris egyenlet megoldása szükséges. Nevezetesen, bevezetve a

$$g(y_n) = y_n - y_{n-1} - hf(t_n, y_n)$$

függvényt, minden időlépésben a $g(y_n) = 0$ egyenletet kell megoldanunk. Ennek megoldását közelíthetjük az egyszerű *függvényiteráció* segítségével, azaz az

$$y_n^{(k)} = y_{n-1} + hf(t_n, y_n^{(k-1)}), \quad k = 1, 2, \ldots, \quad y_n^{(0)} = y_{n-1} \tag{3.150}$$

iterációval.

Vizsgáljuk meg az iteráció konvergenciájának feltételét! Jelölje $e_n^{(k)} = y_n - y_n^{(k)}$ az implicit Euler-módszer numerikus megoldás és a k-ik iterált érték közötti eltérést! Ekkor a (3.145) egyenletből kivonva a (3.150) egyenletet, az

$$e_n^{(k)} = h\left(f(t_n, y_n) - f(t_n, y_n^{(k-1)})\right)$$

egyenlőséget kapjuk. Innen mindkét oldal abszolút értékét véve és felhasználva az f függvény második változó szerinti lipschitzességét,

$$|e_n^{(k)}| \leq hL|e_n^{(k-1)}| \leq \ldots \leq h^k L^k |e_n^{0)}| = h^k L^k |y_n - y_{n-1}|.$$

Ezért $k \to 0$ esetén $hL < 1$ mellett $y_n^{(k)}$ az y_n-hez tart.

Tehát az egyszerű iteráció konvergenciájának feltétele a $h < 1/L$ korlát, amely nagyságrendben megegyezik az explicit Euler-módszer stabilitási feltételével. (Sőt, annál szigorúbb is!) Ezért nagy L esetén az implicit Euler-módszer ilyen realizálása nem előnyös.

Egy másik lehetőség a *Newton-féle iteráció* alkalmazása az egyenlet megoldásának közelítésére. Ez az

$$y_n^{(k)} = y_n^{(k-1)} - \left(g'(y_n^{(k-1)})\right)^{-1} g(y_n^{(k-1)})$$

iterációt jelenti, azaz

$$y_n^{(k)} = y_n^{(k-1)} - \left(1 - h\partial_2 f(t_n, y_n^{(k-1)})\right)^{-1} \cdot \left(y_n^{(k-1)} - y_{n-1} - hf(t_n, y_n^{(k-1)})\right)$$

a módszer algoritmusa. Itt ugyanúgy felmerül a konvergencia kérdése, amely a Newton-módszer esetén a megfelelő kezdeti közelítés megválasztásának kérdéséhez vezet [5].

Egy további lehetőség a 3.1.5 szakaszban tárgyalt trapéz-módszer alkalmazása. A módszer

$$y_n = y_{n-1} + \frac{h_n}{2}\left[f(t_{n-1}, y_{n-1}) + f(t_n, y_n)\right], \quad n = 1, 2 \ldots, \tag{3.151}$$

alakú, ahol $y_0 = u_0$. Korábban megmutattuk, hogy a trapéz-módszer másodrendű módszer.

Határozzuk meg a módszer abszolút stabilitási tartományát! Alkalmazva a (3.151) képletet a (3.125) tesztegyenletre a h lépésközű ekvidisztáns rácshálón, a

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

egyenlőséget kapjuk, amely átrendezve az

$$y_n = \frac{2 + h\lambda}{2 - h\lambda} y_{n-1} \tag{3.152}$$

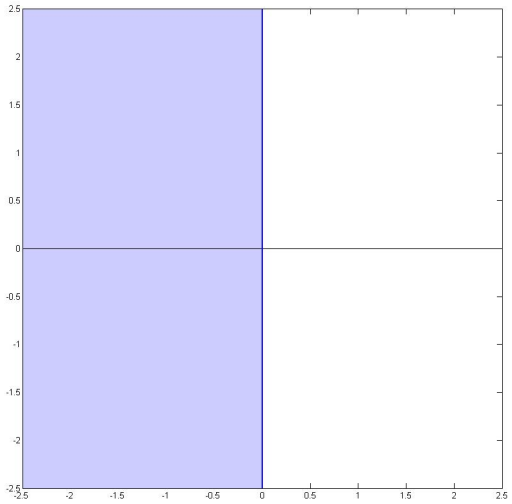
egylépéses iterációt eredményezi. A szokásos $z = \lambda h \in \mathbb{C}$ jelöléssel, az

$$R(z) = \frac{2 + z}{2 - z} \tag{3.153}$$

stabilitási függvénnyel a módszer ismét a korábbi $y_n = R(z)y_{n-1}$ alakot ölti. Ezért az trapéz-módszer abszolút stabilitási tartománya az $|R(z)| \leq 1$ feltétel következtében a

$$\{z \in \mathbb{C} : \quad \left|\frac{2 + z}{2 - z}\right| \leq 1\} \tag{3.154}$$

ponthalmaz. Ez azon z pontokban teljesül, amelyekre érvényes a $|2+z| \leq |2-z|$ egyenlőtlenség. Ez pedig pontosan akkor áll fenn, amikor $\mathrm{Re}(z) \leq 0$.[16] Tehát trapéz-módszer abszolút stabilitási tartománya a nem-pozitív valós részű komplex számok halmaza, azaz a $\mathbb{C}_0^+$ halmaz, amelya 3.3 ábrán látható.



3.3. ábra. A trapéz-módszer abszolút stabilitási tartománya.

Nyilvánvalóan a trapéz-módszer is implicit módszer, azaz realizálása során időrétegenként egy nemlineáris egyenlet megoldása szükséges. Nevezetesen, bevezetve a

$$g(y_n) = y_n - y_{n-1} - \frac{h_n}{2}\left[f(t_{n-1}, y_{n-1}) + f(t_n, y_n)\right]$$

függvényt, minden időlépésben a $g(y_n) = 0$ egyenletet kell megoldanunk.

Oldjuk meg az egyszerű függvényiteráció segítségével, azaz az

$$y_n^{(k)} = y_{n-1} + \frac{h_n}{2}\left[f(t_{n-1}, y_{n-1}) + f(t_n, y_n^{(k-1)})\right], \quad k = 1, 2, \ldots \tag{3.155}$$

iterációval, ahol $y_n^{(0)} = y_{n-1}$.

Vizsgáljuk meg az iteráció konvergenciájának feltételét! Jelölje ismét $e_n^{(k)} = y_n - y_n^{(k)}$ a trapéz-módszer numerikus megoldás és a k-ik iterált érték közötti eltérést! Ekkor

[16] Legyen ugyanis $z = x + iy$, ekkor a feltételünk $|2 + (x+iy)|^2 \leq |2 - (x+iy)|^2$, azaz $(2+x)^2 + y^2 \leq (2-x)^2 + y^2$. Ez pedig nyilvánvalóan csak $x \leq 0$ esetén igaz.

a (3.145) egyenletből kivonva a (3.150) egyenletet, az

$$e_n^{(k)} = \frac{h}{2}\left(f(t_n, y_n) - f(t_n, y_n^{(k-1)})\right)$$

egyenlőséget kapjuk. Innen mindkét oldal abszolút értékét véve és felhasználva az f függvény második változó szerinti lipschitzességét,

$$|e_n^{(k)}| \leq \frac{hL}{2}|e_n^{(k-1)}| \leq \ldots \leq \left(\frac{hL}{2}\right)^k |e_n^{0)}|$$

becslést kapjuk. Ezért $k \to 0$ esetén $hL < 2$ mellett $y_n^{(k)}$ az y_n-hez tart.

Tehát trapéz-módszer esetén az egyszerű iteráció konvergenciájának feltétele a $h < 2/L$ korlát, amely nagyságrendben megegyezik az explicit Euler-módszer stabilitási feltételével, de az implicit Euler-módszer esetére nyert becslés korlátjának kétszerese.

3.7. Megjegyzés A trapéz-módszer esetén az egyszerű iteráció során szokás az $y_n^{(0)}$ közelítésre y_{n-1} helyett a nagyobb pontosságot eredményező $y_n^{(0)} = y_{n-1} + hf(t_{n-1}, y_{n-1})$ érték megválasztás. Megjegyezzük továbbá, hogy a trapéz-módszer esetén is természetesen alkalmazható a Newton-módszer, a trapéz-módszerra definiált $g(y_n)$ függvény alkalmazásával.

3.3.3. Az A-stabilitás fogalma

Egy numerikus módszer alkalmazásánál alapvető elvárás, hogy a lehető legjobban kövesse a folytonos feladat megoldását, egyben megőrizve a folytonos feladat megoldásának legfontosabb kvalitatív tulajdonságait. Ennek bemutatására tekintsük a (3.125) tesztegyenletet! A pontos megoldás $u(t) = e^{\lambda t}u_0$, ahol $u_0 = u(0)$ és az egyszerűség kedvéért feltesszük, hogy $u_0 > 0$. (Ez nem jelent megszorítást.) Ekkor $\mathrm{Re}(\lambda) \leq 0$ esetén a megoldás (abszolút értékben) monoton csökkenő, míg $\mathrm{Re}(\lambda) > 0$ esetén növekszik, és kinő a végtelenbe. Ezért megköveteljük, hogy egy adott numerikus módszer által származtatott numerikus megoldás is örökölje ezeket a tulajdonságokat. Tehát, ha a tesztegyenletben $\mathrm{Re}(\lambda) \leq 0$, akkor a numerikus megoldásra teljesüljön az$|y_n| \leq |y_{n-1}| \leq \ldots \leq |y_0|$ egyenlőtlenség. Ez a módszerhez tartozó $R(z)$ stabilitási függvényre az $|R(z)| \leq 1$ feltételt eredményezte. Gyakorlati szempontból előnytelen, ha ez a feltétel h nagyságára valamilyen felső korlátot jelent. Ezért kiemelten fontosak azok a numerikus módszerek, amelyek azzal a tulajdonsággal rendelkeznek, hogy $\mathrm{Re}(\lambda) \leq 0$ esetén a tesztegyenlethez tartozó numerikus megoldások *tetszőleges* pozitív h esetén monoton csökkenő numerikus megoldást szolgáltatnak, azaz a $z = h\lambda$ tetszőleges h esetén benne van a módszer abszolút stabilitási tartományában.

Ez motiválja az alábbi fogalom bevezetését.

3.8. Definíció *Egy numerikus módszert A-stabil módszernek nevezünk, ha az abszolút stabilitási tartománya tartalmazza $\mathbb{C}_0^-$ bal oldali komplex félsíkot.*

A fenti definíciónak megfelelően az implicit Euler-módszer és a trapéz-módszer A-stabil, az explicit Euler-módszer viszont nem.

Fontos megjegyezni a következőt: ha egy numerikus módszer A-stabil, akkor az csak azt biztosítja, hogy a módszer a tesztegyenleten $\mathrm{Re}(\lambda) \leq 0$ esetén megfelelően viselkedik. Viszont a $\mathrm{Re}(\lambda)$ 0 esetre nem ad választ!

3.9. Példa *Vizsgáljuk meg az implicit Euler-módszer viselkedését az*

$$u' = 10^6 u, \quad u(0) = 1$$

feladaton!

Mivel a feladat pontos megoldása $u(t) = e^{10^t}$, ezért a megoldás $t \to \infty$ esetén szigorúan monoton módon kinő a végtelenbe. Nézzük meg, hogy az implicit Euler-módszer milyen feltételek esetén örökli át ezt a tulajdonságot? Az implicit Euler-módszer alkalmazása esetén a numerikus megoldást az

$$y_n = \frac{1}{1 - 10^6 h} y_{n-1} \equiv R(10^6 h) y_{n-1}$$

iterációval állítjuk elő. Az $|y_n| > |y_{n-1}|$ reláció teljesülésének feltétele: $h \leq 2 \cdot 10^{-6}$. Tehát a numerikus megoldás abszolút értékben csak ebben az esetben növekszik. Ez azt jelenti, hogy az ennél nagyobb h nem alkalmas a numerikus módszerre.

Vegyük észre, hogy tetszőleges $\lambda \in \mathbb{R}$ esetén a pontos megoldás szigorúan monoton és pozitív. Ugyanakkor $h > 10^{-6}$ esetén $R(h) < 0$, ezért időlépésenként y_n előjelet vált. Ez azt jelenti, hogy a numerikus megoldás hamisan oszcillál.

Foglaljuk össze az implicit Euler-módszer viselkedését a különböző rácshálókon a fenti feladatra!

- $h \in (0, 10^{-6})$ esetén szigorúan monoton nő (tehát pozitív) és nem oszcillál,
- $h \in (10^{-6}, 2 \cdot 10^{-6})$ esetén abszolút értékben szigorúan monoton nő és oszcillál,
- $h \in (2 \cdot 10^{-6}, \infty)$ esetén abszolút értékben szigorúan monoton csökken és oszcillál.

Ugyanakkor hangsúlyozzuk, hogy ez a probléma (tehát, hogy csak a $h < 10^{-6}$ lépésközű rácshálókon viselkedik jól az implicit Euler-módszer) csak a tesztfeladatban lévő pozitív valós részű λ esetén fordul elő. Ha λ valós része nem-pozitív, akkor bármely $h > 0$ esetén megfelelő a numerikus megoldás.

Az előző problémában már érintettük az numerikus megoldás oszcillációját, amely a tesztfeladaton nemkívánatos jelenség. (Hiszen a pontos megoldás a $\lambda \in \mathbb{R}$ értékekre

monoton, azaz nem vált előjelet az egyes időrétegek között.) Ezt a tulajdonságot egy adott numerikus módszerre pontosan akkor tudjuk biztosítani, amikor a numerikus módszerhez tartozó $R(z)$ stabilitási függvényre az $R(\lambda h) > 0$ feltétel teljesül. A vizsgált módszereinkre ez a következő feltételeket jelenti.

1. Az explicit Euler-módszerre $\lambda \geq 0$ esetén minden $h > 0$ esetén oszcilláció-mentes a megoldás, és $\lambda < 0$ esetén $h < 1/(-\lambda)$ a feltétel.

2. Az implicit Euler-módszerre $\lambda \leq 0$ esetén minden $h > 0$ esetén oszcilláció-mentes a megoldás, és $\lambda > 0$ esetén pedig $h < 1/\lambda$ a feltétel.

3. Az trapéz-módszerra $\lambda = 0$ esetén nincs feltétel, $\lambda < 0$ esetén $h < 2/(-\lambda)$, $\lambda > 0$ esetén pedig $h < 2/\lambda$ a feltétel.

Tehát az A-stabil módszerek közül a trapéz-módszernak $\lambda \leq 0$ esetén is van hátránya: ha $|\lambda h|$ nagy, akkor a numerikus megoldás oszcillál. (Ráadásul, ekkor $|R(\lambda h)|$ közel van az egyhez, és ezért a numerikus megoldás nem fog a kellő sebességgel lecsengeni.) Az implicit Euler-módszer alkalmazása pedig pozitív nagy értékű λ esetén nem ajánlott. Ezek az észrevételek az

$$\mathbf{u}' = \mathbf{A}\mathbf{u} + \mathbf{f}$$

lineáris rendszerek megoldására is vonatkoznak, amikor az $\mathbf{A}$ mátrix sajátértékei a fenti tulajdonságúak.

Ezért ha az $\mathbf{A}$ mátrix mindegyik sajátértékének valós része nem-pozitív, akkor az A-stabil módszerek közül a trapéz-módszer kevésbé ajánlott, amikor a mátrix spektrálsugarára $\rho(\mathbf{A}) \gg 1$. Ugyanakkor ha a spektrum (mindegyik sajátérték) a képzetes tengelyen van, akkor a trapéz-módszer különösen jó. Ugyanis, ebben az esetben a sajátértékek $\lambda_k = i\beta_k$ alakúak, ahol i a szokásos képzetes egység, ezért a tesztegyenletek $u' = i\beta_k u$ alakúak. Ekkor a megoldások $u(t) = e^{i\beta_k t}u(0)$ alakúak, azaz $|u(t)| = |u(0)|$ tetszőleges $t > 0$ időpontban, azaz a megoldás abszolút értékben nem változik. Ennek a tulajdonságnak az numerikus módszer $|R(z)| = 1$ tulajdonsága felel meg. Mivel a trapéz-módszer esetén

$$|R(i\beta_k)| = \left|\frac{2 + ih\beta_k}{2 - ih\beta_k}\right| = 1$$

(ugyanis a számláló és nevező egymás komplex konjugáltja), ezért ez a módszer megfelelő. Emellett, mivel az eredeti folytonos megoldás oszcillál, ezért a trapéz-módszer oszcillációs tulajdonsága is jól követi az eredeti megoldás ezen tulajdonságát.

3.10. Megjegyzés Ha $\mathbf{A}$ egy úgynevezett ferdén szimmetrikus, valós mátrix, azaz $\mathbf{A} = -\mathbf{A}^T$, akkor a spektruma az $\mathrm{Im}(\mathbb{C})$ tengelyen van. Ez közvetlenül adódik a $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ sajátérték tulajdonság melletti $\lambda(\mathbf{v}, \mathbf{v}) = (\mathbf{A}\mathbf{v}, \mathbf{v}) = (\mathbf{v}, \mathbf{A}^T\mathbf{v}) = -(\mathbf{v}, \mathbf{A}\mathbf{v}) = -(\mathbf{v}, \lambda\mathbf{v}) = -\overline{\lambda}(\mathbf{v}, \mathbf{v})$ relációból következő $\lambda = -\overline{\lambda}$ egyenlőségből.

Mutassuk meg, hogy ha $\mathbf{A}$ egy ferdén szimmetrikus, valós mátrix, akkor tetszőleges $s \in \mathbb{R}$ esetén az euklideszi normára érvényes az

$$\|(\mathbf{I} + s\mathbf{A}) \cdot (\mathbf{I} - s\mathbf{A})^{-1}\|_2 = 1$$

egyenlőség! Felhasználjuk, hogy ha $\mathbf{T}$ egy ortogonális mátrix, akkor az euklideszi normában normatartó leképezés, azaz $(\mathbf{Tx}, \mathbf{Ty}) = (\mathbf{x}, \mathbf{y})$ minden $\mathbf{x}, \mathbf{y}$ vektor esetén. Ezért a λ sajátértékeire, a $\mathbf{Tv} = \lambda\mathbf{v}$ reláció alapján érvényes a $(\mathbf{v}, \mathbf{v}) = (\mathbf{Tv}, \mathbf{Tv}) = (\lambda\mathbf{v}, \lambda\mathbf{v}) = \lambda\overline{\lambda}(\mathbf{v}, \mathbf{v}) = |\lambda|^2(\mathbf{v}, \mathbf{v})$ összefüggés, azaz sajátértékeinek abszolút értéke egy.

Jelölje $\mathbf{B} = (\mathbf{I}+s\mathbf{A})(\mathbf{I}-s\mathbf{A})^{-1}$ mátrixot. Elegendő megmutatnunk, hogy ez a mátrix ortogonális, hiszen ebből az állításunk már következik. Felhasználva, hogy az $\mathbf{A}$ mátrix ferdén szimmetrikus, érvényes az alábbi egyenlőség: $\mathbf{B}^{-1} = (\mathbf{I} - s\mathbf{A})(\mathbf{I} + s\mathbf{A})^{-1} = (\mathbf{I} + s\mathbf{A}^T)(\mathbf{I} - s\mathbf{A}^T)^{-1}$ Mivel $(\mathbf{I}+s\mathbf{A}^T) = (\mathbf{I}+s\mathbf{A})^T$, ezért tehát $\mathbf{B}^{-1} = (\mathbf{I}+s\mathbf{A})^T(\mathbf{I}-s\mathbf{A})^{-T} = [(\mathbf{I} - s\mathbf{A})^{-1})(\mathbf{I} + s\mathbf{A})]^T$. A zárójelben lévő szorzat két mátrixa egymással kommutál, ezért $\mathbf{B}^{-1} = [(\mathbf{I} + s\mathbf{A})(\mathbf{I} - s\mathbf{A})^{-1})]^T = \mathbf{B}^T$. Így $\mathbf{B}$ valóban ortogonális mátrix, és ezzel az állításunkat beláttuk.

Következésképpen, ferdén szimmetrikus mátrixok esetén a trapéz-módszer az euklideszi normában normatartó.

4. fejezet

Runge-Kutta típusú módszerek kezdetiérték-feladatok megoldására

Az előző fejezetben tárgyalt egylépéses módszerek algoritmikusan viszonylag egyszerűek, ugyanakkor csak alacsony rendben pontosak. Ha nagy pontosságú közelítést szeretnénk elérni, akkor ezek a módszerek általában nem megfelelőek. A túlságosan kis lépésköz megválasztása ugyanis a számítások hatékonysága miatt nem mindig célravezető: egy adott időréteg eléréséhez túl sok lépést kell megtennünk, amely egyrészt lényegesen megnövelheti a szükséges számítási igényt, másrészt az egyes lépések során keletkező numerikus hibák jelentősen felhalmozódhatnak.

Ezért célszerű magasabb rendű módszerek bevezetése. Ezt kétféleképpen is elérhetjük.

- Más felépítésű egylépéses módszerekkel, amelyek két rácspont közötti újabb értékek kiszámításával növelik a rendet.
- Az egylépéses módszerek helyett többlépéses módszerek alkalmazása, amelyek az új időpontbeli közelítés meghatározásánál nem csak a közvetlenül megelőző rácshálóbeli közelítést, hanem több, korábbi időpontbeli közelítést is figyelembe vesz.

Ebben a fejezetben az első megoldással foglalkozunk, amelyeket összefoglalóan Runge-Kutta típusú módszereknek nevezünk.

4.1. A Runge-Kutta típusú módszerek alapjai

Kettőnél magasabb rendű numerikus módszerek megkonstruálása a (3.1)-(3.2) Cauchy-feladatra az előzőekben ismertetett egylépéses módszerek segítségével akadályokba ütközik: az egyszerűbb módszerek (explicit Euler-módszer, implicit Euler-módszer, trapézmódszer) legfeljebb másodrendűek, a Taylor-módszerek viszont egy meglehetősen bonyo-

lult előzetes analízist (a parciális deriváltak meghatározását) és azok kiértékelését igénylik. Ebben a részben megmutatjuk, hogy a parciális deriváltak kiszámításának feladata – egy viszonylag egyszerű ötlet segítségével – megkerülhető.

4.1.1. További másodrendű módszerek

Tekintsük ismét a (3.1)-(3.2) Cauchy-feladatot. Először a Runge-Kutta típusú módszerek bevezetéséhez határozzunk meg egy, a trapéz-módszertől különböző, másodrendű, egylépéses módszert.
Írjuk ki az $u(t)$ (3.6) alakú Taylor-sorának első tagjait a $t = t^\star + h$ pontban. Ekkor

$$u(t^\star + h) = u(t^\star) + hu'(t^\star) + \frac{h^2}{2!}u''(t^\star) + \mathcal{O}(h^3). \tag{4.1}$$

Felhasználva a (3.4) deriváltakat, bevezetve az

$$f = f(t^\star, u(t^\star)), \quad \partial_i f = \partial_i f(t^\star, u(t^\star)), \quad \partial_{ij} f = \partial_{ij} f(t^\star, u(t^\star)), \quad \text{stb.}$$

egyszerűsítő jelöléseket, (4.1) átírható az

$$\begin{aligned} u(t^\star + h) =& u(t^\star) + hf + \frac{h^2}{2!}(\partial_1 f + f\partial_2 f) + \mathcal{O}(h^3) \\ =& u(t^\star) + \frac{h}{2}f + \frac{h}{2}[f + h\partial_1 f + hf\partial_2 f] + \mathcal{O}(h^3) \end{aligned} \tag{4.2}$$

alakra. Mivel[1]

$$f(t^\star + h, u(t^\star) + hf(t^\star, u(t^\star)) = f + h\partial_1 f + hf\partial_2 f + \mathcal{O}(h^2), \tag{4.3}$$

ezért (4.2) felírható az

$$u(t^\star + h) = u(t^\star) + \frac{h}{2}f + \frac{h}{2}(f(t^\star + h, u(t^\star) + hf(t^\star, u(t^\star))) + \mathcal{O}(h^3) \tag{4.4}$$

alakban. Tehát egy ω_h rácsháló tetszőleges $t_{n-1} = t^\star$ pontjában felírva a (4.4) egyenlőséget, definiálhatjuk az

$$y_n = y_{n-1} + \frac{h}{2}f(t_{n-1}, y_{n-1}) + \frac{h}{2}f(t_n, y_{n-1} + hf(t_{n-1}, y_{n-1})) \tag{4.5}$$

egylépéses, explicit numerikus módszert. Vezessük be a

$$\begin{aligned} k_1 =& f(t_{n-1}, y_{n-1}), \\ k_2 =& f(t_n, y_{n-1} + hf(t_{n-1}, y_{n-1})) = f(t_{n-1} + h, y_{n-1} + hk_1) \end{aligned} \tag{4.6}$$

[1]Emlékeztetünk, hogy a kétváltozós $f : Q_T \to \mathbb{R}$ függvény (t, u) pont körüli elsőfokú Taylor-sorba fejtése tetszőleges $c_1, c_2 \in \mathbb{R}$ esetén $f(t + c_1 h, u + c_2 h) = f(t, u) + c_1 h\partial_1 f(t, u) + c_2 h\partial_2 f(t, u) + \mathcal{O}(h^2)$ alakú.

jelöléseket. Ekkor a (4.5) módszer felírható

$$y_n = y_{n-1} + \frac{h}{2}(k_1 + k_2) \tag{4.7}$$

alakban. A (4.6)-(4.7) módszert *Heun-módszernek* nevezzük.

4.1. Megjegyzés Mivel (4.4) alapján

$$u(t^\star + h) - u(t^\star) - \frac{h}{2}f - \frac{h}{2}(f(t^\star + h, u(t^\star) + hf(t^\star, u(t^\star))) = \mathcal{O}(h^3), \tag{4.8}$$

ezért a megoldás $\mathcal{O}(h^3)$ rendben elégíti ki a (4.5) képletet, azaz *a Heun-módszer másodrendben konzisztens.*

A Heun-módszer néhány részlete megtalálható a
http://math.fullerton.edu/mathews/n2003/Heun'sMethodMod.html
linken. Az ugyanitt lévő animáción a módszer viselkedése az $u' = 1 - t\sqrt[3]{u}$ differenciálegyenletre is jól látható.
Megadhatók-e egyéb másodrendű módszerek? A (4.4) összefüggés általánosítása (megfelelően sima megoldások esetén) a következő paraméteres alak:

$$\begin{aligned} u(t^\star + h) =& u(t^\star) + b_1 hf(t^\star, u(t^\star)) + \\ & b_2 hf(t^\star + c_2 h, u(t^\star) + a_{21} hf(t^\star, u(t^\star))) + \mathcal{O}(h^3), \end{aligned} \tag{4.9}$$

ahol b_1, b_2, c_2 és a_{21} egyelőre tetszőleges paraméterek. Felírva a $t = t_{n-1}$ pontban a (4.9) egyenletet, az

$$\begin{aligned} y_n =& y_{n-1} + b_1 hf(t_{n-1}, y_{n-1}) + \\ & b_2 hf(t_{n-1} + c_2 h, y_{n-1} + a_{21} hf(t_{n-1}, y_{n-1})) \end{aligned} \tag{4.10}$$

egylépéses numerikus módszert kapjuk.

4.2. Megjegyzés A (4.12) általános alakban felírt módszer paramétereit célszerű csoportosítva a következő alakban felírni:

$$\begin{array}{c|cc} 0 & & \\ c_2 & a_{21} & \\ \hline & b_1 & b_2 \end{array} \tag{4.11}$$

Fejtsük Taylor-sorba a (4.9) jobb oldalát! Ekkor az

$$\begin{aligned} & u(t^\star + h) = \\ & u(t^\star) + b_1 hf + b_2 h[f + c_2 h\partial_1 f + a_{21} hf\partial_2 f] + \mathcal{O}(h^3) = \\ & u(t^\star) + (b_1 + b_2)hf + h^2[c_2 b_2 \partial_1 f + b_2 a_{21} f\partial_2 f] + \mathcal{O}(h^3) \end{aligned} \tag{4.12}$$

egyenlőséget kapjuk. A módszerek rendjére vonatkozó 4.1. megjegyzést alkalmazva, a (4.2) és a (4.12) képletek összevetéséből azt kapjuk, hogy a (4.10) által meghatározott numerikus módszer pontosan akkor másodrendű, amikor

$$\begin{aligned} b_1 + b_2 &= 1 \\ c_2 b_2 &= 0.5 \\ a_{21} b_2 &= 0.5. \end{aligned} \tag{4.13}$$

A (4.10) képlet egyszerű átírásával eredményeinket az alábbi tételben összegezhetjük.

4.3. Tétel *Tegyük fel, hogy a b_1, b_2, c_2 és a_{21} paraméterek megoldásai a (4.13) egyenletnek. Ekkor a*

$$k_1 = f(t_{n-1}, y_{n-1}), \quad k_2 = f(t_{n-1} + c_2 h, y_{n-1} + h a_{21} k_1), \tag{4.14}$$

$$y_n = y_{n-1} + h(b_1 k_1 + b_2 k_2) \tag{4.15}$$

képletekkel definiált egylépéses explicit numerikus módszer másodrendű.

A (4.13) feltételeket kielégítő (4.14)-(4.15) módszert *másodrendű Runge-Kutta típusú módszernek* nevezzük és RK2 szimbólummal jelöljük.
Vizsgáljuk meg az RK2 módszert meghatározó (4.13) egyenletrendszert! Mivel a négy ismeretlenre három egyenletünk van, ezért a megoldása nem egyértelmű. Könnyen látható, hogy tetszőleges $\sigma \neq 0$ esetén (4.13) megoldásai a következő alakúak:

$$b_2 = \sigma, \quad b_1 = 1 - \sigma, \quad c_2 = a_{21} = 1/2\sigma. \tag{4.16}$$

Tehát az RK2 módszerek egy egyparaméteres módszercsaládot alkotnak, amelynek paramétereit a (4.11) táblázat alapján a

$$\begin{array}{c|cc} 0 & & \\ 1/2\sigma & 1/2\sigma & \\ \hline & 1-\sigma & \sigma \end{array} \tag{4.17}$$

szerint kell megválasztani.

4.4. Megjegyzés A $\sigma = 0.5$ értékhez tartozó RK2 módszer éppen a Heun-módszert eredményezi. Érdekes megválasztás a $\sigma = 1$. Ekkor $\sigma_1 = 0$, $\sigma_2 = 1$ és $a_2 = b_{21} = 0.5$ és így a származtatott numerikus módszer

$$\begin{aligned} k_1 &= f(t_{n-1}, y_{n-1}), \quad k_2 = f(t_{n-1} + 0.5h, y_{n-1} + 0.5h k_1), \\ y_n &= y_{n-1} + h k_2. \end{aligned} \tag{4.18}$$

A (4.18) másodrendű módszert *javított explicit Euler-módszernek* nevezzük.

Vegyük észre, hogy a Heun-módszer és a javított explicit Euler-módszer is bevezethető a korábbiakban már ismertetett módszerek módosításával. Nevezetesen,

- Ha a trapéz-módszer (3.51)

$$y_n - y_{n-1} = \frac{h_n}{2}\left[f(t_{n-1}, y_{n-1}) + f(t_n, y_n)\right]$$

képletében az $f(t_n, y_n)$ implicit tagban y_n helyett az $\tilde{y}_n = y_{n-1} + hf(t_{n-1}, y_{n-1})$ explicit Euler-módszerrel kiszámított értéket helyezzük, akkor éppen a Heun-módszert kapjuk.

- A javított explicit Euler-módszer esetén a $t = t_{n-1} + 0.5h = t_{n-0.5}$ felezőpontban explicit Euler-módszerrel kiszámoljuk az $u(t)$ függvény közelítő értékét az $\tilde{y}_{n-0.5} = y_n + 0.5hf(t_{n-1}, y_{n-1})$ képlettel, és ezzel az értékkel meghatározott irányban egy újabb explicit Euler-módszert írunk fel a teljes intervallumra.

Tehát a fenti Runge-Kutta módszerek paramétereinek (4.17) szerinti megadása a következő. A Heun-módszer esetén

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & 0.5 & 0.5 \end{array} \tag{4.19}$$

alakban, míg a javított explicit Euler-módszer esetén

$$\begin{array}{c|cc} 0 & & \\ 0.5 & 0.5 & \\ \hline & 0 & 1 \end{array} \tag{4.20}$$

alakban adható meg.

Az elsőrendű explicit Euler-módszer és a másodrendű javított explicit Euler-módszer konvergenciasebességei egymáshoz viszonyítva az alábbi animáción láthatóak:
../animaciok/ee1jee2.gif

4.5. Megjegyzés Felmerülhet a kérdés: lehetséges-e a σ tetszőleges paramétert úgy megválasztani, hogy az RK2 módszer nemcsak másodrendű, hanem harmadrendű legyen? A válasz nemleges, amit a következő példa bizonyít. Legyen a (3.1)-(3.2) Cauchy-feladatban $f(t, u) = u$. Ezért a (3.1) differenciálegyenlet megoldására $u'(t) = u(t)$, amelyet deriválva $u''(t) = u'(t)$. Ezért $u''(t) = u'(t) = u(t)$. Másrészt, az f függvény definíciója miatt $f(t_{n-1}, y_{n-1}) = y_{n-1}$. Ezért a (4.10) módszer erre a feladatra az

$$\begin{aligned} y_n &= y_{n-1} + b_1 h y_n + b_2 h(y_{n-1} + a_{21} h f(t_{n-1}, y_{n-1})) \\ &= y_{n-1} + b_1 h y_n + b_2 h(y_{n-1} + a_{21} h y_{n-1}) \\ &= y_{n-1} + h y_{n-1}[b_1 + b_2 + h b_2 a_{21}] \\ &= y_{n-1}[1 + (b_1 + b_2)h + b_2 a_{21} h^2] \end{aligned} \tag{4.21}$$

alakot ölti. Behelyettesítve a másodrendűséghez szükséges (4.16) értékeket, az RK2 módszer erre a feladatra az

$$y_n = y_{n-1}\left(1 + h + \frac{h^2}{2}\right). \tag{4.22}$$

algoritmust eredményezi, amely független a σ szabad paramétertől. Helyettesítsük be az $u(t)$ pontos megoldást a (4.22) képletbe, azaz számítsuk ki a pontos megoldásból induló, egy lépésben keletkező hibát!(V.ö. a 3.17. megjegyzéssel.) Ekkor

$$d_n = u(t_n) - u(t_{n-1})(1 + h + \frac{h^2}{2}). \tag{4.23}$$

Az $u(t_n)$ kifejezés $t = t_{n-1}$ pontbeli sorbafejtése a deriváltakra vonatkozó $u''(t_{n-1}) = u'(t_{n-1}) = u(t_{n-1})$ egyenlőség következtében $u(t_n) = u(t_{n-1})(1 + h + h^2/2) + \mathcal{O}(h^3)$. Ezért tehát $g_n = \mathcal{O}(h^3)$ a σ tetszőleges megválasztása mellett, azaz minden RK2 módszer legfeljebb másodrendű erre a feladatra.

4.2. Explicit, magasabb rendű módszerek

A korábbiakban már bevezettük és részletesen megvizsgáltuk az explicit Euler-módszert, implicit Euler-módszert és a trapéz-módszert. Ezeket a módszereket a megoldásfüggvény Taylor-sorba fejtésével származtattuk. Ugyanakkor a módszerek más módon is bevezethetők, és ennek általánosítása ad majd lehetőséget a magasabb rendű módszerek megkonstruálására.

Legyen $u(t)$ a

$$\frac{du}{dt} = f(t, u), \quad u(0) = u_0 \tag{4.24}$$

feladat megoldása, azaz érvényes rá az

$$u'(t) = f(t, u(t)), \quad u(0) = u_0 \tag{4.25}$$

azonosság a $[0, T]$ intervallum mindegyik pontjában. Az azonosság mindkét oldalát integrálva az ω_h rácsháló két tetszőleges szomszédos pontja között az

$$u(t_n) - u(t_{n-1}) = \int_{t_{n-1}}^{t_n} f(t, u(t))dt, \quad t \in [t_{n-1}, t_n] \tag{4.26}$$

($n = 1, \ldots, N$ tetszőleges) egyenlőséget nyerjük. A közelítő módszerek megkonstruálásához a jobb oldalon lévő integrált valamely közelítő formulával számoljuk ki a $[t_{n-1}, t_n]$ intervallumon. A legegyszerűbb numerikus integrálási formula, amikor az intervallum valamely végpontjában, avagy mindkettőben felvett értéke szerepel csak a numerikus integráló formulákban. A különböző numerikus integráló formulák elvezetnek a fenti módszerekhez. Nevezetesen,

- A legegyszerűbb módszer, amikor a téglalapszabályt alkalmazzuk a bal oldali végpontjában felvett értékének felhasználásával, azaz

$$\int_{t_{n-1}}^{t_n} f(t,u(t))dt \approx h_n f(t_{n-1},u(t_n)). \tag{4.27}$$

  Ekkor a (4.26) és a (4.27) összefüggések a (3.26) képlettel megadott explicit Euler-módszert eredményezik.

- Egy további lehetséges módszer a (4.26) azonosságban szereplő integrál közelítő meghatározására, hogy az intervallum jobb oldali végpontbeli függvényértéket felhasználva a téglalapszabályt alkalmazzuk, azaz

$$\int_{t_{n-1}}^{t_n} f(t,u(t))dt \approx h_n f(t_n,u(t_n)). \tag{4.28}$$

  Ekkor (4.26) és (4.28) felhasználásával a (3.145) alakú implicit Euler-módszert kapjuk.

- Ha a (4.26) közelítő integrálására a trapézszabályt alkalmazzuk, azaz

$$\int_{t_{n-1}}^{t_n} f(t,u(t))dt \approx \frac{h_n}{2}\left[f(t_{n-1},u(t_{n-1}))+f(t_n,u(t_n))\right], \tag{4.29}$$

  akkor a (4.26) és a (4.29) képletek felhasználásával a (3.51) alakú trapéz-módszert kapjuk. (Emiatt szokás a módszert *trapéz módszernek* nevezni.)

- Ha a (4.26) közelítő integrálására a középponti szabályt alkalmazzuk, azaz

$$\int_{t_{n-1}}^{t_n} f(t,u(t))dt \approx h_n f(t_{n-0.5},u(t_{n-0.5})), \tag{4.30}$$

  akkor az

$$y_n - y_{n-1} = h_n f(t_{n-0.5},y_{n-0.5}) \tag{4.31}$$

  formulát kapjuk, ahol $y_{n-0.5}$ az $u(t_{n-0.5})$ érték közelítése.

- A (4.26) jobb oldalának numerikus integrálásra alkalmazhatjuk a Simpson-féle szabályt is. Ekkor ezt az integrált három pontbeli értékkel közelítjük, nevezetesen a

$$\frac{h_n}{6}\left[f(t_{n-1},u(t_{n-1}))+4f(t_{n-0.5},u(t_{n-0.5}))+f(t_n,u(t_n))\right]$$

  formulával. Ezzel a

$$y_n - y_{n-1} = \frac{h_n}{6}\left[f(t_{n-1},y_{n-1})+4f(t_{n-0.5},y_{n-0.5})+f(t_n,y_n)\right], \tag{4.32}$$

  numerikus integráló formulát kapjuk.

4.1. Megjegyzés Egy másik lehetséges mód a fenti módszerek származtatására, amikor a (4.25) azonosságot valamely rögzített rácspontban felírva, a bal oldalon lévő deriváltra egy numerikus deriválási formulát alkalmazunk.

- Ha a $t = t_{n-1}$ pontban írjuk fel a (4.25) azonosságot, akkor az $u'(t_{n-1}) = f(t_{n-1}, u(t_{n-1}))$ egyenlőséget kapjuk. A bal oldalon szereplő deriváltat a haladó véges differenciával közelítve az $u'(t_{n-1}) \simeq (u(t_n) - u(t_{n-1}))/h_n$ közelítést kapjuk. Ez a két formula pedig az explicit Euler-módszert generálja.
- Ha a $t = t_n$ pontban írjuk fel a (4.25) azonosságot, és a retrográd numerikus deriválást alkalmazzuk, akkor az implicit Euler-módszert kapjuk.
- A fenti két képlet számtani közepe szolgáltatja a trapéz-módszer képletét.

A középponti szabályon illetve a Simpson-képleten alapuló (4.31) és (4.32) módszerek tartalmazzák az ismeretlen $y_{n-0.5}$ értékeket is. Ezeket az $y_{n-0.5} \approx 0.5(y_{n-1} + y_n)$ közelítés alkalmazásával egy újabb sémához jutunk. Így a (4.31) helyett tekinthetjük a

$$y_n - y_{n-1} = h_n f(t_{n-0.5}, \frac{y_{n-1} + y_n}{2}) \tag{4.33}$$

sémát, amelyet *középponti módszernek* szokásos nevezni.

Az implicit Euler-módszer, trapéz-módszer, továbbá a középponti illetve a Simpson-képleten alapuló numerikus módszerek implicit sémát jelentenek, minden időrétegen nemlineáris egyenletek (rendszer esetén: egyenletrendszerek) megoldását igénylik. Ennek elkerülésére szolgál a következő ötlet. Az implicit módszerek esetén első lépésben a jobb oldalon szereplő y_n értékét közelítsük egy $\tilde{y}_n$ értékkel valamely explicit módszer segítségével, majd második lépésben az $\tilde{y}_n$ közelítést behelyettesítve az eredeti formulában, már explicit képletet nyerve meghatározzuk az $t = t_n$ pontbeli y_n közelítést.

4.2. Példa *Írjuk fel ekvidisztáns rácshálón az implicit Euler-módszer explicitté tett változatát, és vizsgáljuk meg az így nyert explicit módszer rendjét!*

Mivel az implicit Euler-módszer

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

ezért legyen

$$\tilde{y}_n = y_{n-1} + hf(t_{n-1}, y_{n-1}). \tag{4.34}$$

Ezután az

$$y_n = y_{n-1} + hf(t_n, \tilde{y}_n) \tag{4.35}$$

képlettel állítjuk elő a megoldást. Ezen módszer lokális approximációs hibáját a következő módon lehet meghatározni. A (4.34) és a (4.35) módszer felírható a

$$y_n = y_{n-1} + hf(t_{n-1}, y_{n-1} + hf(t_{n-1}, y_{n-1})) \tag{4.36}$$

egylépéses explicit módszer alakjában, amelynek kanonikus alakja

$$\frac{y_n - y_{n-1}}{h} = f(t_n, y_{n-1} + hf(t_{n-1}, y_{n-1})). \tag{4.37}$$

Ezért a lokális approximációs hibára érvényes a következő.

$$d_n = \frac{u(t_n) - u(t_{n-1})}{h} - f(t_n, u(t_{n-1}) + hf(t_{n-1}, u(t_{n-1}))). \tag{4.38}$$

Mivel

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

továbbá az $f_{n-1} = f(t_{n-1}, u(t_{n-1}))$ jelöléssel

$$\begin{aligned}
&f(t_n, u(t_{n-1}) + hf(t_{n-1}, u(t_{n-1}))) = \\
&f(t_{n-1} + h, u(t_{n-1}) + hf_{n-1}) = f_{n-1} + h\partial_1 f_{n-1} + h(\partial_2 f_{n-1})f_{n-1} + \mathcal{O}(h^2) = \\
&f_{n-1} + h\left(\partial_1 f_{n-1} + (\partial_2 f_{n-1})f_{n-1}\right) + \mathcal{O}(h^2)
\end{aligned}$$

Mivel $u'(t_{n-1}) = f_{n-1}$ és $u''(t_{n-1}) = \partial_1 f_{n-1} + (\partial_2 f_{n-1})f_{n-1}$, ezért a d_n lokális approximációs hiba

$$d_n = -\frac{h}{2}u''(t_{n-1}) + \mathcal{O}(h^2) = \mathcal{O}(h),$$

azaz ez a módszer elsőrendű.

Tekintsünk most egy újabb explicitté tett implicit módszert!

4.3. Példa *Írjuk fel a (4.31) középponti módszer explicit változatát, és határozzuk meg a módszer rendjét!*

Mivel a módszer alakja

$$y_n - y_{n-1} = hf(t_{n-0.5}, y_{n-0.5}), \tag{4.39}$$

ezért explicit alakja a következő.

$$\tilde{y}_{n-0.5} = y_{n-1} + 0.5hf(t_{n-1}, y_{n-1}), \tag{4.40}$$

$$y_n = y_{n-1} + hf(t_{n-0.5}, \tilde{y}_{n-0.5}). \tag{4.41}$$

A (4.40)-(4.41) explicit módszer közvetlenül átírható az alábbi egylépéses módszerként:

$$y_n = y_{n-1} + hf(t_{n-0.5}, y_{n-1} + 0.5hf(t_{n-1}, y_{n-1})). \tag{4.42}$$

Ezért a lokális approximációs hibára érvényes a következő.

$$d_n = \frac{u(t_n) - u(t_{n-1})}{h} - f(t_{n-0.5}, u(t_{n-1}) + 0.5hf(t_{n-1}, u(t_{n-1}))). \tag{4.43}$$

Mivel

$$\begin{aligned} &f(t_{n-0.5}, u(t_{n-1}) + 0.5hf(t_{n-1}, u(t_{n-1}))) = f_{n-1} + 0.5h\partial_1 f_{n-1} + \\ &0.5h(\partial_2 f_{n-1})f_{n-1} + \mathcal{O}(h^2) = f_{n-1} + 0.5h(\partial_1 f_{n-1} + f_{n-1}\partial_2 f_{n-1}) \\ &+ \mathcal{O}(h^2) = u'(t_{n-1}) + 0.5u''(t_{n-1}) + \mathcal{O}(h^2), \end{aligned} \tag{4.44}$$

ezért a (4.40)-(4.41) explicit módszer másodrendű.

4.4. Példa *Írjuk fel a (3.51) trapéz módszer explicit változatát, és határozzuk meg a módszer rendjét!*

Mivel a módszer alakja

$$y_n - y_{n-1} = 0.5h\left[f(t_{n-1}, y_{n-1}) + f(t_n, y_n)\right], \tag{4.45}$$

ezért explicit alakja a következő.

$$\tilde{y}_n = y_{n-1} + hf(t_{n-1}, y_{n-1}), \tag{4.46}$$

$$y_n = y_{n-1} + 0.5h\left[f(t_{n-1}, y_{n-1}) + f(t_n, \tilde{y}_n)\right]. \tag{4.47}$$

Vagyis a (4.46)-(4.47) explicit módszer közvetlenül átírható az alábbi egylépéses módszerként:

$$y_n = y_{n-1} + 0.5h\left[f(t_{n-1}, y_{n-1}) + f(t_n, y_{n-1} + hf(t_{n-1}, y_{n-1}))\right]. \tag{4.48}$$

Ezért a lokális approximációs hiba a következő.

$$\begin{aligned} d_n = &\frac{u(t_n) - u(t_{n-1})}{h} - \\ &0.5\left[f(t_{n-1}, u(t_{n-1})) + f(t_n, u(t_{n-1}) + hf(t_{n-1}, u(_{n-1})))\right] \end{aligned} \tag{4.49}$$

Mivel

$$\begin{aligned} &0.5\left[f(t_{n-1}, u(t_{n-1})) + f(t_n, u(t_{n-1}) + hf(t_{n-1}, u(_{n-1})))\right] = \\ &0.5\left[f_{n-1} + (f_{n-1} + h\partial_1 f_{n-1} + h(\partial_2 f_{n-1})f_{n-1}\right] + \mathcal{O}(h^2) = \\ &f_{n-1} + 0.5h(\partial_1 f_{n-1} + f_{n-1}\partial_2 f_{n-1}) \\ &+ \mathcal{O}(h^2) = u'(t_{n-1}) + 0.5u''(t_{n-1}) + \mathcal{O}(h^2), \end{aligned} \tag{4.50}$$

ezért a (4.46)-(4.47) explicit módszer másodrendű.

A Simpson-képleten alapuló (4.32) módszer explicitté tétele már jóval bonyolultabb. Részletezés nélkül megadjuk a szokásos explicit módszert:

$$\begin{aligned} \tilde{y}_n^1 &= y_{n-1}, \\ \tilde{y}_n^2 &= y_{n-1} + 0.5hf(t_{n-1}, \tilde{y}_n^1), \\ \tilde{y}_n^3 &= y_{n-1} + 0.5hf(t_{n-0.5}, \tilde{y}_n^2), \\ \tilde{y}_n^4 &= y_{n-1} + 0.5hf(t_{n-0.5}, \tilde{y}_n^3), \end{aligned} \tag{4.51}$$

és ezután

$$y_n = y_{n-1} + \frac{h}{6}\left[f(t_{n-1}, \tilde{y}_n^1) + 2f(t_{n-0.5}, \tilde{y}_n^2) + 2f(t_{n-0.5}, \tilde{y}_n^3) + f(t_n, \tilde{y}_n^4)\right]. \tag{4.52}$$

Hosszadalmas számolással megmutatható, hogy a (4.51)-(4.52) explicit módszer negyedrendű [7].

4.2.1. A Runge-Kutta típusú módszerek rendje

Láthattuk, hogy számos explicit módszer állítható elő a Cauchy-feladatok megoldására. Célunk ezen módszerek paraméterezéssel történő egységes felírása. Ezért először írjuk át a módszereinket a köztes értékek új jelölésével!

- explicit Euler-módszer:

$$Y_1 = y_{n-1} \tag{4.53}$$

$$y_n = y_{n-1} + hf(t_{n-1}, Y_1). \tag{4.54}$$

- explicitté tett implicit Euler-módszer:

$$\begin{aligned} Y_1 &= y_{n-1}, \\ Y_2 &= y_{n-1} + hf(t_{n-1}, Y_1), \end{aligned} \tag{4.55}$$

$$y_n = y_{n-1} + hf(t_{n-1} + h, Y_2). \tag{4.56}$$

- explicit középponti módszer:

$$\begin{aligned} Y_1 &= y_{n-1}, \\ Y_2 &= y_{n-1} + 0.5hf(t_{n-1}, Y_1), \end{aligned} \tag{4.57}$$

$$y_n = y_{n-1} + hf(t_{n-1} + 0.5h, Y_2). \tag{4.58}$$

- explicit trapéz-módszer:

$$\begin{aligned} Y_1 &= y_{n-1}, \\ Y_2 &= y_{n-1} + hf(t_{n-1}, Y_1), \end{aligned} \tag{4.59}$$

$$y_n = y_{n-1} + 0.5h\left[f(t_{n-1}, Y_1) + f(t_{n-1}+h, Y_2)\right]. \tag{4.60}$$

- A Simpson-képleten alapuló explicit módszer:

$$\begin{aligned} Y_1 &= y_{n-1}, \\ Y_2 &= y_{n-1} + 0.5hf(t_{n-1}, Y_1), \\ Y_3 &= y_{n-1} + 0.5hf(t_{n-1}+0.5h, Y_2), \\ Y_4 &= y_{n-1} + 0.5hf(t_{n-1}+0.5h, Y_3), \end{aligned} \tag{4.61}$$

és ezután

$$\begin{aligned} &y_n = y_{n-1} + \\ &\frac{h}{6}\left[f(t_{n-1}, Y_1) + 2f(t_{n-1}+0.5h, Y_2) + 2f(t_{n-1}+0.5h, Y_3) + f(t_{n-1}+h, Y_4)\right]. \end{aligned} \tag{4.62}$$

Ezek alapján könnyen felírható az *általános paraméteres* alak:

4.5. Definíció *Legyen $s \geq 1$ egy adott egész szám. A rögzített a_{ij}, c_i, b_i $(i, j = 1, 2, \ldots, s)$ számok melletti*

$$Y_i = y_{n-1} + h\sum_{j=1}^{s} a_{ij} f(t_{n-1} + c_j h, Y_j), \quad i = 1, 2, \ldots, s \tag{4.63}$$

$$y_n = y_{n-1} + h\sum_{i=1}^{s} b_i f(t_{n-1} + c_i h, Y_i), \quad i = 1, 2, \ldots, s \tag{4.64}$$

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

Néhány észrevétel a (4.63)-(4.64) Runge-Kutta típusú módszer kapcsán.

- A módszer továbbra is egylépéses, hiszen y_{n-1} értékéből származtatja y_n értékét. Az s paraméter a lépcsőszámot jelenti: hány köztes pontban számolunk ki értékeket, amelyekből az y_n értékét meghatározzuk.

- A módszert adott s esetén a paraméterei határozzák meg, amelyek száma $s^2 + 2s$. Bevezetve az $\mathbf{A} = (a_{ij})_{i,j=1}^{s} \in \mathbb{R}^{s\times s}$, valamint a $\mathbf{c} = (c_i)_{i=1}^{s}, \mathbf{b} = (b_i)_{i=1}^{s} \in \mathbb{R}^s$

jelöléseket[2], a Runge-Kutta típusú módszer paramétereit jól áttekinthetően megadhatjuk a

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^\top \end{array} \tag{4.65}$$

alakban. Az ilyen típusú táblázatot *Butcher-táblázatnak* nevezzük.[3]

- A szakaszban említett explicit módszerekben mindenütt $Y_1 = y_{n-1}$ szerepelt, ezért ezekre a $\mathbf{c}$ vektor első koordinátája, valamint az $\mathbf{A}$ mátrix első minden sorának eleme egyaránt nulla.

- Az előzőekben említett explicit módszerek Butcher-táblázatai tehát a következők.

  1. explicit Euler-módszer:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \tag{4.66}$$

  2. explicitté tett implicit Euler-módszer:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 0 & 1 \end{array} \tag{4.67}$$

  3. explicit középponti módszer:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 0.5 & 0.5 & 0 \\ \hline & 0 & 1 \end{array} \tag{4.68}$$

  4. explicit trapéz-módszer:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 0.5 & 0.5 \end{array} \tag{4.69}$$

  5. A Simpson-képleten alapuló explicit módszer:

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} \tag{4.70}$$

[2] A vektorokat oszlopvektorként értelmezzük.

[3] John Butcher (1933–), Új-Zélandon élő és dolgozó matematikus.

- A fenti módszerek közül az explicitté tett középponti módszer és az explicit trapézmódszer másodrendűek. Mint az előző szakaszban azt már említettük, ezekre a módszerekre a következő módon hivatkozunk: a (4.57)-(4.58) alakú explicit középponti módszert *javított explicit Euler-módszernek*, a (4.59)-(4.60) alakú explicit trapéz-módszert pedig *Heun-módszernek* nevezzük. A Simpson-képleten alapuló negyedrendű, a (4.61)-(4.62) alakú explicit módszert *RK4-módszernek* szokás nevezni.

- Az explicit Runge-Kutta típusú módszerek sajátossága, hogy a módszerek $\mathbf{A}$ mátrixa szigorúan alsó háromszögű, azaz $a_{ij} = 0$ minden $j \geq i$ esetén. Ez azt teszi lehetővé, hogy a köztes értékeket közvetlenül, egyenletek megoldása nélkül is meghatározhatók: az egyes Y_i értékek közvetlenül kiszámolhatók a megelőző $Y_1, Y_2, \ldots, Y_{i-1}$ elemek ismeretében.

A fenti egylépéses, de többlépcsős módszerek ki is próbálhatóak. A program az alábbi linkről tölthető le:
../programok/onestep.exe

4.6. Megjegyzés A fenti Butcher táblázatok érdekes tulajdonsága, hogy az $\mathbf{A}$ mátrix tetszőleges i-ik sorának összege megegyezik c_i értékével. Vizsgáljuk meg ennek okát! A (4.63)-(4.64) Runge-Kutta típusú módszerek úgy vannak felépítve, hogy a kiszámolt Y_i értékek az $u(t)$ megoldásfüggvényt approximálják valamilyen $q \geq 1$ rendben a $t = t_{n-1} + c_i h$ pontokban. Ezért, ha a megoldásfüggvény konstans avagy elsőrendű polinom, akkor a módszernek pontosnak kell lennie. Tekintsük az

$$u' = 1, \quad u(0) = 0$$

feladatot. Mivel a feladat pontos megoldása $u(t) = t$, ezért a fentiek alapján a (4.63)-(4.64) által meghatározott Y_i értékekre teljesülni kell az $Y_i = u(t_{n-1} + c_i) = t_{n-1} + c_i h$ egyenlőségnek. Erre a feladatra $f(t, u) = 1$. Ezért az első lépésben (tehát $n = 1$, $t_0 = 0$ és $y_{n-1} = y_0 = 0$) a (4.63) alapján kiszámolt értékek

$$Y_i = y_{n-1} + h \sum_{j=1}^{s} a_{ij} f(t_{n-1} + c_j h, Y_j) = h \sum_{j=1}^{s} a_{ij}.$$

Ezért az $Y_i = c_i h$ feltétel a kívánt $c_i = \sum_{j=1}^{s} a_{ij}$, relációt eredményezi. Mindez azt jelenti, hogy a Butcher-táblázatban valójában $\mathbf{c}$ értéke az $\mathbf{A}$ mátrix által már meghatározott: az $\mathbf{e} = (1, 1, \ldots, 1)^T \in \mathbb{R}^s$ jelöléssel $\mathbf{c} = \mathbf{Ae}$, és így a Butcher-táblázat

$$\begin{array}{c|c} \mathbf{Ae} & \mathbf{A} \\ \hline & \mathbf{b}^\top \end{array} \tag{4.71}$$

alakú.

A gyakorlatban alkalmazott háromlépcsős, egyparaméteres módszer a következő:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{2}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{2}{3} - \frac{1}{4\alpha} & \frac{1}{4\alpha} & 0 \\ \hline & \frac{1}{4} & \frac{3}{4} - \alpha & \alpha \end{array} \tag{4.72}$$

A 4.6. megjegyzés szerint a (4.63)-(4.64) Runge-Kutta típusú módszer esetén az Y_i értékek a köztes pontokban az u megoldást approximálják. Az adott módszer egy másik lehetséges felírása, amikor nem a megoldásfüggvényt, hanem annak deriváltját, azaz $u'(t) = f(t, u(t))$ függvényt approximáljuk köztes értékekkel a $t_{n-1} + c_i h$ pontokban. Ekkor a módszer algoritmusa a következő:

$$k_i = f(t_{n-1} + c_i h, y_{n-1} + h \sum_{j=1}^{s} a_{ij} k_j), \quad i = 1, 2, \ldots, s, \tag{4.73}$$

$$y_n = y_{n-1} + h \sum_{i=1}^{s} b_i k_i. \tag{4.74}$$

Megjegyezzük, hogy a (4.73)-(4.74) és a (4.63)-(4.64) módszerek Butcher-táblázatai illetve pontossági feltételei megegyeznek, tehát ezeket a módszereket, mit egylépéses módszereket ekvivalensnek tekintjük.

A továbbiakban megvizsgáljuk a Runge-Kutta típusú módszerek konzisztenciájának feltételét.

Először szükséges feltételt adunk a p-ed rendű konzisztenciára. Ehhez alkalmazzuk a (4.63)-(4.64) Runge-Kutta típusú módszert az

$$u' = u + 1, \quad u(0) = 0$$

feladatra, amelynek pontos megoldása az $u(t) = e^t - 1$ függvény. Ezért a megoldás a $t = h$ pontban

$$u(h) = e^h - 1 = h + \frac{1}{2!}h^2 + \ldots + \frac{1}{p!}h^p + \mathcal{O}(h^{p+1}). \tag{4.75}$$

(Tehát itt a lokális approximációs hibát számoljuk ki.) Határozzuk meg az első lépés utáni közelítő megoldást. (Ekkor $n = 1$, $y_0 = 0$ és $t_0 = 0$.) Mivel erre a feladatra $f(t, u) = u + 1$, ezért a (4.63) alapján a köztes értékekre

$$Y_i = h \sum_{j=1}^{s} a_{ij}(Y_j + 1), \quad i = 1, 2, \ldots, s. \tag{4.76}$$

Bevezetve az $\mathbf{Y} = (Y_i)_{i=1,s} \in \mathbb{R}^s$ vektort, a korábban már használt $\mathbf{I}$ egységmátrix és $\mathbf{e}$ egyesekből álló vektor jelöléssel (4.76) felírható

$$\mathbf{Y} = h\mathbf{A}(\mathbf{Y} + \mathbf{e}) = h\mathbf{A}\mathbf{Y} + h\mathbf{A}\mathbf{e}$$

alakban, ahonnan

$$(\mathbf{I} - h\mathbf{A})\mathbf{Y} = h\mathbf{A}\mathbf{e}. \tag{4.77}$$

Mint ismeretes, megfelelően kis h esetén az $\mathbf{I} - h\mathbf{A}$ mátrix invertálható, és inverze előállítható az ún. Neumann-sor segítségével:

$$(\mathbf{I} - h\mathbf{A})^{-1} = \mathbf{I} + h\mathbf{A} + h^2\mathbf{A}^2 + \ldots + h^p\mathbf{A}^p + \ldots. \tag{4.78}$$

Ezért a (4.77) és a (4.78) összefüggések alapján $\mathbf{Y}$ kifejezhető a következő alakban:

$$\begin{aligned} \mathbf{Y} =& (\mathbf{I} - h\mathbf{A})^{-1}h\mathbf{A}\mathbf{e} = \\ & (\mathbf{I} + h\mathbf{A} + h^2\mathbf{A}^2 + \ldots + h^p\mathbf{A}^p + \ldots)h\mathbf{A}\mathbf{e} = \\ & (h\mathbf{A} + h^2\mathbf{A}^2 + \ldots + h^p\mathbf{A}^p + \ldots)\mathbf{e}. \end{aligned} \tag{4.79}$$

Ezért y_1 (4.64) értéke, a (4.79) összefüggés alapján:

$$\begin{aligned} & y_1 = h\sum_{i=1}^{s} b_i(Y_i + 1) = h\mathbf{b}^\top(\mathbf{Y} + \mathbf{e}) = \\ & h\mathbf{b}^\top\left[(h\mathbf{A} + h^2\mathbf{A}^2 + \ldots + h^p\mathbf{A}^p + \ldots)\mathbf{e} + \mathbf{e}\right] = \\ & h\mathbf{b}^\top\left[\mathbf{I} + h\mathbf{A} + h^2\mathbf{A}^2 + \ldots + h^p\mathbf{A}^p + \ldots)\mathbf{e}\right] = \\ & h\mathbf{b}^\top\mathbf{e} + h^2\mathbf{b}^\top A\mathbf{e} + h^3\mathbf{b}^\top\mathbf{A}^2\mathbf{e} + \ldots + h^p\mathbf{b}^\top\mathbf{A}^{p-1}\mathbf{e} + \ldots. \end{aligned} \tag{4.80}$$

Összevetve a (4.80) és a (4.75) kifejezéseket látható, hogy a Runge-Kutta típusú módszer p-ed rendűségének szükséges feltétele a következő:

$$\frac{1}{k!} = \mathbf{b}^\top\mathbf{A}^{k-1}\mathbf{e}, \quad k = 1, 2, \ldots, p. \tag{4.81}$$

Ez $k = 1$ esetén a

$$\mathbf{b}^\top \cdot \mathbf{e} = 1,$$

és $k = 2$ esetén, az $\mathbf{A}\mathbf{e} = \mathbf{c}$ összefüggés miatt, a

$$\mathbf{b}^\top \cdot \mathbf{c} = 0.5$$

feltételt jelenti.

Bevezetve az

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

jelöléseket, felírható egy Runge-Kutta típusú módszer p-ed rendű konzisztenciájának elégséges feltétele az $\mathbf{A}$ mátrix illetve a $\mathbf{b}$ sorvektor segítségével[4]

rend (p)	feltétel
1	$\mathbf{b}^\top \cdot \mathbf{e} = 1$
2	$\mathbf{b}^\top \cdot \mathbf{c} = 1/2$
3	$\mathbf{b}^\top \cdot \mathbf{c}^2 = 1/3, \quad \mathbf{b}^\top \cdot \mathbf{Ac} = 1/6$
4	$\mathbf{b}^\top \cdot \mathbf{c}^3 = 1/4, \quad \mathbf{b}^\top \cdot \mathbf{CAc} = 1/8,$ $\mathbf{b}^\top \cdot \mathbf{Ac}^2 = 1/12, \quad \mathbf{b}^\top \cdot \mathbf{A}^2\mathbf{c} = 1/24.$

(4.82)

Így, érvényes az alábbi

4.7. Tétel *A 4.65 Butcher-táblázatú Runge-Kutta típusú módszer pontosan akkor konzisztens, amikor teljesülnek a*

$$\boldsymbol{A}\boldsymbol{e} = \boldsymbol{c}; \quad \boldsymbol{b}^\top \cdot \boldsymbol{e} = 1 \tag{4.83}$$

feltételek, azaz

$$\sum_{k=1}^{s} a_{ik} = c_i \quad i = 1, 2, \ldots, s \quad \textit{és emellett} \quad \sum_{k=1}^{s} \boldsymbol{b}_k = 1. \tag{4.84}$$

4.8. Megjegyzés Mivel a (4.63)-(4.64) Runge-Kutta típusú módszerek esetén Y_i belső lépcső is approximál a megfelelő pontban, ezért értelmes egy Runge-Kutta típusú módszer belső rendjét is definiálni. Azt mondjuk, hogy egy s-lépcsős Runge-Kutta típusú módszer belső rendje q, ha a belső lépések soráni hiba minimális rendje q, azaz $\min_{i=1,\ldots,s}(u(t_{n-1} + c_i h) - Y_i) = \mathcal{O}(h^q)$. Nyilvánvalóan minden módszere esetén $q \leq p$, és a reális módszerek esetén $q \ll p$.

4.9. Példa *A (4.82) táblázatba való közvetlen behelyettesítéssel könnyen ellenőrizhető, hogy az alábbi háromlépcsős explicit Runge-Kutta típusú módszer harmadrendű.*

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 \\ 2/3 & 0 & 2/3 & 0 \\ \hline & 1/4 & 0 & 3/4 \end{array} \tag{4.85}$$

[4]Ezek a feltételek a korábban már ismertetett, de csak alacsonyabb lépcsőszámra ($s = 1, 2$) kiszámolt Taylor-sorba fejtéssel nyerhetők. A számolások már $s = 4$ esetén is meglehetősen bonyolultak. Magasabb rendű pontosság feltétele is megadható, de ekkor a feltételek száma is lényegesen megnő. (V.ö. a (4.91) táblázattal.)

4.10. Példa *Szintén harmadrendű a*

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \tag{4.86}$$

módszer.

Megjegyezzük, hogy a fenti harmadrendű módszerek gyakran szerepelnek az alkalmazásokban. Emellett $f(t,u) = f(t)$ esetén a (4.86) együtthatókkal definiált Runge-Kutta típusú módszer $\mathcal{O}(h^5)$ rendben pontos a Simpson-formulával, ezért alkalmazása abban az esetben különösen előnyös, amikor a $\partial_2 f$ parciális derivált közel van a nullához.

A (4.70) Butcher-táblázatú Runge-Kutta típusú módszer nyilván explicit, emellett negyedrendben pontos. (Ennek belátása a (4.82) táblázat feltételeinek egy meglehetősen hosszú, de matematikailag nem nehéz ellenőrzésével önállóan elvégezhető.)

Ezt a módszert szokás a (4.73)-(4.74) alakban alkalmazni. Ebben a felírásban a módszer algoritmusa (azaz a t_n pontbeli y_n közelítés meghatározása a már kiszámolt t_{n-1} pontbeli y_{n-1} közelítésből) a következő:

- A
$$\begin{aligned} k_1 &= f(t_{n-1}, y_{n-1}) \\ k_2 &= f(t_{n-1} + 0.5h, y_{n-1} + 0.5hk_1) \\ k_3 &= f(t_{n-1} + 0.5h, y_{n-1} + 0.5hk_2) \\ k_4 &= f(t_{n-1} + h, y_{n-1} + hk_3) \end{aligned} \tag{4.87}$$
képletekkel rendre kiszámoljuk a k_1, k_2, k_3 és k_4 értékeket.

- Az
$$y_n = y_{n-1} + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \tag{4.88}$$
képlettel meghatározzuk az új közelítést.

4.11. Megjegyzés Az RK4 módszer ezen alakja már közvetlenül láthatóan kapcsolatba hozható a Simpson-féle numerikus integrálással. Ugyanis, ha a (4.26) jobb oldalának numerikus integrálására alkalmazzuk a Simpson-féle szabályt, akkor a

$$\frac{h_n}{6}\left[f(t_{n-1}, u(t_{n-1})) + 4f(t_{n-0.5}, u(t_{n-0.5})) + f(t_n, u(t_n))\right]$$

közelítést nyerjük. Ezzel tehát az

$$y_n - y_{n-1} = \frac{h_n}{6}\left[f(t_{n-1}, y_{n-1}) + 4f(t_{n-0.5}, y_{n-0.5}) + f(t_n, y_n)\right], \tag{4.89}$$

numerikus integráló formulát kapjuk. Mivel k_i a megoldásfüggvény deriváltjának közelítése a $t_{n-1} + c_i h$ pontban, és $u'(t_{n-1} + c_i h) = f(t_{n-1} + c_i h, u(t_{n-1} + c_i h))$, ezért $k_1 \approx u'(t_{n-1})$, $k_4 \approx u'(t_n)$, és k_2, k_3 két különböző közelítése az $u'(t_{n-0.5})$ értéknek. Legyen tehát ezek számtani közepe az approximáció, azaz $(k_2 + k_3)/2 \approx u'(t_{n-0.5})$. Ezen approximációkkal a (4.89) képlet éppen a (4.88) formulát adja.

Mi a kapcsolat az explicit Runge-Kutta típusú módszerek lépcsőszáma (s) és rendje (p) között? Azt láttuk, hogy az egylépcsős explicit Euler-módszer elsőrendű, a kétlépcsős módszerek (Heun-módszer, javított explicit Euler-módszer) másodrendűek, a háromlépcsős (4.85)-(4.86) Butcher-táblázatú módszerek harmadrendűek, a négylépcsős (4.87)-(4.88) módszer pedig negyedrendben pontos. Tehát $s = 1, 2, 3, 4$ esetén a az elérhető konzisztenciarend elérhető a lépcsőszámmal, azaz a $p = s$ pontosság biztosítható. Ugyanakkor, $s \geq 5$ esetén ez már nem érvényes, a rend alatta marad a lépcsőszámnak, azaz $p < s$. A közöttük lévő kapcsolat az első tíz lépcsőszámú explicit Runge-Kutta típusú módszerre a következő:

s	1	2	3	4	5	6	7	8	9	10
$p(s)$	1	2	3	4	4	5	6	6	7	7

(4.90)

Az explicit Runge-Kutta típusú módszer esetén nem ismeretes, hogy egy tetszőleges p rend eléréséhez hány lépcső szükséges. Ez egyelőre csak $p = 8$ értékig ismert. Az alábbi érdekes táblázatban megadjuk, hogy ezen értékig hány feltétel, illetve hány paraméter szerepel.

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

(4.91)

A táblázat választ ad arra is, hogy miért nem lehetséges ötödrendű explicit Runge-Kutta típusú módszert adni az ötlépcsős módszerek körében: erre az esetre a feltételek száma 17, míg a szabad paraméterek száma csak 15.

4.12. Megjegyzés Az explicit Runge-Kutta típusú módszerek adott rendjét biztosító feltételek megadása a megfelelő Taylor-sorba fejtésből adódnak, es ezért általában nagyon bonyolult, nemlineáris egyenleteket jelentenek. Mivel a feltételek függetlensége általában nem mutatható ki, ezért általában nem igaz, hogy a szabad paraméterek száma (azaz s értéke) előre kiszámíthatóan megadja az elérhető pontosságot. Ezért nem mondható meg, hogy egy tetszőleges p-ed rendű módszer hány lépcsős explicit Runge-Kutta típusú módszerrel érhető el. A fenti táblázatból kiolvasható a pontos válasz $p = 8$-ig. A $p = 9$ esetre csak az ismeretes, hogy $12 \leq s \leq 17$, míg $p = 10$ esetén $13 \leq s \leq 17$. A $p > 10$ esetre nem ismerjük a választ. [7]

4.2.2. Az explicit Runge-Kutta típusú módszerek konvergenciája

A Runge-Kutta típusú módszerek felírhatók a már korábban tárgyalt (3.57) alakban, azaz az

$$y_n = y_{n-1} + h\Phi(h, t_{n-1}, y_{n-1}, y_n) \tag{4.92}$$

egylépéses módszerként, ahol Φ a numerikus módszert meghatározó adott függvény. Ha explicit Runge-Kutta típusú módszert vizsgálunk, akkor Φ alakja $\Phi(h, t_{n-1}, y_{n-1})$, azaz a módszerünk

$$\frac{y_n - y_{n-1}}{h} = \Phi(h, t_{n-1}, y_{n-1}) \tag{4.93}$$

alakú. Emlékeztetünk, hogy a 3.24. tételben megmutattuk: ha a módszer p-ed rendben konzisztens, emellett Φ folytonos, és a harmadik változójában lipschitzes, azaz valamely $h_0 > 0$ és $L_\Phi > 0$ állandóval

$$|\Phi(h, t_n, s_1) - \Phi(h, t_n, s_2)| \leq L_\Phi |s_1 - s_2| \tag{4.94}$$

tetszőleges $t_n \in \omega_h$ és $h < h_0$ esetén, akkor a módszer p-ed rendben konvergens.

Mindez azt jelenti, hogy elegendő megmutatni, hogy az egyes módszerekhez tartozó Φ függvények ilyen tulajdonságúak.

4.13. Példa *Mutassuk meg, hogy az explicit Euler-módszerhez tartozó Φ függvény kielégíti a 3.24. tétel feltételeit!*

Az explicit Euler-módszer alakjából nyilvánvalóan

$$\Phi(h, t_{n-1}, y_{n-1}) = f(t_{n-1}, y_{n-1}).$$

Ezért f folytonosságából Φ folytonossága következik. Emellett, f második változója szerinti lipschitzessége az $L > 0$ állandóval pontosan a 4.94 tulajdonságot jelenti, és emellett $L_\Phi = L$. Így erre az esetre a tétel alkalmazható.

4.14. Példa *Mutassuk meg, hogy az explicit trapéz-módszerhoz tartozó Φ függvény kielégíti a 3.24. tétel feltételeit!*

A (4.59) szerinti explicit trapéz-módszer esetén a Φ függvény alakja:

$$\Phi(h, t_{n-1}, y_{n-1}) = 0.5\left[f(t_{n-1}, y_{n-1}) + f(t_{n-1} + h, y_{n-1} + hf(t_{n-1}, y_{n-1}))\right]$$

Ha f folytonos, akkor Φ is az. Tegyük fel megint, hogy az f függvény a második változója szerinti lipschitzes az $L > 0$ állandóval. Ekkor

$$\begin{aligned}
&\Phi(h, t_{n-1}, s_1) - \Phi(h, t_{n-1}, s_2) = \\
&0.5\left[f(t_{n-1}, s_1) + f(t_{n-1} + h, s_1 + hf(t_{n-1}, s_1))\right] - \\
&0.5\left[f(t_{n-1}, s_2) + f(t_{n-1} + h, s_2 + hf(t_{n-1}, s_2))\right] \leq \\
&0.5\left[f(t_{n-1}, s_1) - f(t_{n-1}, s_2)\right] + \\
&0.5\left[f(t_n, s_1 + hf(t_{n-1}, s_1)) - f(t_n, s_2 + hf(t_{n-1}, s_2))\right].
\end{aligned} \tag{4.95}$$

Nyilvánvalóan

$$|f(t_{n-1}, s_1) - f(t_{n-1}, s_2)| \leq L|s_1 - s_2|, \tag{4.96}$$

továbbá

$$\begin{aligned} &|f(t_n, s_1 + hf(t_{n-1}, s_1)) - f(t_n, s_2 + hf(t_{n-1}, s_2))| \leq \\ &L|\,(s_1 + hf(t_{n-1}, s_1)) - (s_2 + hf(t_{n-1}, s_2))\,| \leq L|s_1 - s_2| + \\ &Lh|f(t_{n-1}, s_1) - f(t_{n-1}, s_2)| \leq L|s_1 - s_2| + L^2h|s_1 - s_2|. \end{aligned} \tag{4.97}$$

A (4.96) és a (4.96) becslések felhasználásával (4.95) összefüggésből látható, hogy Φ függvényre teljesül a 4.94 tulajdonság, és emellett $L_\Phi = 0.5(2L + hL^2)$.

Teljes indukcióval megmutatható, hogy az f függvény folytonossága és második változója szerint lipschitzessége esetén explicit Runge-Kutta típusú módszeresetén a Φ függvény kielégíti a 3.24. tétel feltételeit.

Ezért érvényes az alábbi állítás.

4.15. Tétel *Tegyük fel, hogy a (4.24) kezdetiérték-feladatban f folytonos és a második változójában lipschitzes. Ekkor egy p-ed rendben konzisztens explicit Runge-Kutta típusú módszer p-ed rendben konvergens is.*

Egy Φ-módszer konzisztenciáját általában nehéz ellenőrizni, pontosabban, meglehetősen munkaigényes. Ugyanakkor ha csak a módszer konvergenciáját akarjuk belátni, és a rendjére nem vagyunk kíváncsiak, akkor egyszerűbben is eljárhatunk. Ugyanis, tegyük fel, hogy a módszer $\Phi(h, t_{n-1}, y_{n-1})$ függvénye az első változójában folytonos a $h = 0$ pontban, $\Phi(0, t_{n-1}, y_{n-1}) = f(t_{n-1}, y_{n-1})$, és érvényes a (4.94) harmadik változó szerinti lipschitzesség. Ezen feltételek mellett megmutatható a módszer konvergenciája a $t = t^\star$ pontban. Ugyanis a korábbi jelölések megtartásával a $t_n = t^\star$ pontbeli e_n globális hibára érvényes az

$$|e_n| \leq e^{t^\star L_\Phi} \left(|e_0| + t^\star \max_{1 \leq i \leq n} |d_i| \right)$$

szokásos becslés. Mivel $e_0 = 0$, ezért tehát

$$|e_n| \leq e^{t^\star L_\Phi} t^\star \max_{1 \leq i \leq n} |d_i|.$$

A lokális approximációs hiba felírható

$$\begin{aligned} d_i = &\frac{u(t_i) - u(t_{i-1})}{h} - \Phi(h, t_{i-1}, u(t_{i-1})) = \\ &u'(t_{i-1}) + \frac{h}{2}u''(t_{i-1}) + \mathcal{O}(h^2) - \Phi(h, t_{i-1}, u(t_{i-1})) \end{aligned} \tag{4.98}$$

alakban. A feltételek miatt

$$\begin{aligned} \lim_{h \to 0} \Phi(h, t_{i-1}, u(t_{i-1})) = &\Phi(0, t_{i-1}, u(t_{i-1})) = \\ &f(t_{i-1}, u(t_{i-1})) = u'(t_{i-1}). \end{aligned} \tag{4.99}$$

Ezért tehát $\lim_{h\to 0} d_i = 0$, és a rendjére $d_i = \mathcal{O}(h^p)$. Tehát

$$|e_n| \leq e^{t^\star L_\Phi} t^\star \mathcal{O}(h^p),$$

ami a kívánt $\lim_{h\to 0} e_n = 0$ relációt jelenti.

Vegyük észre, hogy az f szokásos tulajdonságai mellett az explicit Runge-Kutta típusú módszerre a Φ függvény $h = 0$ pontban folytonos, és $h = 0$ esetén a kívánt $\Phi(0, t_{n-1}, y_{n-1}) = f(t_{n-1}, y_{n-1})$ feltétel teljesül. Mivel a (4.94) feltétel szintén teljesül, ezért az előbbi levezetéssel beláttuk az alábbi állítást.

4.16. Tétel *Tegyük fel, hogy a (4.24) kezdetiérték-feladatban f folytonos és a második változójában lipschitzes. Ekkor egy explicit Runge-Kutta típusú módszer konvergens is.*

4.17. Megjegyzés Mint a levezetésből is látható, nem szükséges az $e_0(h) = 0$ feltétel, azaz nem feltétlenül kell az y_0 értékét u_0-nak választani. A konvergenciához elegendő a $\lim_{h\to 0} e_0(h) = 0$ teljesülése. (A p-ed rendű konvergencia megőrzéséhez pedig az $e_0(h) = \mathcal{O}(h^p)$ feltétel.)

4.2.3. A Runge-Kutta típusú módszerek lépéshosszának megválasztása. Hibaanalízis.

A Runge-Kutta típusú módszerek hibabecslése általában bonyolult. A konvergencia rendjének eléréséhez megfelelően kis h érték megválasztása szükséges, de konkrét feladatok esetén ennek meghatározása nem kézenfekvő. Másrészt, mint azt az előzőekben is láttuk, a Runge-Kutta típusú módszerekre a hibaanalízis meglehetősen összetett feladat. A továbbiakban megvizsgáljuk, hogy egy adott módszerhez hogyan lehet hatékonyan lépéshosszúságot választani.

Egyik lehetséges eljárás, amikor az adott, p-ed rendben konzisztens numerikus módszerrel két módon lépünk át a t_{n-1} rácspontból a t_n pontba: egyszer egy h lépéssel, másodszor pedig két, egymást követő $h/2$ lépéssel. A két numerikus eredmény eltérésének analízise lehetőséget ad a lépésköz hatékony megválasztására, nevezetesen az l_n (vagy d_n) lokális diszkretizációs hiba vizsgálatával. (V.ö. a 3.17. definícióval.) Jelölje y_n a h lépésközzel nyert számolt numerikus megoldást, míg $\hat{y}_n$ a két, $h/2$ lépésközű numerikus módszer által nyert, és a pontos kezdeti értékből számolt közelítést. Ekkor

$$u(t_n) = y_n + Ch^{p+1} + \mathcal{O}(h^{p+2}), \tag{4.100}$$

$$u(t_n) = \hat{y}_n + 2C\left(\frac{h}{2}\right)^{p+1} + \mathcal{O}(h^{p+2}), \tag{4.101}$$

tehát

$$u(t_n) = \hat{y}_n + C\frac{h^{p+1}}{2^p} + \mathcal{O}(h^{p+2}). \tag{4.102}$$

Ezekből az összefüggésekből

$$y_n - \hat{y}_n = Ch^{p+1}\left(1 - \frac{1}{2^p}\right) + \mathcal{O}(h^{p+2}), \tag{4.103}$$

azaz C ún. *hibakonstans* egy jó becslése

$$C = \frac{|y_n - \hat{y}_n|}{(1 - 2^{-p})\, h^{p+1}}. \tag{4.104}$$

Mivel ez a becslés egyben a megválasztására is használható, ezért a (4.102) képletbe behelyettesítve a C állandó (4.104) szerinti értékét, az

$$u(t_n) = \hat{y}_n + \varepsilon + \mathcal{O}(h^{p+2}) \tag{4.105}$$

összefüggést nyerjük, ahol

$$\varepsilon = \frac{|y_n - \hat{y}_n|}{2^p - 1}. \tag{4.106}$$

A (4.106) képlet szerinti ε számot a csonkolási hiba mutatójának is szokás nevezni.

Vegyük észre, hogy (4.105) és (4.106) következtében az

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

közelítésre

$$u(t_n) = y_n^* + \mathcal{O}(h^{p+2}),$$

azaz a lokális diszkretizációs hiba rendje erre a közelítésre egyel nagyobb.

Egy adott ε_0 pontossági kritériumhoz a lépéshossz megválasztására is felhasználható ez az eljárás. Határozzuk meg azt a h_{uj} új lépéshosszt, amellyel a (4.101) szerint kiszámított $\hat{y}_n$ hibája ε_0-nál kisebb lesz! Mivel a hiba a képlet szerint $2C\left(\frac{h_{uj}}{2}\right)^{p+1}$ nagyságú, ezért a feltételünk

$$C\frac{h^{p+1}}{2^p} \leq \varepsilon_0.$$

Figyelembe véve C (4.104) értékét, ez a

$$\frac{|y_n - \hat{y}_n|}{2^p - 1} \cdot \frac{h_{uj}^{p+1}}{h^{p+1}} \leq \varepsilon_0$$

feltételt jelenti, ami a (4.106) jelöléssel az

$$\varepsilon \cdot \frac{h_{uj}^{p+1}}{h^{p+1}} \leq \varepsilon_0.$$

Ezek alapján az új lépéshosszúságra a

$$h_{uj} = h \cdot \left(\frac{\varepsilon_0}{\varepsilon}\right)^{\frac{1}{p+1}} \tag{4.107}$$

feltételt eredményezi. A fenti becslésekben közelítéseket használtunk. Ezért megbízhatósági meggondolásból a szakirodalomban egy egynél kisebb, de ahhoz közeli β számmal szokásos módosítani a becslést, és a

$$h_{uj} = \beta h \cdot \left(\frac{\varepsilon_0}{\varepsilon}\right)^{\frac{1}{p+1}} \tag{4.108}$$

képlettel számolják az új lépésközt, ahol $\beta \in (0.8, 0.9)$. A (4.108) azt jelenti, hogy $\varepsilon > \varepsilon_0$ esetén a következő lépésben csökkenteni kell a lépésközt, míg $\varepsilon < \varepsilon_0$ esetén növelni lehet a következő lépésben. Megjegyezzük, hogy ez az eljárás a lépésköz megválasztásának adaptivitását eredményezi, azaz a meglévő eredményekből tudunk következtetni az új lépésre.

A hiba és a lépéshossz megválasztásának egy másik lehetséges módja, amikor két különböző módszer eredményeiből következtetünk a megoldás pontosságára, illetve a lépésköz megválasztására. Tipikus megválasztás, amikor két *különböző rendű* módszert választunk, amelyek rendje p és $p+1$. Jelölje y_n és $\hat{y_n}$ az ezen módszerekhez tartozó numerikus megoldásokat. Ekkor a kevésbé pontos megoldás hibáját, illetve a megfelelő lépésközt a két numerikus megoldás különbségéből becsüljük. (Lásd pl. [19]) Az alapötlet, hogy a magasabb rendű ($p+1$-ed rendű) módszert használjuk az alacsonyabb rendű megoldás pontosságának becslésére, mégpedig úgy, hogy a magasabb rendű módszer eredményét tekintjük a pontos megoldás egy jó közelítésének, ezért ezzel számoljuk ki az alacsonyabb rendű módszer globális hibáját. A számítási munka minimalizálása céljából célszerű olyan Runge-Kutta típusú módszeret választani, amelyek $\mathbf{A}$ mátrixa megegyezik, és csak a köztes értékeket súlyozó vektorok különböznek. Jelölje tehát a $\mathbf{b}$ vektor a p-ed rendű módszer, és $\hat{\mathbf{b}}$ vektor a $p+1$-ed rendű módszer súlyait tartalmazó vektort. A két módszert összefoglalóan az alábbi Butcher-táblázattal írhatjuk fel:

$$\begin{array}{c|c} \mathbf{Ae} & \mathbf{A} \\ \hline & \mathbf{b}^\top \\ & \hat{\mathbf{b}}^\top \end{array} \tag{4.109}$$

Ilyenkor azt mondjuk, hogy az alacsonyabb rendű módszer *beágyazott módszere* a magasabb rendű Runge-Kutta típusú módszernek. Az ilyen módszerpárok esetén a globális

hibát az $e_n = y_n - \hat{y_n}$ képlettel becsüljük, azaz az explicit Runge-Kutta típusú módszer (4.73)-(4.74) alakjából

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

ahol b_i és $\hat{b}_i$ a $\mathbf{b}$ és $\hat{\mathbf{b}}$ vektorok koordinátái. Ennek nagyságából egyrészt a hibára, másrészt az új lépésköz megválasztására is tudunk következtetni.

4.18. Megjegyzés Mutassuk meg, hogyan lehet a lépésköz megválasztására használni a beágyazott módszereket![5] Mivel az alacsonyabb rendű módszer p-ed rendű, a magasabb pedig $p+1$, ezért a lokális csonkolási hiba egy lépés után $u(t_n) = y_n + C_1 h^{p+1} + ...$, illetve $u(t_n) = \hat{y}_n + C_1 h^{p+2} + ...$ Az első összefüggés alapján a p-ed rendű módszer globális hibájára $|e_n| \approx C_1 h^{p+1}$. Másrészt, a két reláció különbségéből szintén az $|\hat{y}_n - y_n| = C_1 h^{p+1} + ...$ becslést kapjuk. Ezért egy adott ε_0 pontosság eléréséhez az $|\hat{y}_n - y_n| < \varepsilon_0$ egyenlőtlenséget ellenőrizzük. ha ez teljesül, akkor y_n értékét elfogadjuk az új közelítésnek. Ha nem teljesül, akkor egy új, kisebb h_{uj} lépésközt választunk. Ennek megválasztására ismét felhasználhatjuk a fenti relációt, ugyanis olyan h_{uj} szükséges, amelyre $e_n \approx C_1 h_{uj}^{p+1} < \varepsilon_0$. Mivel $|e_n| \approx C_1 h^{p+1}$, ezért

$$\frac{C_1 h_{uj}^{p+1}}{C_1 h^{p+1}} < \frac{\varepsilon_0}{e_n},$$

ahonnan a korábbi esethez hasonló

$$h_{uj} = h \cdot \left(\frac{\varepsilon_0}{e_n} \right)^{\frac{1}{p+1}} \tag{4.110}$$

feltételt kapjuk.

Tekintsünk néhány példát!

4.19. Példa *Adjunk példát egy másodrendű explicit Runge-Kutta típusú módszer elsőrendű beágyazott módszerére!*

Tekintsük az explicit trapéz-módszert! Könnyen láthatóan ennek beágyazott módszere az explicit Euler-módszer. A módszerpár Butcher-táblázata a következő:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1 & 0 \\ & 0.5 & 0.5 \end{array} \tag{4.111}$$

[5] Ez a megjegyzés természetesen nem csak a beágyazott módszerekre vonatkozik. A lépéshossz ezen adaptív megválasztása *bármely* két, megfelelő rendű módszerre alkalmazható.

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

4.1. táblázat. A beágyazott Dormand-Prince RK-módszer paraméterei (a nulla elemeket nem jelöltük.)

4.20. Példa *Adjunk beágyazott (másodrendű) módszert a (4.86) harmadrendű explicit Runge-Kutta típusú módszerhez!*

A konzisztencia rendjére vonatkozó (4.82) feltételből könnyen látható, hogy olyan b_1, b_2 és b_3 értékeket kell megválasztanunk, amelyekre

$$b_1 + b_2 + b_3 = 1$$

$$\frac{b_2}{2} + b_3 = \frac{1}{2}.$$

Ennek megoldása:

$$b_1 = b_3 = b, \quad b_2 = 1 - 2b,$$

ahol b tetszőleges paraméter, de $b \neq 1/6$. Így tehát a beágyazott párok Butcher-táblázata

0	0	0	0
$1/2$	$1/2$	0	0
1	-1	2	0
	b	$1-2b$	b
	$1/6$	$2/3$	$1/6$

(4.112)

(Speciálisan, $b = 0$ esetén a javított explicit Euler-módszert, míg $b = 1/3$ esetén az ún. szimmetrikus sémát kapjuk.)

Nagyon fontos és a Matlab alkalmazza az ún. Dormand–Prince beágyazott módszerpárt, amelyek négy illetve ötödrendűek. Ezekre később a 8.1 fejezetben még visszatérünk, ezért itt csak a 4.1 táblázatban csak Butcher-táblázatát adjuk meg.

0	0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
$\frac{3}{4}$	0	$\frac{3}{4}$	0	0
1	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	0
	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	0
	$\frac{7}{24}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{8}$

4.2. táblázat. A beágyazott Bogacki-Shampine RK-módszer paraméterei.

Egy másik, ugyancsak gyakran használt beágyazott Runge-Kutta típusú módszer a Bogacki-Shampine-módszer, amelynek Butcher-táblázata a következő:

Befejezésül megjegyezzük, hogy nem minden módszernek van beágyazott módszere. Például, a már ismertetett RK4-módszernek nincs megfelelő beágyazott módszerpárja. (Lásd [19].)

4.2.4. Az explicit Runge-Kutta típusú módszerek abszolút stabilitása

Ebben a szakaszban megvizsgáljuk az explicit Runge-Kutta típusú módszerek abszolút stabilitását, és megvizsgáljuk azt a kérdést, hogy vannak-e A-stabil explicit Runge-Kutta típusú módszerek.

Ezért a továbbiakban az explicit Runge-Kutta típusú módszerek viselkedését vizsgáljuk az

$$u' = \lambda u, \quad t > 0 \tag{4.113}$$

tesztegyenleten. A stabilitási tartomány felírásához első lépésben a tesztfeladatra alkalmazott explicit Runge-Kutta típusú módszerhez tartozó $R(z)$ stabilitási függvény meghatározása szükséges.

A (4.63)-(4.64) Runge-Kutta típusú módszer t alkalmazva a (4.113) tesztegyenletre, a korábbi jelöléseink megtartásával (4.63) alapján érvényes az

$$\mathbf{Y} = y_{n-1}\mathbf{e} + h\mathbf{A}\lambda\mathbf{Y} = y_{n-1}\mathbf{e} + z\mathbf{A}\mathbf{Y}$$

összefüggés, ahol $z = h\lambda$. Innen

$$(\mathbf{I} - z\mathbf{A})\mathbf{Y} = y_{n-1}\mathbf{e}. \tag{4.114}$$

Mint korábban szerepelt, megfelelően kis z (azaz h) esetén az $\mathbf{I}-z\mathbf{A}$ mátrix invertálható, és inverze előállítható a Neumann-sora segítségével. Tehát

$$\mathbf{Y} = (\mathbf{I} - z\mathbf{A})^{-1}\mathbf{e}y_{n-1} \tag{4.115}$$

Másrészt, a (4.64) alapján

$$\begin{aligned} y_n = y_{n-1} + h\sum_{i=1}^{s} b_i \lambda Y_i = y_{n-1} + z\mathbf{b}^\top \mathbf{Y} = \\ y_{n-1} + z\mathbf{b}^\top(\mathbf{I} - z\mathbf{A})^{-1}\mathbf{e}y_{n-1} = \left[1 + z\mathbf{b}^\top(\mathbf{I} - z\mathbf{A})^{-1}\mathbf{e}\right] y_{n-1}. \end{aligned} \tag{4.116}$$

Tehát a Runge-Kutta típusú módszerek stabilitási függvénye

$$R(z) = 1 + z\mathbf{b}^\top(\mathbf{I} - z\mathbf{A})^{-1}\mathbf{e}.$$

Felírva $(\mathbf{I} - z\mathbf{A})^{-1}$ Neumann sorát, érvényes az alábbi összefüggés:

$$\begin{aligned} R(z) = 1 + z\mathbf{b}^\top(\mathbf{I} + z\mathbf{A} + z^2\mathbf{A}^2 + \ldots + ...)\mathbf{e} = \\ 1 + z\mathbf{b}^\top\left(\sum_{k=0}^{\infty} z^k\mathbf{A}^k\right)\mathbf{e} = 1 + \mathbf{b}^\top\left(\sum_{k=0}^{\infty} z^{k+1}\mathbf{A}^k\right)\mathbf{e} = \\ 1 + \mathbf{b}^\top\left(\sum_{k=1}^{\infty} z^k\mathbf{A}^{k-1}\right)\mathbf{e} = 1 + \sum_{k=1}^{\infty} z^k\left(\mathbf{b}^\top\mathbf{A}^{k-1}\mathbf{e}\right). \end{aligned} \tag{4.117}$$

Mint megmutattuk, (v.ö. (4.81)) a Runge-Kutta típusú módszerek p-ed rendűségének szükséges feltétele a

$$\frac{1}{k!} = \mathbf{b}^\top\mathbf{A}^{k-1}\mathbf{e}, \quad k = 1, 2, \ldots, p. \tag{4.118}$$

Ezért egy p-ed rendű Runge-Kutta típusú módszer stabilitási függvénye a (4.117) és a (4.118) alapján:

$$R(z) = 1 + \sum_{k=1}^{p} z^k\frac{1}{k!} + \sum_{k=p+1}^{\infty} z^k\left(\mathbf{b}^\top\mathbf{A}^{k-1}\mathbf{e}\right). \tag{4.119}$$

Tehát *egy tetszőleges, p-ed rendű Runge-Kutta típusú módszer* stabilitási függvénye (4.119) alakú. Most tegyük fel, hogy a módszerünk s lépcsős, p-ed rendben pontos explicit Runge-Kutta típusú módszer! Ekkor $\mathbf{A} \in \mathbb{R}^{s\times s}$ egy szigorúan alsó háromszögű mátrix, ezért

$$\mathbf{A}^k = 0, \quad k = s+1, s+2, \ldots$$

esetén. Így az ilyen módszerek stabilitási függvénye

$$R(z) = 1 + \sum_{k=1}^{p} z^k\frac{1}{k!} + \sum_{k=p+1}^{s} z^k\left(\mathbf{b}^\top\mathbf{A}^{k-1}\mathbf{e}\right). \tag{4.120}$$

Fontos megjegyeznünk, hogy a maximális rendű explicit Runge-Kutta típusú módszerek stabilitási függvényei $p = 1, 2, 3, 4$ esetén ugyanolyan alakúak, hiszen a $p = s$ reláció miatt ekkor mindegyik módszerre

$$R(z) = 1 + \sum_{k=1}^{p} z^k \frac{1}{k!}, \quad p = 1, 2, 3, 4, \tag{4.121}$$

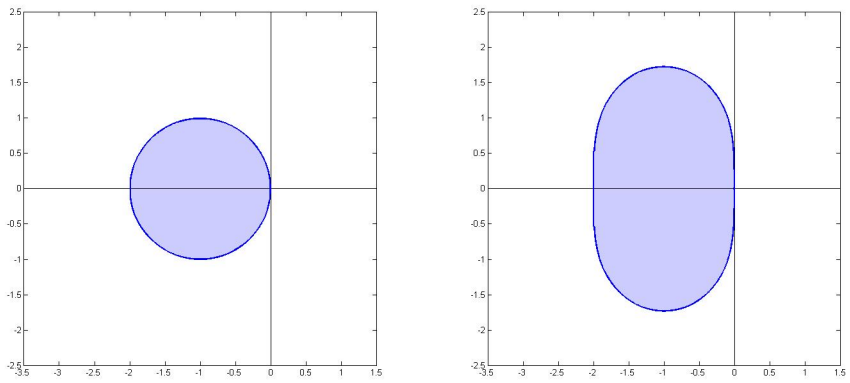
mivel ebben az esetben (4.120) jobb oldalán a második szumma eltűnik.

A stabilitási függvény a (4.120) alakú előállításából látszik, hogy az explicit Runge-Kutta típusú módszer stabilitási függvénye egy legfeljebb s-ed fokú polinom. Ezért $|z| \to \infty$ esetén a $\lim |R(z)| = \infty$, ezért ebben az esetben a stabilitási tartomány nem tartalmazhatja a teljes $\mathbb{C}^-$ félsíkot. Így érvényes a következő állítás.

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

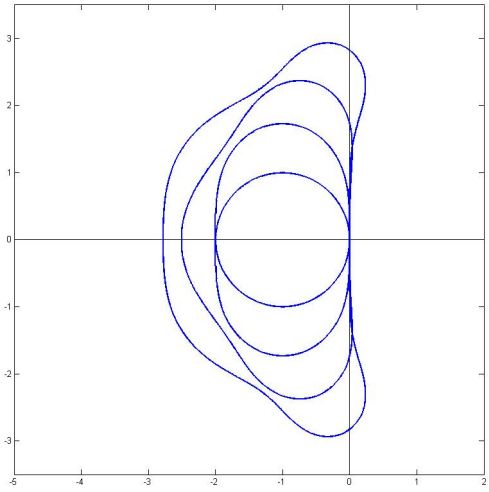
Az explicit Runge-Kutta típusú módszerekről további részletek megtalálhatóak a
http://math.fullerton.edu/mathews/n2003/RungeKuttaMod.html
linken.

Az explicit Runge-Kutta típusú módszer abszolút stabilitási tartománya a $p = 1$ és $p = 2$ rendre a 4.1 ábrán látható.



4.1. ábra. Az ERK1 és ERK2 explicit Runge-Kutta típusú módszer abszolút stabilitási tartománya.

Az ábrából látható, hogy a magasabb rendű módszer abszolút stabilitási tartomány bővebb. Növelve p értékét, az abszolút stabilitási tartomány tovább növekszik. Az explicit Runge-Kutta típusú módszer abszolút stabilitási tartománya a $p = 1, 2, 3, 4$ rendekre összefoglalóan a 4.2 ábrán látható.



4.2. ábra. Az explicit Runge-Kutta-módszerek abszolút stabilitási tartományai a $p = 1, 2, 3, 4$ esetekben.

4.3. Az implicit Runge-Kutta típusú módszerek

A 4.21. tételben kimondtuk, hogy az explicit Runge-Kutta típusú módszerek nem A-stabilak. Ez a tény az alkalmazhatóságukat erősen bekorlátozza, különösen merev feladatok megoldása esetén. Ezért célszerű az implicit módszerek vizsgálata.

Vegyük észre, hogy általánosan a Runge-Kutta típusú módszerek a következő módon definiálhatók. Legyen $\mathbf{A} \in \mathbb{R}^{s\times s}$ egy tetszőleges mátrix, $\mathbf{b} \in \mathbb{R}^s$ egy adott vektor, és $\mathbf{c} = \mathbf{Ae}$. Egy tetszőleges Runge-Kutta típusú módszert ekkor az ezen elemekkel definiált Butcher-táblázattal adhatunk meg. Az olyan módszereket, amelyekre $\mathbf{A}$ nem szigorúan alsó háromszögű mátrix, *implicit Runge-Kutta típusú módszernek* (IRK) nevezzük. Amikor $\mathbf{A}$ alsó (de nem szigorúan alsó) háromszögű mátrix, akkor a módszert *diagonálisan implicit Runge-Kutta típusú módszernek* (DIRK) nevezzük. A DIRK esetén a k_i (vagy Y_i) értékének kiszámolása, az explicit Runge-Kutta típusú módszerektől eltérően, egy (általában nemlineáris) egyenlet megoldását, míg az implicit Runge-Kutta típusú módszer esetén egy m ismeretlenes (általában nemlineáris) egyenletrendszer megoldását igényli. Ez a módszer alkalmazását bonyolultabbá teszi, viszont az ilyen típusú módszerek általában jobb stabilitási tulajdonságokkal rendelkeznek.

Mint majd megmutatjuk, az implicit Runge-Kutta típusú módszerek fontosak és a gyakorlatban használatosak, elsősorban a parciális differenciálegyenletek numerikus megoldása során. Ennek alapvetően két oka van:

- általában jó stabilitási tulajdonságokkal rendelkeznek,
- ugyanolyan lépcsőszám mellett magasabb rendben pontosak.

Az explicit Runge-Kutta típusú módszerekkel összehasonlítva, ennek természetesen "ára" van, nevezetesen a nagyobb számítási munka.

Ebben a szakaszban ismertetünk több nevezetes implicit Runge-Kutta típusú módszert, megadva a módszer Butcher-táblázatát, és megvizsgálva, hogy milyen numerikus algoritmus tartozik az adott táblázathoz.

4.3.1. Bevezetés az implicit Runge-Kutta típusú módszerekbe

Egy implicit Runge-Kutta típusú módszerek általános egylépéses módszer, mint általános egylépéses módszer a (4.92) szerint a következő alakban adható meg:

$$y_n = y_{n-1} + h\Phi(h, t_{n-1}, y_{n-1}, y_n), \tag{4.122}$$

ahol Φ a numerikus módszert meghatározó adott függvény,. Mint látni fogjuk, a Φ függvény meghatározása az implicit Runge-Kutta típusú módszerek esetén már jóval összetettebb feladat, mint az explicit módszerek esetében volt.

Tekintsünk néhány példát!

4.1. Példa *Határozzuk meg a*

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \tag{4.123}$$

Butcher-táblázatú módszerhez tartozó Φ függvényt!

A (4.123) egy egylépcsős Runge-Kutta típusú módszert határoz meg. Részletesen kiírva a következőt jelenti:

$$\begin{aligned} k_1 &= f(t_{n-1} + h, y_{n-1} + hk_1) \\ y_n &= y_{n-1} + hk_1. \end{aligned} \tag{4.124}$$

Tehát algoritmikus realizálása következőt jelenti: első lépésben megoldjuk a k_1 ismeretlenre az első egyenletet[6], majd a megoldást behelyettesítjük a második képletbe. Határozzuk meg a módszer Φ függvényét! Mivel a második képletből $hk_1 = y_n - y_{n-1}$, ezt behelyettesítve az első egyenletbe $k_1 = f(t_{n-1}+h, y_{n-1}+(y_n-y_{n-1})) = f(t_{n-1}+h, y_n)$. Ezért, ismét a második összefüggésből, $y_n = y_{n-1} + hf(t_{n-1} + h, y_n)$, azaz $\Phi(h, t_{n-1}, y_{n-1}, y_n) = f(t_{n-1} + h, y_n)$. Így a (4.123) Butcher-táblázatú módszer az jól ismert és már tárgyalt implicit Euler-módszert eredményezi. (Így ez a módszer elsőrendű.)

[6] A megoldáshoz valamely, már korábban ismertetett iterációs módszert (tipikusan a Newton-féle iterációt) alkalmazzuk.

4.2. Példa *Határozzuk meg a*

$$\begin{array}{c|c} 0.5 & 0.5 \\ \hline & 1 \end{array} \tag{4.125}$$

Butcher-táblázatú módszerhez tartozó Φ *függvényt!*

Ez szintén egy egylépcsős diagonálisan implicit Runge-Kutta típusú módszer, amely a következőt jelenti:

$$\begin{aligned} k_1 &= f(t_{n-1} + 0.5h, y_{n-1} + 0.5hk_1) \\ y_n &= y_{n-1} + hk_1. \end{aligned} \tag{4.126}$$

Határozzuk meg ennek a módszernek is a Φ függvényét! A második képletből $hk_1 = y_n - y_{n-1}$. Behelyettesítve az első egyenletbe $k_1 = f(t_{n-1}+0.5h, y_{n-1}+0.5(y_n - y_{n-1})) = f(t_{n-1}+0.5h, 0.5(y_{n-1}+y_n))$. Tehát $\Phi(h, t_{n-1}, y_{n-1}, y_n) = f(t_{n-1}+0.5h, 0.5(y_{n-1}+y_n))$ és a megfelelő egylépéses módszer

$$y_n = y_{n-1} + hf(t_{n-1} + 0.5h, 0.5(y_{n-1} + y_n)) \tag{4.127}$$

alakú. A (4.127) diagonálisan implicit Runge-Kutta típusú módszert *implicit középponti szabálynak* nevezzük. A módszer rendjét a

$$g_n = u(t_n) - u(t_{n-1}) - hf(t_{n-1} + 0.5h, 0.5(u(t_{n-1}) + u(t_n)))$$

kifejezés nagyságrendjének meghatározásával nyerjük. A szokásos sorbafejtéssel a következőket kapjuk. Az első kifejezés sorbafejtésével

$$u(t_n) - u(t_{n-1}) = hu'(t_{n-1}) + \frac{h^2}{2}u''(t_{n-1}) + \mathcal{O}(h^3).$$

Másrészt

$$\begin{aligned} &f(t_{n-1} + 0.5h, 0.5(u(t_{n-1}) + u(t_n))) = \\ &f(t_{n-1} + 0.5h, u(t_{n-1}) + 0.5hu'(t_{n-1}) + \mathcal{O}(h^2)) = f(t_{n-1}, u(t_{n-1}))+ \\ &0.5h\partial_1 f(t_{n-1}, u(t_{n-1})) + 0.5hu'(t_{n-1})\partial_2 f(t_{n-1}, u(t_{n-1})) + \mathcal{O}(h^2) = \\ &f(t_{n-1}, u(t_{n-1}))+ \\ &\frac{h}{2}[\partial_1 f(t_{n-1}, u(t_{n-1})) + f(t_{n-1}, u(t_{n-1}))\partial_2 f(t_{n-1}, u(t_{n-1}))] + \mathcal{O}(h^2). \end{aligned}$$

Behelyettesítve ezeket az értékeket g_n kifejezésébe, és figyelembe véve (3.4) második összefüggését, azaz az

$$u''(t) = \partial_1 f(t, u(t)) + \partial_2 f(t, u(t))u'(t) = \partial_1 f(t, u(t)) + \partial_2 f(t, u(t))f(t, u(t))$$

képletet, a $g_n = \mathcal{O}(h^3)$ nagyságrendet kapjuk. Tehát az implicit középponti szabály másodrendű.

4.3. Példa () *Határozzuk meg a*

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0.5 & 0.5 \\ \hline & 0.5 & 0.5 \end{array} \tag{4.128}$$

Butcher-táblázatú módszerhez tartozó Φ *függvényt!*

Ez egy kétlépcsős implicit Runge-Kutta típusú módszer, amely a következőt jelenti:

$$\begin{aligned} k_1 &= f(t_{n-1}, y_{n-1}) \\ k_2 &= f(t_{n-1} + h, y_{n-1} + 0.5k_1 + 0.5k_2) \\ y_n &= y_{n-1} + 0.5hk_1 + 0.5hk_2. \end{aligned} \tag{4.129}$$

A harmadik képletből $0.5h(k_1+k_2) = y_n - y_{n-1}$. Ezt és az első összefüggést behelyettesítve a második egyenletbe, a $k_2 = f(t_{n-1}+h, y_{n-1}+(y_n-y_{n-1})) = f(t_{n-1}+h, y_n)$ összefüggést kapjuk. Tehát ezt a k_2 értéket, illetve az első összefüggésbeli k_1 értéket behelyettesítve a harmadik egyenletbe, $\Phi(h, t_{n-1}, y_{n-1}, y_n) = 0.5[f(t_{n-1}, y_{n-1}) + f(t_{n-1}+h, y_n)]$. Ezért a (4.128) Butcher-táblázat az implicit trapéz-módszert jelenti, amely másodrendű módszer.

A 4.2. példa azt mutatja, hogy az implicit Runge-Kutta típusú módszerek esetén a pontosság rendszáma (p) meghaladhatja a lépcsők számát (s), hiszen (4.125) egy másodrendben pontos, egylépcsős módszert definiál.

Mielőtt áttérünk az implicit Runge-Kutta típusú módszerek vizsgálatára, egy rövid összefoglalót adunk a szükséges és felhasznált matematikai háttérről.

Először az interpolációs formulákon alapuló numerikus integráló képletek alapjait ismertetjük.

Legyen $g(t)$ egy adott, az $[a, b]$ intervallumon értelmezett, megfelelően sima függvény. Célunk az $\int_a^b g(t)dt$ integrál közelítő értékének kiszámolása. Legyenek $t_1, t_2, \ldots, t_s$ az $[a, b]$ intervallum s-darab, különböző pontjai. (Mindig feltesszük, hogy indexelésük növekvő sorrendű, azaz $a \le t_1 < t_2 < \ldots < t_s \le b$.) A (t_k), $k = 1, 2, \ldots, s$ pontokat *integrációs alappontoknak* nevezzük. Ekkor az alappontokra fektetett *Lagrange-féle interpolációs polinom*[7]

$$g(t) \approx g_h(t) = \sum_{j=1}^{s} g(t_j) L_j(t),$$

ahol $L_j(t)$ a Lagrange-féle approximációs polinom, amelynek alakja

$$L_j(t) = \prod_{\substack{1 \le i \le s \\ i \ne j}} \frac{t - t_i}{t_j - t_i}.$$

[7]Joseph-Louis Lagrange (1736-1813) (aki Giuseppe Luigi Lagrancia néven született), olasz - francia matematikus és csillagász. Kiemelkedő tudományos munkássága a matematikai analízis szinte valamennyi ágához, továbbá a számelmélethez, a klasszikus és az égi mechanikához kapcsolódik.

Ekkor g_h egy $s-1$-ed fokú polinom (jelölésben: $g_h \in \mathcal{P}_{s-1}$), és $g \in \mathcal{P}_{s-1}$ esetén $g = g_h$, hiszen az $s-1$-ed rendű polinomokra az approximáció pontos.

A numerikus integráló képletet az $\int_a^b g(t)dt \approx \int_a^b g_h(t)dt$ közelítésből állítjuk elő. Ekkor tehát

$$\int_a^b g(t)dt \approx \int_a^b g_h(t)dt = \int_a^b \left(\sum_{j=1}^{s} g(t_j) L_j(t) \right) dt.$$

Így a közelítés

$$\int_a^b g(t)dt \approx \sum_{j=1}^{s} w_j g(t_j) \tag{4.130}$$

alakú, ahol $w_j = \int_a^b L_j(t)dt$. Megjegyezzük, hogy az alappontok optimális megválasztásával (az ún. Gauss-féle alappontokkal) a numerikus interpolációs kvadratúraformula rendje növelhető.

A Lagrange-féle approximáció tehát olyan függvényközelítő eljárás, amikor egy s számú pontban ismert függvényre fektetünk olyan polinomot, amely átmegy ezeken az adott pontokon. Egy másik feladat, amikor egyetlen adott pontban, az $l = 0, 1, \ldots, s-1$ deriválási rendig bezárólag adott értékekre illeszkedő polinomot határozunk meg. Ez az ún. Taylor-féle interpoláció. (Mindkét megközelítést magába foglalja a Hermite-féle interpoláció.)

Foglalkozzunk a függvényközelítés egy másik módszerével, az ún. *Padé-típusú approximációval*.[8] Ez a Taylor-módszer természetes általánosítása: egy adott $g(z)$, megfelelően sima függvényhez olyan racionális törtfüggvényt keresünk, amely egy adott pontban (általában a $z = 0$ pontban) maximális rendben illeszkedik a függvényhez.[9] Tehát az approximációt

$$g(z) \approx R_{k,m}(z) := \frac{P_k(z)}{Q_m(z)} \tag{4.131}$$

alakban keressük, ahol P_k, Q_m valamely k- illetve m-ed fokú polinomok, és az egyértelműség miatt feltesszük, hogy $Q(0) = 1$. (Az $R_{k,m}$ approximációs függvény ezen alakja nyilván általánosítása a polinomoknak, hiszen a $Q_m(z) = 1$ megválasztás esetén $R_{k,0} = P_k \in \mathcal{P}_k$ polinom alakú a közelítés.) Mivel feltételeink mellett az $R_{k,m}(z)$ racionális törtfüggvénynek $k+m+1$ szabad paramétere van, ezért az elérhető maximális rend $p = k + m$. Ezt a

$$g^{(l)}(0) = R_{k,m}^{(l)}(0), \quad l = 0, 1, \ldots, p \tag{4.132}$$

feltétel biztosítja.

[8] Henri Eugene Padé (1863-1953) francia matematikus, aki elsődlegesen a racionális függvényközelítés területén ért el eredményeket.

[9] Megjegyezzük, hogy a racionális approximáció ötlete Ferdinand Georg Frobenius (1849-1917) kiemelkedő német matematikustól ered.

Térjünk át egy általános alakú operátoregyenlet lehetséges megoldására, ezen belül a *kollokációs módszer* rövid bevezetésére. Legyenek $L, B : \mathbf{X} \to \mathbf{Y}$ adott operátorok, ahol B lineáris. Keressük az adott $g \in \mathbf{Y}$ mellett az $Lu = g$ olyan megoldását, amelyre $Bu = 0$. (Ez utóbbi általában valamilyen peremfeltételt jelent.) Tehát a megoldást valamely $\mathbf{C} \subset \mathbf{X}$ vektor-altéren keressük, amelynek elemei az $x \in \Omega$ független változótól, illetve m darab paramétertől függnek, és kielégítik a peremfeltételt, azaz

$$\mathbf{C} = \{w \in \mathbf{X} : \ w = w(x; p_1, p_2, \ldots, p_m), \ Bw = 0\}.$$

Tipikusan a w elem $w = \sum_{k=1}^{m} p_k w_k(x)$ alakú, ahol $w_k(x)$ egy bázis $\mathbf{C}$-ben.

Célunk, hogy valamilyen elv alapján a paraméterek jó megválasztásával egy jó közelítő megoldását adjuk az eredeti operátoregyenletnek. Ezt általában az

$$E(x; p_1, p_2, \ldots, p_m) = (\mathbf{L}w)(x; p_1, p_2, \ldots, p_m) - g(x)$$

kifejezés kicsivé tételével szoktuk meghatározni.

Több módon is megválaszthatjuk az "optimális" paramétereket. A kollokációs módszer lényege, hogy rögzítünk m darab pontot, és megköveteljük, hogy ezekben a pontokban E legyen nulla, azaz teljesüljön az

$$E(x_i; p_1, p_2, \ldots, p_m) = 0, \quad i = 1, 2, \ldots, m$$

feltétel, ahol az x_i pontok az ún. kollokációs alappontok.

4.4. Megjegyzés Általában az m darab pontra kitűzött kollokáció esetén w egy $m-1$-ed fokú polinom, azaz

$$w(x; p_1, p_2, \ldots, p_m) = \sum_{k=1}^{m} p_k x^{k-1}.$$

Ekkor a cél a polinom együtthatóinak megfelelő megválasztása. Vegyük észre, hogy a Lagrange-féle interpoláció értelmezhető kollokációs módszerként is, az $\mathbf{L}$ =identitás megválasztással.

Léteznek más módszerek is a p_k paraméterek meghatározására. Például bevezethetjük a

$$G(p_1, p_2, \ldots, p_m) = \int_\Omega |E(\zeta; p_1, p_2, \ldots, p_m)|^2 d\zeta$$

függvényt, és azt minimalizáljuk az m darab változójában. (Ez az ún. legkisebb négyzetek módszere.) Egy másik ötlet a Galjorkin-módszer, amikor a paramétereket az

$$\int_\Omega E(\zeta; p_1, p_2, \ldots, p_m) \cdot \psi_i(\zeta) d\zeta = 0, \quad i = 1, 2, \ldots, m$$

feltételből határozzuk meg, ahol $\{\psi_i(\zeta)\}$ valamely jó tulajdonságú függvényrendszer. (Például, $\psi_i(\zeta) = w_i(\zeta)$.)

4.3.2. A kollokációs módszeren alapuló implicit Runge-Kutta típusú módszerek

Legyen $u(t)$ az

$$u' = f(t, u) \tag{4.133}$$

egyenlet megoldás, azaz az

$$u'(t) = f(t, u(t)) \tag{4.134}$$

egyenlőséget kielégítő függvény. A (4.134) tulajdonság felírható az alábbi operátoros alakban: jelölje $L : C^1 \to C$ az $(Lw)(t) = w'(t) - f(t, w(t))$ operátort. Az egylépéses numerikus módszerek megadása azt jelenti, hogy megadjuk azt az algoritmust, amely szerint a $t = t_{n-1}$ pontbeli y_{n-1} közelítésből a $t = t_n$ pontbeli y_n közelítés meghatározható.

Ezt az algoritmust a kollokációs módszer segítségével adjuk meg.

Legyenek $t_{n-1} \leq t_1 < t_2 < \ldots < t_s \leq t_n$ a $[t_{n-1}, t_n]$ intervallumon definiált s darab pont. Keressük az ezen pontokra felépített kollokációs módszerrel az olyan $w(t) \in \mathcal{P}_s$ s-ed fokú polinomot, amelyre

$$w(t_{n-1}) = y_{n-1}. \tag{4.135}$$

Továbbá a kollokációs módszer alapján

$$(Lw)(t_i) = w'(t_i) - (Lw)(t_i) = 0, \quad i = 1, 2, \ldots, s. \tag{4.136}$$

Tehát feladatunk olyan $w(t) \in \mathcal{P}_s$ meghatározása, amelyre

a. $w(t_{n-1}) = y_{n-1}$;

b. $w'(t_i) = f(t_i, w(t_i)), \quad i = 1, 2, \ldots, s.$ (Vegyük észre, hogy a. és b. $s + 1$ feltételt jelent az $s+1$ szabad paraméterrel rendelkező ismeretlen s-ed fokú polinomra, azaz egyértelműen meghatározható a w polinom.)

c. Meghatározva a w polinomot, legyen

$$y_n = w(t_n). \tag{4.137}$$

A kollokációs alappontok felírhatók

$$t_i = t_{n-1} + c_i h, \quad i = 1, 2, \ldots, s, \quad h = t_n - t_{n-1} \tag{4.138}$$

alakban, ahol $0 \leq c_1 < c_2 < \cdots < c_s \leq 1$.

Jelölje

$$K_i = w'(t_i), \quad i = 1, 2, \ldots, s. \tag{4.139}$$

Mivel $w'(t)$ egy $s-1$-ed fokú polinom, ezért az s darab kollokációs alappontra felépített Lagrange-interpolációja pontosan előállítja, azaz

$$w'(t) = \sum_{j=1}^{s} K_j L_j(t), \tag{4.140}$$

ahol

$$L_j(t) = \prod_{\substack{k=1 \\ k\neq j}}^{s} \frac{t - t_k}{t_j - t_k}.$$

Integráljuk a (4.140) egyenlőséget a $[t_{n-1}, t_i]$ intervallumon! Ekkor

$$\int_{t_{n-1}}^{t_i} w'(t)dt = \sum_{j=1}^{s} K_j \int_{t_{n-1}}^{t_i} L_j(t)dt, \tag{4.141}$$

azaz

$$w(t_i) - w(t_{n-1}) = \sum_{j=1}^{s} \tilde{a}_{ij} K_j, \tag{4.142}$$

ahol

$$\tilde{a}_{ij} = \int_{t_{n-1}}^{t_i} L_j(t)dt = \int_{t_{n-1}}^{t_i} \left(\prod_{\substack{k=1 \\ k\neq j}}^{s} \frac{t - t_k}{t_j - t_k} \right) dt.$$

Innen a $t = t_{n-1} + \tau h$ transzformációval és a (4.138) jelöléssel

$$\tilde{a}_{ij} = \int_0^{c_i} \left(\prod_{\substack{k=1 \\ k\neq j}}^{s} \frac{(t_{n-1} + \tau h) - (t_{n-1} + c_k h)}{(t_{n-1} + c_j h) - (t_{n-1} + c_k h)} \right) h d\tau = h a_{ij}, \tag{4.143}$$

ahol

$$a_{ij} = \int_0^{c_i} \left(\prod_{\substack{k=1 \\ k\neq j}}^{s} \frac{\tau - c_k}{c_j - c_k} \right) d\tau = \int_0^{c_i} L_j(\tau)d\tau. \tag{4.144}$$

Ezért a (4.135) figyelembevételével a (4.142) és a (4.143) összefüggések alapján

$$w(t_i) = y_{n-1} + h \sum_{j=1}^{s} a_{ij} K_j, \tag{4.145}$$

ahol a_{ij} értékét (4.144) alapján határozhatjuk meg.

Integráljuk most a (4.140) egyenlőséget a $[t_{n-1}, t_n]$ intervallumon! Ekkor

$$\int_{t_{n-1}}^{t_n} w'(t)dt = \sum_{i=1}^{s} K_i \int_{t_{n-1}}^{t_n} L_i(t)dt, \tag{4.146}$$

azaz

$$w(t_n) - w(t_{n-1}) = h \sum_{j=1}^{s} b_i K_i, \tag{4.147}$$

ahol

$$b_i = \int_0^1 \left(\prod_{\substack{k=1 \\ k \neq j}}^{s} \frac{\tau - c_k}{c_i - c_k} \right) d\tau = \int_0^1 L_i(\tau)d\tau. \tag{4.148}$$

Ezért, figyelembe véve a 4.139 jelölést és a (4.145) összefüggést, érvényes a

$$K_i = w'(t_i) = f(t_i, w(t_i)) = f(t_i, y_{n-1} + h \sum_{j=1}^{s} a_{ij} K_j) \tag{4.149}$$

egyenlőség. Másrészt, (4.147), (4.135) és (4.137) alapján

$$y_n = y_{n-1} + h \sum_{j=1}^{s} b_i K_i. \tag{4.150}$$

Vegyük észre, hogy (4.149) és (4.150) éppen egy implicit Runge-Kutta típusú módszert definiál.

4.5. Példa *Határozzuk meg a* $[0, 1]$ *intervallumon (tehát* $h = 1$ *esetén) azt a implicit Runge-Kutta típusú módszert, amelyet a kollokációs módszerrel, az intervallum két végpontjának alappontként való alkalmazásával kapunk!*

Példánkban $s = 2$, $t_1 = 0$ és $t_2 = 1$, azaz $c_1 = 0$ és $c_2 = 1$. Ezért $L_1(t) = 1 - t$, $L_2(t) = t$. Mindezek alapján az implicit Runge-Kutta típusú módszer paraméterei a következők.

$$\begin{aligned} a_{11} &= \int_0^0 L_1(\tau)d\tau = 0, \quad a_{12} = \int_0^0 L_2(\tau)d\tau = 0, \\ a_{21} &= \int_0^1 L_1(\tau)d\tau = 0.5, \quad a_{22} = \int_0^1 L_2(\tau)d\tau = 0.5, \\ b_1 &= \int_0^1 L_1(\tau)d\tau = 0.5, \quad b_2 = \int_0^1 L_2(\tau)d\tau = 0.5. \end{aligned} \tag{4.151}$$

Így a megfelelő implicit Runge-Kutta típusú módszer Butcher-táblázata

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0.5 & 0.5 \\ \hline & 0.5 & 0.5 \end{array} \tag{4.152}$$

alakú, tehát a generált módszer a 4.3. példában már tárgyalt implicit trapéz módszert jelenti.

Az implicit Runge-Kutta típusú módszerek kollokációs módszerként való előállítása lehetőséget ad a módszer hibabecslésére is. Ugyanis a módszer lokális csonkolási hibája az egy lépésnél (a t_{n-1} pontból a t_n pontba a numerikus integráló formula hibájával egyenlő. Mint ismeretes (ld. pl. [5]) az s alappontos kvadratúraformula akkor és csak akkor pontos minden legfeljebb $s-1$-edfokú polinomra, ha interpolációs kvadratúraformula. Mivel a deriváltfüggvényre alkalmazzuk kvadratúrát, ezért s pont esetén a lokális csonkolási hiba rendje a függvényre $\mathcal{O}(h^{s+1})$, azaz az implicit Runge-Kutta típusú módszer pontossága *tetszőleges alappontok* esetén $p = s$. Ugyanakkor speciális alappontmegválasztással a rend növelhető! Például, ha alappontként a Gauss-féle pontokat választjuk, akkor $p = 2s$ rend érhető el. Mivel ennél magasabb kvadratúra nem adható meg, a $p = 2s$ pontosságú módszerek a legmagasabb rendűek, ezért az ilyen módszereket *maximális rendű implicit Runge-Kutta típusú módszereknek* nevezzük.[10]

A kollokációs alappontok speciális megválasztásával elért magasabb rendű implicit Runge-Kutta típusú módszerek közül ismertetünk néhányat.

A *Radau-módszer* olyan alappontmegválasztású módszer, amikor az intervallum egyik végpontja hozzátartozik az alappontokhoz. (Tehát $c_1 = 0$ vagy $c_s = 1$.) A módszer pontossága $p = 2s - 1$.[11]

4.6. Példa *Az implicit Euler-módszer egylépcsős Radau-módszer.*

Ez könnyen látható, hiszen a jobb oldali végpont kollokációs alappont.

4.7. Példa *Legyen a módszer Butcher-táblázata a következő:*

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array} \tag{4.153}$$

Nyilvánvalóan ez az implicit Runge-Kutta típusú módszer szintén Radau-módszer: ebben az esetben is a jobb oldali végpont szerepel az alappontok között. Erre a módszerre $s = 2$

[10] Az irodalomban az ilyen módszereket s-lépcsős Gauss-Legendre-módszernek is szokás nevezni.

[11] Az irodalomban szokásos a módszerekre a Radau IA és Radau IIA megkülönböztetés is: az első esetben a bal oldali végpont, a második esetben pedig a jobb oldali végpont tartozik az alappontokhoz.

és $p = 3$. (A rend a Runge-Kutta típusú módszerek rendjének már ismertetett (4.82) feltételeiből közvetlenül ellenőrizhető.)

A *Lobatto-módszer* olyan alappont-megválasztású módszer, amikor az intervallum mindkét végpontja hozzátartozik az alappontokhoz. (Tehát $c_1 = 0$ és $c_s = 1$.) A módszer pontossága $p = 2s - 2$.[12]

Megjegyezzük, hogy a (4.152) Butcher-táblázatú implicit trapéz szabály Lobatto-módszer, amelyre $p = s = 2$.

4.8. Példa *Az alábbi Butcher-táblázatú implicit Runge-Kutta típusú módszer egy jól ismert Lobatto-módszer:*

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array} \tag{4.154}$$

Erre a módszerre $s = 3$ és $p = 4$.

A legnagyobb pontosságot az eredményezi, amikor alappontként a Gauss-féle alappontokat választjuk. Azt így nyert módszert *Gauss-Legendre–módszernek* nevezzük. Ezek pontossága $p = 2s$. A középponti szabály mellett az alábbi módszer elterjedt.

4.9. Példa *Tekintsük az alábbi Butcher-táblázatú implicit Runge-Kutta típusú módszert:*

$$\begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \tag{4.155}$$

Ez egy Gauss-Legendre-típusú módszer, amelyre $s = 2$ és $p = 4$.

Ezen pont befejezéseként térjünk vissza a Runge-Kutta típusú módszerek rendjének vizsgálatához merev feladatok esetén. Mint beláttuk, egy numerikus módszer p-ed rendűsége azt jelenti, hogy megfelelően kis lépésköz (h) esetén a globális hiba h^p rendűen, azaz, ha h a differenciálegyenlet időskálájához képest megfelelően kicsi, akkor a hiba const$\cdot h^p$ alakú.

4.10. Példa *Tekintsük az*

$$u'(t) = \lambda(u(t) - g(t)), \quad t \in (0, 1] \tag{4.156}$$

feladatot, ahol λ egy külső paramétertől függő mennyiség (pl. a térbeli diszkretizációtól függ), és

$$0 < -\frac{1}{\operatorname{Re}\lambda} \ll h \ll 1. \tag{4.157}$$

[12] Megjegyezzük, ha az $\mathbf{A}, \mathbf{b}$ és $\mathbf{c}$ értékeire néhány további speciális feltételt is kikötünk, akkor szokásos Lobatto IIIA, IIIB és IIIC típusú módszerekről is beszélni. Ezek pontossága szintén $p = 2s - 2$.

Azt is láttuk, hogy a kollokációs implicit Runge-Kutta típusú módszerek hibájának rendjét két tényező határozza meg:

- a (4.156) jobb oldalán az ismeretlen $u(t)$ megoldásfüggvény interpolációjának rendje,
- kvadratúra alkalmazása a differenciálegyenlet kiintegrálására.

Például, tekintsük a szokásos $u' = cu$ tesztfeladatot, ahol c egy adott állandó. Ekkor a differenciálegyenlet $[t_{n-1}, t_n]$ intervallumon való szokásos integrálása után az

$$u(t_n) - u(t_{n-1}) = c \int_{t_{n-1}}^{t_n} u(t)dt$$

egyenlőséget kapjuk. Ha a nulladrendű $u(t) \approx u(t_{n-1})$ interpolációt alkalmazzuk, majd utána az interpolációs kvadratúraformulát, akkor az elsőrendű explicit Euler-módszert kapjuk.

Amint ez a példa is mutatja (de a numerikus integrálás elméletéből is ismeretes), a kvadratúrák alkalmazása egy renddel növeli az approximáció rendjét. Ha visszatérünk a (4.156) feladatra, és azt átírjuk az

$$\frac{1}{\lambda}u'(t) = u(t) - g(t), \quad t \in (0, 1] \tag{4.158}$$

alakra, akkor a (4.157) reláció alapján látható, hogy (4.158) bal oldala gyakorlatilag eltűnik. Ezért a (4.158) feladat tulajdonképpen az

$$u - g(t) = 0, \quad t \in (0, 1] \tag{4.159}$$

feladat megoldását jelenti. Ez viszont a szokásos interpolációs feladathoz vezet, és így elmarad a numerikus integrálási rész. Mindez azt jelenti, hogy az ilyen feladatokra a Runge-Kutta típusú módszerek rendje nem az elmélet szerinti rendet, hanem annál közelítőleg egyel kisebb rendet eredményez csak. Ezt a jelenséget a szakirodalomban *rendcsökkenésnek* szokásos nevezni.

4.3.3. Egylépéses módszerek a tesztfeladaton

Az eddigiekben több numerikus módszert ismertettünk, amelyek a pontosságukban (rendjükben), illetve a megoldás kiszámításának módjában (explicit/implicit) különböztek egymástól. Ebben a részben egy modellfeladatra alkalmazva a módszereinket újabb tulajdonságokat állapítunk meg.
Oldjuk meg rögzített $\lambda < 0$ mellett az

$$u' = \lambda u, \quad u(0) = 1 \tag{4.160}$$

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

4.3. táblázat. A különböző λ értékekkel kitűzött tesztfeladat numerikus megoldásának hibája maximumnormában

ún. *tesztfeladatot* a $[0, 1]$ intervallumon az explicit és implicit Euler-módszerekkel! (A (4.160) feladat pontos megoldása $u(t) = \exp(\lambda t)$.) 4.3 táblázat tartalmazza eredményeinket.

Figyeljük meg, hogy $\lambda = -9$ mellett mindkét módszer az elméletnek megfelelő módon viselkedik (azaz mindegyik h érték mellett a hiba elsőrendű). Ugyanakkor a $\lambda = -99$ és a $\lambda = -999$ esetekben ez már nem így van: a kezdeti h megválasztások mellett az explicit Euler-módszer nem közelíti a megoldást, az implicit Euler-módszer viszont jó közelítést ad. A h paraméter további csökkentésével viszont már mindkét módszer az elméletnek megfelelően viselkedik.[13] Mindebből arra következtethetünk, hogy ezekre a feladatokra az implicit Euler-módszer h megválasztásától függetlenül mindig jól viselkedik, az explicit Euler-módszer viszont csak valamely $h_0 > 0$ melletti $h < h_0$ feltétel mellett alkalmazható. Ez utóbbi viszont problémát jelent: ha a negatív λ értékét tovább csökkentjük, akkor már csak olyan kis h_0 értékek mellett működik az explicit Euler-módszer, amelynél

1. h_0 közel van a gépi epszilonhoz, ezért a számítógépes realizálás eleve nem lehetséges,

2. ha h_0 nagyobb ugyan a gépi epszilonnál, de nagyon kicsi, akkor egy rögzített $t^\star$ időrétegen való $y_{n_{t^\star}}$ közelítés meghatározásához rendkívül sok lépés ($n_{t^\star} \approx t^\star/h_0$) végrehajtása szükséges. Ez egyrészt a számítások idejének megnövekedését, másrészt pedig a számítások során fellépő hibák felhalmozódásának a lehetőségét eredményezi.

A jelenség oka a következő. A (4.160) tesztfeladat megoldása az explicit Euler-módszerrel az

$$y_{i+1} = (1 + h\lambda)y_i, \quad i = 0, 1, \ldots, \quad y_0 = 1,$$

[13] Magyarázatra szorul, hogy a táblázat ezen két oszlopában h csökkentése esetén *miért növekszik* a hiba bizonyos helyeken? Ennek magyarázata, hogy a hibát a maximumnormában mérjük a rácsháló pontjaiban, ezért h csökkentésével új rácspontok kerülnek be. Ezekben a pontokban a durvább rácshálón nem mértük a hibát, viszont a finomabban igen. Így, ha a megoldás nagyon meredek (azaz a derivált értéke nagy), akkor előfordul, hogy ezekben az újonnan hozzávett pontokban még kisebb h esetén is nagyobb a hiba, mint a nagyobb h melletti pontok bármelyikében.

az implicit Euler-módszerrel pedig az

$$y_{i+1} = \frac{1}{1-h\lambda} y_i, \quad i = 0, 1, \ldots, \quad y_0 = 1,$$

egylépéses iterációkat jelenti. Egységesen felírva, az egylépéses módszereknek a tesztfeladatra történő alkalmazása egy

$$y_{i+1} = R(h\lambda) y_i \tag{4.161}$$

iterációt jelent, ahol $R(h\lambda)$ az alkalmazott numerikus módszer által meghatározott függvény. Mivel a tesztfeladat $u(t) = \exp(\lambda t)$ megoldására

$$u(t_{i+1}) = \exp(h\lambda) u(t_i), \tag{4.162}$$

ezért az alkalmazott módszert az jellemzi, hogy az $R(z)$ függvény milyen jól közelíti az $\exp(z)$ függvényt.[14]

4.11. Megjegyzés Vegyük észre, hogy az $\exp(z) - R(z)$ eltérés a numerikus módszer képlethibáját jellemzi a tesztfeladaton! Ugyanis a képlethiba (3.60) definíciója, a (4.161) és a (4.162) összefüggések alapján

$$g_i(h) = u(t_{i+1}) - R(h\lambda) u(t_i) = (\exp(h\lambda) - R(h\lambda)) u(t_i).$$

Mivel az $u(t)$ megoldás korlátos, ezért a konzisztencia rendje meghatározható az $(\exp(h\lambda) - R(h\lambda))$ rendjével, azaz p-ed rendű konzisztencia esetén $\exp(h\lambda) - R(h\lambda) = \mathcal{O}(h^{p+1})$. Mivel az explicit Euler-módszer esetén $R_{EE}(z) = 1 + z$, az implicit Euler-módszer esetén $R_{IE}(z) = 1/(1-z)$, a trapéz-módszer esetén pedig $R_{CN}(z) = (1 + z/2)/(1 - z/2)$, ezért könnyen ellenőrizhetően $\exp(z) - R_{EE}(z) = \mathcal{O}(h^2)$, $\exp(z) - R_{IE}(z) = \mathcal{O}(h^2)$, $\exp(z) - R_{CN}(z) = \mathcal{O}(h^3)$, amely természetesen összhangban áll az ezen módszerek rendjére vonatkozó korábbi megállapításainkkal.

A tesztfeladaton tetszőleges $\lambda < 0$ esetén a megoldás korlátos. Ezért nyilvánvalóan csak azok a numerikus megoldások tudják a pontos megoldást approximálni, amelyekre az előállított numerikus megoldás szintén korlátos, azaz a (4.161) módszerben az

$$|R(h\lambda)| \leq 1 \tag{4.163}$$

feltétel teljesül. Mivel valós λ esetén

$$|R_{EE}(h\lambda)| \leq 1 \quad \Leftrightarrow h \leq 2/(-\lambda), \tag{4.164}$$

[14] Emlékeztetünk rá, hogy az előző 3 fejezetben ezt $R(z)$ egyváltozós függvényt a numerikus módszer *stabilitási függvényének* neveztük, és néhány egyszerűbb egylépéses módszerre részletesen megvizsgáltuk.

ezért az explicit Euler-módszerre a már említett $h \leq h_0$ feltételben $h_0 = 2/(-\lambda)$. Ugyanakkor az implicit Euler-módszer esetén

$$|R_{IE}(h\lambda)| \leq 1 \quad \text{minden } h > 0 \tag{4.165}$$

esetén. A (4.164) és a (4.165) összefüggések magyarázatot adnak arra, hogy a 4.3 táblázatban miért viselkednek ennyire eltérően az explicit és implicit Euler-módszerek bizonyos h értékek esetén.

Fontos megjegyeznünk, hogy valamely numerikus módszer konvergenciája a módszer $h \to 0$ melletti viselkedését jellemzi, azaz csak azt garantálja, hogy *elegendően kis h esetén* a numerikus megoldás közel kerül a pontos megoldáshoz. Ugyanakkor nem ad információt a megoldásról valamely *rögzített rácshálón*. Ezért kiemelkedően fontosak azok az $R(z)$ stabilitási függvénnyel rendelkező numerikus módszerek, amelyek tetszőleges rögzített rácshálón is jól követik a pontos megoldást.

4.12. Definíció *Azon $z \in \mathbb{C}$ komplex számok halmazát, amelyekre az*

$$|R(z)| \leq 1 \tag{4.166}$$

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

Az A-stabilitás tehát azt jelenti, hogy minden olyan $z = a + ib$ komplex számra, amelyre $a < 0$, érvényes az $|R(z)| \leq 1$ egyenlőtlenség. Könnyen megmutatható, hogy az implicit Euler-módszer abszolút stabil. (Az explicit Euler-módszer nyilván nem, hiszen, mint láttuk, (4.166) már az $\mathbb{R}^- \subset \mathbb{C}^-$ halmazon sem igaz.)

4.13. Megjegyzés Tekintsük a trapéz-módszert, amely felírható

$$y_n = R_{CN}(h\lambda)y_{n-1} = \frac{1 + \lambda h/2}{1 - \lambda h/2} y_{n-1} = \frac{2 + \lambda h}{2 - \lambda h} y_{n-1}$$

alakban. Mint megmutattuk, a módszer másodrendű, és könnyen látható tetszőleges $h > 0$ esetén minden valós $\lambda < 0$ esetén $|R_{CN}(h\lambda)| \leq 1$. (Megmutatható, hogy abszolút stabil is.) Ugyanakkor a tesztfeladaton tetszőleges $h > 0$ mellett mégsem viselkedik jól a módszer. Ugyanis $h > 2/(-\lambda)$ esetén $R_{CN}(h\lambda) \in (-1, 0)$, ezért az ilyen rácshálókon az y_i értékei lépésenként előjelet váltanak, azaz bár abszolút értékben csökkennek, de emellett oszcillálnak is, ami ellentmond a pontos megoldás szigorú monoton csökkenésének.

4.3.4. Az implicit Runge-Kutta típusú módszerek abszolút stabilitása

Ebben a szakaszban megvizsgáljuk az implicit Runge-Kutta típusú módszerek abszolút stabilitását. Megmutatjuk, hogy az explicit Runge-Kutta típusú módszerrekkel ellentétben léteznek A-stabil implicit Runge-Kutta típusú módszerek, és megadunk néhány ilyen módszert.

Az A-stabilitás vizsgálatához határozzuk meg a módszerek stabilitási tartományát, azaz vizsgáljuk az implicit Runge-Kutta típusú módszer viselkedését az

$$u' = \lambda u, \quad t > 0 \tag{4.167}$$

tesztegyenleten. A stabilitási tartomány felírásához első lépésben a tesztfeladatra alkalmazott implicit Runge-Kutta típusú módszerhez tartozó $R(z)$ stabilitási függvény meghatározása szükséges.

Emlékeztetünk, hogy amikor a (4.63)-(4.64) Runge-Kutta típusú módszert alkalmazzuk a (4.167) tesztegyenletre, akkor a (4.63) alapján, a korábbi jelöléseink megtartásával érvényes az

$$\mathbf{Y} = y_{n-1}\mathbf{e} + h\mathbf{A}\lambda\mathbf{Y} = y_{n-1}\mathbf{e} + z\mathbf{A}\mathbf{Y} \tag{4.168}$$

összefüggés, ahol $z = h\lambda$. Innen

$$(\mathbf{I} - z\mathbf{A})\mathbf{Y} = y_{n-1}\mathbf{e}. \tag{4.169}$$

Mint korábban, most is megfelelően kis z (azaz h) esetén az $\mathbf{I} - z\mathbf{A}$ mátrix invertálható, és inverze előállítható a Neumann-sora segítségével. Tehát

$$\mathbf{Y} = (\mathbf{I} - z\mathbf{A})^{-1}\mathbf{e}y_{n-1} \tag{4.170}$$

Másrészt, a (4.64) alapján

$$\begin{aligned} y_n = y_{n-1} + h\sum_{i=1}^{s} b_i\lambda Y_i = y_{n-1} + z\mathbf{b}^\top\mathbf{Y} = \\ y_{n-1} + z\mathbf{b}^\top(\mathbf{I} - z\mathbf{A})^{-1}\mathbf{e}y_{n-1} = \left[1 + z\mathbf{b}^\top(\mathbf{I} - z\mathbf{A})^{-1}\mathbf{e}\right] y_{n-1}. \end{aligned} \tag{4.171}$$

Tehát az implicit Runge-Kutta típusú módszer stabilitási függvénye

$$R(z) = 1 + z\mathbf{b}^\top(\mathbf{I} - z\mathbf{A})^{-1}\mathbf{e}.$$

Feladatunk az

$$S = \{z \in \mathbb{C}, |R(z) \leq 1\} \tag{4.172}$$

stabilitási tartomány meghatározása.

4.14. Tétel *Egy tetszőleges implicit Runge-Kutta típusú módszer $R(z)$ stabilitási függvénye felírható*

$$R(z) = \frac{\det\left(\boldsymbol{I} - z\boldsymbol{A} + z\boldsymbol{e}\cdot\boldsymbol{b}^{\top}\right)}{\det\left(\boldsymbol{I} - z\boldsymbol{A}\right)} \tag{4.173}$$

alakban.

Bizonyítás. A (4.168) összefüggésből érvényes az

$$Y_i - z\sum_{j=1}^{s} a_{ij}Y_j = y_{n-1}, \quad i = 1, 2, \ldots, s \tag{4.174}$$

egyenlőség. Másrészt, a (4.171) összefüggésből

$$y_n - z\sum_{j=1}^{s} b_j Y_j = y_{n-1}. \tag{4.175}$$

Ekkor (4.174)-(4.175) egy lineáris algebrai egyenletrendszert jelent az Y_1, Y_2, $\ldots$, Y_s, y_n $s+1$ darab ismeretlenre nézve. Jelölje $\mathbf{y} \in \mathbb{R}^{s+1}$ ezen ismeretlenek vektorát! Ekkor az egyenletrendszer felírható

$$\mathbf{B}\mathbf{y} = \mathbf{f}$$

mátrixos alakban, ahol a rendszer $\mathbf{f} \in \mathbb{R}^{s+1}$ jobb oldala $\mathbf{f} = y_{n-1}\mathbf{e}$ alakú, míg a rendszer $\mathbf{B} \in \mathbb{R}^{(s+1)\times(s+1)}$ együtthatómátrixa a következő alakú:

$$\mathbf{B} = \begin{pmatrix} 1 - za_{11} & -za_{12} & \ldots & -za_{1s} & 0 \\ -za_{21} & 1 - za_{22} & \ldots & -za_{2s} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -za_{s1} & -za_{2s} & \ldots & 1 - za_{ss} & 0 \\ -zb_1 & -zb_2 & \ldots & -zb_s & 1 \end{pmatrix}. \tag{4.176}$$

Határozzuk meg ebből a rendszerből az utolsó, y_n ismeretlent a Cramer-szabállyal! Ekkor

$$y_n = \frac{\det \mathbf{B}_f}{\det \mathbf{B}}, \tag{4.177}$$

ahol

$$\mathbf{B}_f = \begin{pmatrix} 1 - za_{11} & -za_{12} & \ldots & -za_{1s} & y_{n-1} \\ -za_{21} & 1 - za_{22} & \ldots & -za_{2s} & y_{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -za_{s1} & -za_{2s} & \ldots & 1 - za_{ss} & y_{n-1} \\ -zb_1 & -zb_2 & \ldots & -zb_s & y_{n-1} \end{pmatrix}. \tag{4.178}$$

Könnyen láthatóan $\det \mathbf{B}$ értékét az utolsó oszlop szerinti kifejtéssel közvetlenül meghatározhatjuk, és ekkor

$$\det \mathbf{B} = \det(\mathbf{I} - z\mathbf{A}). \tag{4.179}$$

Ezért $\det \mathbf{B}_f$ értékének meghatározásához vonjuk ki az utolsó sorát rendre a felette lévő sorokból! Ezután az utolsó oszlop szerinti kifejtéssel már látható, hogy

$$\det \mathbf{B}_f = y_{n-1} \det(\mathbf{I} - z\mathbf{A} - +\mathbf{e} \cdot \mathbf{b}^\top). \tag{4.180}$$

A (4.179) és a (4.180) összefüggések (4.177) képletbe való behelyettesítése a tétel állítását eredményezi. □

4.15. Megjegyzés Vegyük észre, hogy a számlálóban lévő $\mathbf{e} \cdot \mathbf{b}^\top$ a két vektor diadikus szorzata, azaz egy $\mathbb{R}^{s \times s}$ méretű mátrix.

A tételből következik, hogy egy általános Runge-Kutta típusú módszer stabilitási függvénye két polinom hányadosa, azaz ún. racionális törtfüggvény. (Ugyanakkor, ha a Runge-Kutta típusú módszer explicit, akkor $\mathbf{A}$ szigorúan alsó háromszögű mátrix, ezért $\det \mathbf{B} = \det(\mathbf{I} - z\mathbf{A}) = 1$, azaz ebben az esetben $R(z)$ polinom. Ez természetesen megegyezik az explicit Runge-Kutta típusú módszer stabilitási függvényének már vizsgált tulajdonságával.

A tesztegyenlet megoldása $u(t) = e^{\lambda t}u(0)$. Ezért $u(t_n) = e^{\lambda h}u(t_{n-1})$, azaz $u(t_n) = e^z u(t_{n-1})$. Mivel az egylépéses módszerekre (így a Runge-Kutta típusú módszerekre is) $y_n = R(z)y_{n-1}$, ezért az $R(z)$ stabilitási függvénynek az e^z exponenciális függvényt kell a megfelelő módon approximálni. Mint láttuk, az explicit Runge-Kutta típusú módszerek esetén Taylor-típusú volt ez az approximáció, az implicit Runge-Kutta típusú módszerekre pedig Padé-típusú az approximáció. Tehát egy Runge-Kutta típusú módszer pe-d rendű pontosságát a (4.132) feltétellel tudjuk biztosítani, ahol $g(z) = e^z$, azaz itt a feltétel

$$\left(\frac{P_k}{Q_m}\right)^{(l)}(0) = 1, \quad l = 0, 1, \ldots, p \tag{4.181}$$

alakú.

4.16. Példa *Adjuk meg az exponenciális függvény maximális rendű $R_{1,1}$ Padé-típusú approximációját!*

Mivel $p = 1 + 1 = 2$, ezért a (4.181) alapján a feltétel

$$\left(\frac{P_1}{Q_1}\right)^{(l)}(0) = 1, \quad l = 0, 1, 2, \tag{4.182}$$

ahol $P_1(z) = a_0 + a_1 z$ és $Q_1(z) = 1 + b_1 z$. Az $R_{1,1}(z)$ approximációban szereplő három ismeretlenre (4.182) három feltételt jelent. Az így nyert egyenletrendszer megoldása,

mint az könnyen ellenőrizhető, az $a_0 = 1, a_1 = 1/2$ és $b_1 = -1/2$. Tehát a keresett másodrendű Padé-típusú approximáció

$$R_{1,1} = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}.$$

Így az ehhez tartozó implicit Runge-Kutta típusú módszer az implicit trapéz-módszer.

Kiszámíthatók az exponenciális függvény maximális rendű egyéb $R_{k,m}$ Padé-típusú approximációi is. A $k, m = 0, 1, 2$ esetekre az alábbi függvényeket kapjuk.

$$\begin{aligned} &R_{0,0} = 1,\ R_{1,0} = 1 + z,\ R_{2,0} = 1 + z + z^2/2, \\ &R_{0,1} = \frac{1}{1 - z},\ R_{1,1} = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}},\ R_{2,1} = \frac{1 + \frac{2}{3}z + \frac{1}{6}z^2}{1 - \frac{1}{3}z}, \\ &R_{0,2} = \frac{1}{1 - z + \frac{z^2}{2}},\ R_{1,2} = \frac{1 + \frac{z}{3}}{1 - \frac{2}{3}z + \frac{z^2}{6}},\ R_{2,2} == \frac{1 + \frac{z}{2} + \frac{z^2}{12}}{1 - \frac{1}{2}z + \frac{z^2}{12}}. \end{aligned} \tag{4.183}$$

Mikor lesznek A-stabilak a megfelelő Runge-Kutta típusú módszerek?

4.17. Definíció *Egy $r(z)$ racionális törtfüggvényt A-megengedett függvénynek nevezünk, ha minden $z \in \mathbb{C}^-$ esetén $|r(z)| \leq 1$.*

Tehát azok a Runge-Kutta típusú módszerek lesznek A-stabilak, amelyek $R(z)$ stabilitási függvényei A-megengedettek. Ezért fontos az alábbi állítás, amely a rendkorlátot jelenti az implicit Runge-Kutta típusú módszerekre.

4.18. Tétel *Legyen $R_{k,m}(z)$ az exponenciális függvény maximális rendű Padé-típusú approximációja. Ekkor $R_{k,m}(z)$ pontosan akkor A-megengedett, amikor*

$$m - 2 \leq k \leq m. \tag{4.184}$$

4.19. Példa *Közvetlen számolással ellenőrizhető, hogy az $s = 2$ lépcsős Gauss-Legendre-féle implicit Runge-Kutta típusú módszer stabilitási függvénye*

$$R(z) = \frac{1 + \frac{z}{2} + \frac{z^2}{12}}{1 - \frac{1}{2}z + \frac{z^2}{12}}.$$

Tehát erre a módszerre a stabilitási függvény az $R_{2,2}(z)$ maximális rendű Padé-típusú approximáció. Mivel $k = m = 2$, ezért a 4.18. tétel alapján ez a módszer A-stabil.

Megmutathatók a következők.

- Egy s-lépcsős Gauss-Legendre-féle implicit Runge-Kutta típusú módszer stabilitási függvénye az exponenciális függvény $R_{s,s}(z)$ maximális rendű Padé-típusú approximációja.
- Egy s-lépcsős Radau-féle implicit Runge-Kutta típusú módszer stabilitási függvénye az exponenciális függvény $R_{s-1,s}(z)$ maximális rendű Padé-típusú approximációja.
- Egy s-lépcsős Lobatto-féle implicit Runge-Kutta típusú módszer stabilitási függvénye az exponenciális függvény $R_{s-1,s-1}(z)$ maximális rendű Padé-típusú approximációja.

Mindez az alábbi állítást jelenti.

4.20. Tétel *A Gauss-Legendre, a Radau és a Lobatto módszerek egyaránt A-stabilak.*

Ezzel befejeztük az egylépéses módszerek vizsgálatát. A következő részben áttérünk az egylépéses módszerek általánosítására. Mint látni fogjuk, ezek vizsgálata bonyolultabb, és újabb matematikai eszközök alkalmazását is igényli.

5. fejezet

Többlépéses módszerek

Az eddigiekben az egylépéses módszereket vizsgáltuk, vagyis az olyan numerikus módszereket, amelyekkel a megoldásfüggvény valamely rácsháló-pontbeli közelítését az ezen pontot megelőző rácsháló-pontbeli közelítésének segítségével határozzuk meg. A továbbiakban ezt általánosítjuk. A többlépéses módszerek során az új közelítés értékét egynél több megelőző rácspontbeli közelítésből határozzuk meg.
A továbbiakban feltesszük, hogy ezek száma m ($m \geq 1$). Az ilyen módszereket *k-lépéses módszereknek* nevezzük. (A korábbi egylépéses módszereink a többlépéses módszerek speciális eseteinek tekinthetők az $k = 1$ megválasztással.)

Az alábbiakban két egyszerű példán bemutatjuk, hogy a (3.1)-(3.2) Cauchy-feladat megoldásának a megfelelő pontok körüli Taylor-sorba fejtésével hogyan származtathatók ilyen típusú módszerek.

5.21. Példa *Adjunk példát kétlépéses, implicit módszerre!*

Nyilván

$$\begin{aligned} u(t_{n-1}) =& u(t_n) - hu'(t_n) + \frac{h^2}{2}u''(t_n) + \mathcal{O}(h^3), \\ u(t_{n-2}) =& u(t_n) - 2hu'(t_n) + \frac{4h^2}{2}u''(t_n) + \mathcal{O}(h^3). \end{aligned} \tag{5.1}$$

Ezért

$$3u(t_n) - 4u(t_{n-1}) + u(t_{n-2}) = 2hu'(t_n) + \mathcal{O}(h^3) = 2hf(t_i, u(t_i)) + \mathcal{O}(h^3).$$

Így az $f_n = f(t_n, y_n)$ jelöléssel definiálható az

$$y_n - \frac{4}{3}y_{n-1} + \frac{1}{3}y_{n-2} = \frac{2}{3}hf_n, \quad n = 2, 3, \ldots \tag{5.2}$$

módszer. Láthatóan az (5.2) egy kétlépéses implicit módszer, amely másodrendben konzisztens.

5.22. Példa *Adjunk példát kétlépéses, explicit módszerre!*

Először fejtsük Taylor-sorba a megoldásfüggvényt, majd annak deriváltfüggvényét a t_{n-1} pont körül. Ekkor

$$\begin{aligned} u(t_n) =& u(t_{n-1}) + hu'(t_{n-1}) + \frac{h^2}{2}u''(t_{n-1}) + \mathcal{O}(h^3), \\ u'(t_{n-2}) =& u'(t_{n-1}) - hu''(t_{n-1}) + \mathcal{O}(h^2). \end{aligned} \tag{5.3}$$

Mivel a második összefüggésből $hu''(t_{n-1}) = u'(t_{n-1}) - u'(t_{n-2}) + \mathcal{O}(h^2)$, ezt behelyettesítve az első egyenletbe az

$$u(t_n) = u(t_{n-1}) + \frac{h}{2}[3u'(t_{n-1}) - u'(t_{n-2})] + \mathcal{O}(h^3)$$

összefüggést nyerjük. Ez alapján az

$$y_n - y_{n-1} = h[\frac{3}{2}f_{n-1} - \frac{1}{2}f_{n-2}], \quad n = 2, 3, ... \tag{5.4}$$

egy kétlépéses explicit módszer, amely másodrendben konzisztens.

5.23. Megjegyzés Felhívjuk a figyelmet a Runge-Kutta típusú módszerek és többlépéses módszerek közötti alapvető különbségre. A Runge-Kutta típusú módszerek olyan egylépéses módszerek, amikor a t_{n-1} pontbeli közelítésből úgy határozzuk meg a t_n pontbeli közelítést, hogy a két rácspont között s számú pontban köztes értékeket is generálunk. (Ezek általában szintén alacsonyabb rendű approximációk.) A többlépéses módszerek viszont csak a rácspontbeli értékekkel számolnak.

A fenti példáink alapján a k-lépéses módszerek általános alakban is definiálhatók.

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

$$\alpha_0 y_n + \alpha_1 y_{n-1} + \cdots + \alpha_k y_{n-k} = h\left[\beta_0 f_n + \beta_1 f_{n-1} + \ldots + \beta_k f_{n-k}\right], \tag{5.5}$$

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

A továbbiakban mindig feltesszük, hogy $\alpha_0 \neq 0$, hiszen ebben az esetben lehetséges az ismert $y_{n-k}, y_{n-k+1}, \dots, y_{n-1}$ értékekből y_n értékének meghatározása. Mivel $f_n = f(t_n, y_n)$, ezért az (5.5) módszer *explicit*, ha $\beta_0 = 0$, és *implicit*, amikor $\beta_0 \neq 0$. Egy lineáris többlépéses módszer definiálása az α_j és β_j paraméterek ($j = 0, 1, \dots, k$) konkrét értékeinek a megadásával történik. Például, az (5.2) numerikus módszer esetén $k = 2$, $\alpha_0 = 1$, $\alpha_1 = -4/3$, $\alpha_2 = 1/3$, és $\beta_0 = 2/3$, $\beta_1 = 0$, $\beta_2 = 0$. Ha két lineáris többlépéses módszer együtthatói (α_j, β_j) és $(\alpha_j^\star, \beta_j^\star)$, és létezik olyan $c \neq 0$ állandó, amely mellett

$\alpha_j = c\alpha_j^\star$ valamint $\beta_j = c\beta_j^\star$ minden $j = 0, 1, \ldots, k$ értékre, akkor adott kezdeti értékekből mindkét módszer ugyanazt az eredményt szolgáltatja. Ezért az ilyen módszereket nem különböztetjük meg egymástól. Tehát a lineáris többlépéses módszereket leíró $2(k+1)$ paraméter közül egyet érdemes előre rögzítenünk. Ez, konvenció szerint, az α_0 paraméter, és a továbbiakban mindig feltesszük, hogy

$$\alpha_0 = 1. \tag{5.6}$$

Összefoglalóan tehát, a lineáris k-lépéses módszerek alakja

$$\sum_{j=0}^{k} \alpha_j y_{n-j} = h \sum_{j=0}^{k} \beta_j f_{n-j}, \tag{5.7}$$

ahol $\alpha_0 = 1$, és $y_0, y_1, \ldots, y_{k-1}$ adottak.

A lineáris többlépéses módszer kapcsán a következő alapvető kérdések fogalmazhatók meg.

- Hogyan válasszuk meg az α_j, β_j paramétereket?
- Hogyan határozzuk meg az $y_0, y_1, , \ldots, y_{k-1}$ k darab kezdeti feltételt?
- Adott módszer esetén mi mondható el ezen módszer konzisztenciájáról, illetve stabilitásáról?

5.1. Adams-típusú módszerek

Az (5.7) módszer alkalmazása az $u' = f(t,u)$ feladatra azt jelenti, hogy az $\dfrac{1}{h}\displaystyle\sum_{j=0}^{k} \alpha_j y_{n-j}$ kifejezés az u megoldásfüggvény deriváltját, míg $\sum_{j=0}^{k} \beta_j f_{n-j}$ az $f(t,u(t))$ függvényt approximálja a $t = t_n$ pontban.

A lineáris többlépéses módszerek között kiemelkedő szerepet játszanak azok, amelyek a bal oldalon a deriváltat a legegyszerűbb $(y_n - y_{n-1})/h$ véges differenciás approximációval közelítjük, azaz az (5.5) képletben

$$\alpha_0 = 1, \quad \alpha_1 = -1, \quad \alpha_2 = \alpha_3 = \ldots = \alpha_k = 0. \tag{5.8}$$

5.1. Definíció *Az*

$$y_n - y_{n-1} = h \sum_{j=0}^{k} \beta_j f_{n-j} \tag{5.9}$$

alakú módszereket k-lépéses Adams-típusú módszereknek *(vagy röviden* Adams-módszereknek*) nevezzük.*

(Például az (5.4) módszer egy Adams-módszer.) Az Adams-típusú módszereknél a $\beta_0, \beta_1, \ldots, \beta_k$ paraméterek szerepelnek szabad paraméterként. Alapvető különbség létezik a $\beta_0 = 0$ és a $\beta_0 \neq 0$ megválasztású Adams-módszer között: míg az első esetben a módszerek explicit, addig a második esetben implicit.

A továbbiakban megmutatjuk, hogy az Adams-módszerek hogyan kapcsolódnak a numerikus integrálás elméletéhez. Az alapötlet, hogy a $g : [-kh, h] \to \mathbb{R}$ típusú függvényt interpoláljuk az $\omega_h = \{t_l = lh, \quad l = -k, -k+1, \ldots, 0, 1\}$ rácsháló bizonyos pontjaira illeszkedően, majd ezután az interpolációs polinomot integráljuk a $[0, h]$ intervallumon.

5.2. Megjegyzés Ha az interpolációs alappontok a $[-kh, 0]$ intervallumon helyezkednek el, akkor nem a klasszikus numerikus integrálás történik, hiszen a $[0, h]$ intervallumon való integrálással "kimegyünk" az interpolációs alappontok halmazából. (Ez az ún. extrapoláció.)

Az előző részben leírtuk a Lagrange-féle interpoláció alapjait. Most röviden foglaljuk össze a Newton-féle interpolációt! Ha ismerjük a g függvény értékeit a $t = 0, -h, -2h, \ldots, -kh$ pontokban, tehát $k+1$ számú pontban, akkor létezik egyetlen olyan k-ad fokú polinom, amelyre

$$N_k(t_l) = g(t_l), \quad t_l = -lh, \quad l = 0, 1, \ldots, k.$$

Ez a korábban is említett interpolációs polinom, amelynek Lagrange-féle alakját már felírtuk az előző szakaszban, de felírható a következő, Newton-féle osztott differenciás alakban is[1]

$$N_k(t) = \sum_{i=0}^{k} \frac{\nabla^i g(0)}{i! h^i} \omega_i(t), \tag{5.10}$$

ahol

$$\omega_i(t) = \begin{cases} 1, & \text{ha } i = 0 \\ \prod_{l=0}^{i-1} (t - t_l) = t(t+h)\ldots(t+(i-1)h), & \text{ha } i > 0, \end{cases}$$

továbbá

$$\begin{aligned} \nabla^0 g(0) &= g(0), \\ \nabla^1 g(0) &= \nabla g(0) = g(0) - g(-h), \\ \nabla^2 g(0) &= \nabla(\nabla g(0)) = g(0) - 2g(-h) + g(-2h), \\ &\vdots \\ \nabla^i g(0) &= \nabla(\nabla^{i-1} g(0)). \end{aligned}$$

[1] Felhasználjuk, hogy egyenközű felosztás esetén az osztott differencia könnyen meghatározható.

Ekkor

$$g(t) = N_k(t) + r_k(t), \tag{5.11}$$

ahol

$$r_k(t) = \mathcal{O}(h^{k+1}). \tag{5.12}$$

Így

$$\int_0^h g(t)dt = \int_0^h N_k(t)dt + \int_0^h r_k(t)dt = \int_0^h N_k(t)dt + \mathcal{O}(h^{k+2}). \tag{5.13}$$

Behelyettesítve az (5.10) összefüggést (5.13) képletbe, a

$$\int_0^h g(t)dt = \int_0^h \sum_{i=0}^{k} \frac{\nabla^i g(0)}{i!h^i}\omega_i(t)dt + \mathcal{O}(h^{k+2}), \tag{5.14}$$

ezért az

$$\int_0^h g(t)dt = \sum_{i=0}^{k} \frac{\nabla^i g(0)}{h^i} \int_0^h \frac{\omega_i(t)}{i!}dt + \mathcal{O}(h^{k+2}) \tag{5.15}$$

egyenlőséget kapjuk. Mivel a $t = \tau h$ helyettesítéssel

$$\int_0^h \frac{\omega_i(t)}{i!}dt = h\int_0^1 \frac{\tau h(\tau+1)h\ldots(\tau+i-1)h}{i!}d\tau = h^{i+1}\gamma_i, \tag{5.16}$$

ahol

$$\gamma_i = \begin{cases} 1, & \text{ha } i = 0 \\ \displaystyle\int_0^1 \frac{\tau(\tau+1)\ldots(\tau+i-1)}{i!}d\tau, & \text{ha } i > 0, \end{cases} \tag{5.17}$$

ezért, (5.15) és (5.16) alapján

$$\int_0^h g(t)dt = \sum_{i=0}^{k} \frac{\nabla^i g(0)}{h^i}\gamma_i h^{i+1} + \mathcal{O}(h^{k+2}) = h\sum_{i=0}^{k} \gamma_i \nabla^i g(0) + \mathcal{O}(h^{k+2}), \tag{5.18}$$

ahol γ_i az (5.17) képlet alapján számolható.

Az Adams-módszerek közvetlenül nyerhetők az (5.18) numerikus integráló formulából. Ugyanis, legyen $g(t) = u'(t_n + t - h)$. Ekkor

$$\int_0^h u'(t_n + t - h)dt = h\sum_{i=0}^{k} \gamma_i \nabla^i u'(t_n - h) + \mathcal{O}(h^{k+2}), \tag{5.19}$$

azaz

$$u(t_n) - u(t_{n-1}) = h\sum_{i=0}^{k} \gamma_i \nabla^i f(t_n - h, u(t_n - h)) + \mathcal{O}(h^{k+2}). \tag{5.20}$$

Ekkor tehát (5.20), a szokásos $f_n = f(t_n, y_n)$ jelöléssel az

$$y_n - y_{n-1} = h \sum_{i=0}^{k} \gamma_i \nabla^i f_{n-1} \tag{5.21}$$

numerikus sémát generálja.

5.3. Definíció *Az (5.17) képlet szerinti γ_i együtthatókkal definiált (5.21) módszercsaládot Adams-Bashforth-módszernek nevezzük.*

Következésképpen, (5.21) egy explicit Adams-módszer.

Mivel $\nabla^0 f_{n-1} = f_{n-1}$, $\nabla^1 f_{n-1} = f_{n-1} - f_{n-2}$, $\nabla^2 f_{n-1} = f_{n-1} - 2f_{n-2} + f_{n-3}$, stb, ezért

$$\nabla^i f_{n-1} = \sum_{j=0}^{i} (-1)^j \binom{i}{j} f_{n-1-j}, \quad i = 1, 2, \ldots. \tag{5.22}$$

Az (5.22) reláció figyelembevételével (5.21) jobb oldala átírható a következőképpen:

$$\begin{aligned} &\sum_{i=0}^{k} \gamma_i \nabla^i f_{n-1} = \sum_{i=0}^{k} \gamma_i \sum_{j=0}^{i} (-1)^j \binom{i}{j} f_{n-1-j} = \\ &\sum_{j=0}^{k} \sum_{i=j}^{k} \gamma_i (-1)^j \binom{i}{j} f_{n-1-j} = \sum_{j=0}^{k} (-1)^j f_{n-1-j} \sum_{i=j}^{k} \gamma_i \binom{i}{j} = \\ &\sum_{j=1}^{k+1} (-1)^{j-1} f_{n-j} \sum_{i=j-1}^{k} \gamma_i \binom{i}{j-1}. \end{aligned} \tag{5.23}$$

Tehát (5.21) Adams-Bashforth-módszer felírható

$$y_n - y_{n-1} = h \sum_{j=1}^{k+1} \beta_j f_{n-j} \tag{5.24}$$

alakban, ahol

$$\beta_j = (-1)^{j-1} \sum_{i=j-1}^{k} \gamma_i \binom{i}{j-1} \quad j = 1, 2, \ldots, k+1. \tag{5.25}$$

Mivel tetszőleges $m \geq n$ természetes számokra

$$\binom{m}{n} = \frac{m!}{(m-n)!n!} = \frac{m(m-1)\ldots(m-n+1)}{n!},$$

ezért $i > 0$ esetén ezt a jelölést kiterjesztve a valós számokra

$$\gamma_i = \int_0^1 \frac{\tau(\tau+1)\ldots(\tau+i-1)}{i!} d\tau = (-1)^i \int_0^1 \binom{-\tau}{i} d\tau. \tag{5.26}$$

Az (5.25)-(5.26) módszer realizálása során az y_{n-1}, $y_{n-2}, \dots y_{n-k-1}$ értékeiből számoljuk ki y_n értékét, azaz $k+1$ lépéses a módszer. Ezért az (5.9) alak figyelembevételével a k-lépéses Adams-Bashforth-módszerek az alábbi módon írhatók fel:

$$y_n - y_{n-1} = h \sum_{j=0}^{k} \beta_j f_{n-j-1}, \tag{5.27}$$

ahol β_j

$$\beta_j = (-1)^{j-1} \sum_{i=j-1}^{k-1} \gamma_i \binom{i}{j-1} \quad j = 1, 2, \dots, k+1, \tag{5.28}$$

és

$$\gamma_i = (-1)^i \int_0^1 \binom{-\tau}{i} d\tau. \tag{5.29}$$

Mivel a $k+1$ lépéses Adams-Bashforth-módszer pontossága (5.20) alapján $p = k+1$,[2] ezért érvényes az alábbi állítás.

5.4. Tétel *Az (5.27)–(5.29) k-lépéses Adams-Bashforth-módszerek konzisztenciarendje megegyezik a lépésszámokkal (vagyis k-val).*

Tekintsünk néhány példát!

5.5. Példa *Írjuk fel a kétlépéses Adams-Bashforth-módszert!*

Ebben az esetben $k = 2$. Ezért (5.27) alapján $y_n - y_{n-1} = h(\beta_1 f_{n-1} + \beta_2 f_{n-2}$. Határozzuk meg a β_1 és β_2 értékét! Az (5.29) képlet alapján

$$\gamma_0 = \int_0^1 \binom{-\tau}{0} d\tau = 1.$$

$$\gamma_1 = (-1) \int_0^1 \binom{-\tau}{1} d\tau = -\int_0^1 -\tau d\tau = 1/2.$$

Mivel (5.28) alapján

$$\beta_1 = \sum_{i=0}^{1} \gamma_i \binom{i}{0} = \gamma_0 + \gamma_1 = 3/2.$$

$$\beta_2 = -\sum_{i=1}^{1} \gamma_i \binom{i}{1} = -\gamma_1 = -1/2.$$

[2]Mivel a lokális csonkolási hiba rendje $\mathcal{O}(h^{k+2})$.

Tehát a séma:

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

Ezt a módszert a bevezetőben már tárgyaltuk (ld. (5.4)), és megmutattuk, hogy kétlépéses, másodrendben konzisztens módszer. Ez természetesen összhangban áll az 5.4. tétel állításával.

A következő (5.30) képletszámú táblázatban hatodik rendig bezárólag felsoroljuk az ilyen konstrukcióval előállított Adams-Bashforth-módszerek β_j együtthatóit. (A könnyebb áttekinthetőség kedvéért nem törtként írjuk fel ezeket, és a harmadik oszlopban jelezzük, hogy β_j hányszorosa szerepel a táblázatban.)[3]

p	k	β_j	1	2	3	4	5	6
1	1	β_j	1					
2	2	$2\beta_j$	3	-1				
3	3	$12\beta_j$	23	-16	5			
4	4	$24\beta_j$	55	-59	37	-9		
5	5	$720\beta_j$	1901	-2774	2616	-1274	251	
6	6	$1440\beta_j$	4277	-7923	9982	-7298	2877	-475

(5.30)

A táblázatból láthatjuk - az elméletünknek megfelelően- , hogy $p = k$. A $k = 1$ az explicit Euler-módszert jelenti. A $k = 2$ esetén a már ismert (5.4) módszert kapjuk.

Az implicit Adams-módszerek az explicit esethez hasonlóan nyerhetők. Az egyetlen lényeges különbség, hogy a $g(t)$ függvény interpolációjának felépítésénél bevesszük a $t = h$ pontot is. A Newton-féle interpolációs polinomot a $t_l = -lh$, $l = -1, 0, 1, 2, \ldots, k$; $k + 2$ számú pontra fektetjük, és alakja

$$N_{k+1}(t) = \sum_{i=0}^{k+1} \frac{\nabla^i g(0)}{i!h^i}\omega_i(t),$$

ahol

$$\omega_i(t) = \begin{cases} 1, & \text{ha } i = 0 \\ \prod_{l=0}^{i-1}(t - t_l), & \text{ha } i > 0. \end{cases}$$

A fenti interpoláció $[0, h]$ intervallumon való integrálása során, az Adams-Bashforth-módszerekkel megegyező módon definiálhatunk egy implicit Adams-típusú numerikus módszert, amely az

$$y_n - y_{n-1} = h\sum_{j=0}^{k} \tilde{\gamma}_j \nabla^i f_n \tag{5.31}$$

[3] Az így elért rend egyben maximális is, v.ö. a későbbi 5.10. definícióval.

alakú numerikus sémát generálja, ahol

$$\tilde{\gamma}_j = \int_0^1 \frac{(\tau-1)\tau(\tau+1)\dots(\tau+j-1)}{j!} d\tau. \tag{5.32}$$

5.6. Definíció *Az 5.31-(5.32) módon származtatott numerikus módszert Adams-Moulton-módszernek nevezzük.*

Az Adams-Moulton-módszerek pontossága $p = k+1$. Az Adams-Bashforth módszerhez képest történő rendnövekedés az egyel több paraméterrel (azaz a magasabb rendű polinommal való közelítéssel) magyarázható.

Nem részletezve a számításokat, az (5.33) táblázatban ismertetjük az Adams-Moulton-módszerek együtthatóit. A rendjük, az elméletnek megfelelően, $p = k+1$. [4]

p	k	β_j	0	1	2	3	4	5
1	1	β_j	1					
2	1	$2\beta_j$	1	1				
3	2	$12\beta_j$	5	8	-1			
4	3	$24\beta_j$	9	19	-5	1		
5	4	$720\beta_j$	251	646	-264	106	-19	
6	5	$1440\beta_j$	475	1427	-798	482	-173	27

(5.33)

Az első módszer ($k = 1$, $\beta_1 = 0$) az implicit Euler-módszert jelenti. Ez a módszer elsőrendű, így nem maximális rendű. A táblázat második módszere ($k = 1$, $\beta_1 \neq 0$) a (3.51) képletű trapéz-módszert jelenti.

5.7. Megjegyzés Az Adams-Bashforth-módszert gyakran *Adams-féle extrapolációs módszernek*, míg az Adams-Moulton-módszert *Adams-féle interpolációs módszernek* is szokásos nevezni. Ezek a módszerek gyakran együttesen, egymással kombinálva kerülnek felhasználásra a következő módon. Először egy Adams-Bashforth-módszerrel kiszámolt $y_n^\star$ értékkel "megjósoljuk" a t_n időrétegen a közelítő értéket, majd az Adams-Moulton-módszerrel "javítjuk" a numerikus megoldást a következő módon: a jobb oldalon szereplő $\beta_0 f_n$ tagba f_n helyett $f_n^\star = f(t_n, y_n^\star)$ értékét rakjuk. Az így nyert módszert "jósló-javító" (angolul: "predictor-corrector", PC) módszernek nevezzük, de szokásos az összefoglaló *Adams-Bashforth-Moulton-módszer* elnevezés is. Lényeges tulajdonsága, hogy ez a módszer már explicit. (A realizálást illetően ld. a következő szakaszt.)
A PC módszer néhány elméleti részlete megtalálható a
http://math.fullerton.edu/mathews/n2003/AdamsBashforthMod.html
linken, és konkrét formulákkal animáción is megfigyelhetjük a módszer viselkedését az $u'(t) = 1 - t\sqrt[3]{u(t)}$ differenciálegyenleten.

[4] Az így elért rend egyben maximális is, v.ö. a későbbi 5.10. definícióval.

5.2. A retrográd differencia módszerek

Mint láttuk, az Adams-módszerek jellemzője, hogy a kezdetiérték-feladat $u(t)$ megoldására a $t = t_n$ pontban érvényes, a kollokációból eredő $u'(t) = f(t, u(t))$ egyenlőség integrálása után a $t = t_n$ pontban a bal oldalon szereplő deriváltra a szokásos $(u(t_n) - u(t_{n-1}))/h$ véges differenciás approximációt alkalmazzuk, míg a jobb oldalon a $g(t) = f(t, u(t))$ függvényre a Newton-féle interpolációs polinomot fektetjük, és azt integráljuk.

Egy másik lehetséges eljárás, amikor közvetlenül alkalmazzuk az egyenlet megoldására a $t = t_n$ pontbeli kollokációt, azaz az $u'(t_n) = f(t_n, u(t_n))$ egyenlőségből indulunk ki. Ezután a közelítéshez a bal oldalra alkalmazzuk a Newton-féle interpolációt, a jobb oldalt változatlanul hagyjuk. Ez a megközelítés egy olyan lineáris többlépéses módszert eredményez, amelyben $\beta_1 = \beta_2 = \ldots = \beta_k = 0$, azaz csak $\beta_0 \neq 0$. Ez különösen előnyös lehet a merev feladatok numerikus megoldására. Tekintsük ugyanis újra a már tárgyalt 4.10. példában tárgyalt

$$u'(t) = \lambda(u(t) - g(t)), \quad t \in (0, 1] \tag{5.34}$$

egyenletet! A (4.157) feltétel esetén (tehát $0 < -1/\mathrm{Re}\,\lambda \ll h \ll 1$ esetén) egy adott $t = t^\star$ időpontban $u'(t^\star)/\lambda \to 0$, amikor $\lambda \to -\infty$. Ezért a numerikus megoldással szembeni természetes elvárás, hogy az adott $t = t^\star$ időrétegen, rögzített h esetén $y_n - g(t_n) \to 0$, amikor $\mathrm{Re}(\lambda)h \to -\infty$. (Emlékeztetőül, $t_n = t^\star$.) Ha a jobb oldalra a lineáris többlépéses módszer szokásos (5.5) approximációját felírjuk, akkor a

$$\sum_{j=1}^{k} \beta_j (y_{n-j} - g(t^\star - jh))\lambda$$

közelítést kapjuk. Ez pedig a fenti követelményünket csak akkor elégítheti ki, ha a β_j együtthatók a már említett (tehát $\beta_0 \neq 0$, $\beta_1 = \beta_2 = \ldots = \beta_k = 0$) tulajdonságúak.

Tehát célunk az

$$\alpha_0 y_n + \alpha_1 y_{n-1} + \cdots + \alpha_k y_{n-k} = h\beta_0 f_n \tag{5.35}$$

típusú módszer paramétereinek meghatározása. Ezt ismét a Newton-féle interpolációs polinom segítségével adjuk meg. Jelölje $N_k(t)$ a $g(t)$ függvény $t_l = -lh, \quad l = 0, 1, \ldots, k$ pontjaira épülő interpolációs polinomját, azaz $N_k(t) \in \mathcal{P}_k$ és $N_k(t_l) = g(t_l)$, $l = 0, 1, \ldots, k$. ekkor, mint azt az előző szakaszban már láttuk, $g(t) = N_k(t) + \mathcal{O}(h^{k+1})$ ahol

$$N_k(t) = \sum_{i=0}^{k} \frac{\nabla^i g(0)}{i! h^i} \omega_i(t)$$

és

$$\omega_i(t) = \begin{cases} 1, & \text{ha } i = 0 \\ \displaystyle\prod_{l=0}^{i-1} (t - t_l) = t(t+h)\ldots(t+(i-1)h), & \text{ha } i > 0. \end{cases}$$

Ekkor a $t = h\tau$ transzformációval $\omega_i(h\tau) = h^i\tau(\tau+1)\dots(\tau+i-1)$. Tehát

$$\frac{\omega_i(t)}{i!h^i} = \frac{\tau(\tau+1)\dots(\tau+i-1)}{i!} = (-1)^i\binom{-\tau}{i},$$ [5]

és ezért

$$g(h\tau) \approx N_k(h\tau) = \sum_{i=0}^{k}(-1)^i\binom{-\tau}{i}\nabla^i g(0), \tag{5.36}$$

ahol az approximáció $k+1$-ed rendű.

Legyen $g(h\tau) = u(t_n + h\tau)$. Ekkor (5.36) következtében

$$u(t_n + h\tau) \approx \sum_{i=0}^{k}(-1)^i\binom{-\tau}{i}\nabla^i u(t_n). \tag{5.37}$$

Innen $u'(t_n)$ közelítését mindkét oldal τ szerinti deriválásával nyert egyenlőség $\tau = 0$ melletti alakjából nyerhetjük, azaz

$$\begin{aligned}\left(\frac{d}{d\tau}u(t_n+h\tau)\right)_{\tau=0} \approx \left(\frac{d}{d\tau}\sum_{i=0}^{k}(-1)^i\binom{-\tau}{i}\nabla^i u(t_n)\right)_{\tau=0} = \\ \sum_{i=0}^{k}\left(\frac{d}{d\tau}(-1)^i\binom{-\tau}{i}\right)_{\tau=0}\nabla^i u(t_n).\end{aligned} \tag{5.38}$$

Mivel $i = 0$ indexre $\omega_0 = 1$, és ezért a deriváltja nulla, így az (5.38) összefüggés alapján

$$u'(t_n) \approx \sum_{i=1}^{k}\hat{\gamma}_i\nabla^i u(t_n), \tag{5.39}$$

ahol

$$\hat{\gamma}_i = (-1)^i\frac{d}{d\tau}\binom{-\tau}{i}_{\tau=0} = \frac{d}{d\tau}\left[\frac{\tau(\tau+1)\dots(\tau+i-1)}{i!}\right]_{\tau=0}.$$

Ezért

$$\hat{\gamma}_i = \frac{1\cdot 2\cdot\ldots\cdot(i-1)}{i!} = \frac{1}{i} \quad i = 1,2,\dots,k.$$

Tehát

$$u'(t_n) \approx N_k'(t_n) = \frac{1}{h}\sum_{i=1}^{k}\frac{1}{i}\nabla^i u(t_n). \tag{5.40}$$

Az $u'(t_n) = f(t_n, u(t_n))$ egyenlőségből így a numerikus séma

$$\frac{1}{h}\sum_{i=1}^{k}\frac{1}{i}\nabla^i y_n = f(t_n, y_n). \tag{5.41}$$

[5] Mivel alapértelmezés szerint $\binom{m}{0} = 1$, ezért az $i = 0$ esetre is érvényes a formula.

5.1. Definíció *Az (5.41) numerikus módszert* retrográd differencia módszernek *(angolul* backward differentation method (BDF)*) nevezzük.*

A Newton-féle approximációs polinom pontosságából következően a BDF módszer rendje $p = k$.

5.2. Példa *Határozzuk meg az egylépéses retrográd differencia módszert!*

Mivel $k = 1$, ezért

$$N_1'(t_n) = \frac{1}{h}\nabla^1 y_n = \frac{1}{h}(y_n - y_{n-1}).$$

Tehát a numerikus módszer algoritmusa $\dfrac{y_n - y_{n-1}}{h} = f(t_n, y_n)$, azaz az implicit Euler-módszer.

5.3. Példa *Határozzuk meg a kétlépéses retrográd differencia módszert!*

Mivel $k = 2$, ezért

$$N_2'(t_n) = \frac{1}{h}\left(\nabla^1 y_n + \frac{1}{2}\nabla^2 y_n\right) = \frac{1}{h}\left((y_n - y_{n-1}) + \frac{1}{2}(y_n - 2y_{n-1} + y_{n-2})\right).$$

Így

$$N_2'(t_n) = \frac{1}{h}\left(\frac{3}{2}y_n - 2y_{n-1} + \frac{1}{2}y_{n-2}\right).$$

Ezért az (5.6) szerinti $\alpha_0 = 1$ normalizáló feltétellel a numerikus módszer algoritmusa:

$$\frac{y_n - \frac{4}{3}y_{n-1} + \frac{1}{3}y_{n-2}}{h} = \frac{2}{3}f(t_n, y_n).$$

A következő (5.42) táblázatban megadjuk az első hat maximális rendű retrográd differencia módszer együtthatóit.

p	k	β_0	α_0	α_1	α_2	α_3	α_4	α_5	α_6
1	1	1	1	-1					
2	2	$\frac{2}{3}$	1	$-\frac{4}{3}$	$\frac{1}{3}$				
3	3	$\frac{6}{11}$	1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$			
4	4	$\frac{12}{25}$	1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$		
5	5	$\frac{60}{137}$	1	$-\frac{300}{137}$	$\frac{300}{137}$	$-\frac{200}{137}$	$\frac{75}{137}$	$-\frac{12}{137}$	
6	6	$\frac{60}{147}$	1	$-\frac{360}{147}$	$\frac{450}{147}$	$-\frac{400}{147}$	$\frac{225}{147}$	$-\frac{72}{147}$	$\frac{10}{147}$

(5.42)

Az első módszer ($k = 1$) az implicit Euler-módszert jelenti, míg a $k = 2, 3, 4$ módszereket rendre *másod-, harmad- és negyedrendű Curtis-Hirschfeld-módszernek* nevezzük. Megjegyezzük, hogy $k > 6$ esetén a retrográd differencia módszerek már a legalapvetőbb stabilitási tulajdonsággal, a 0-stabilitással sem rendelkeznek, ezért a hatodrendűnél magasabb módszereket nem alkalmazzák.

5.3. A kezdeti közelítések megválasztása

Ebben a rövid szakaszban megvizsgáljuk a kezdeti közelítések meghatározásának lehetőségét. Ez speciálisan a többlépéses módszerek problémája, hiszen egy k lépéses iteráció elindításához ismerni kell az első k darab lépés eredményét. Ugyanakkor az eredeti kezdetiérték-feladatból csak a kezdeti érték, azaz az első érték ismeretes. Vizsgáljuk meg ezt a kérdést részletesebben!

A probléma tehát az, hogy egy k-lépéses lineáris többlépéses módszer feltételezi a kezdeti közelítések ismeretét a rácsháló első k pontjában, azaz feltesszük, hogy $y_0, y_1, \ldots, y_{k-1}$ adottak. Ugyanakkor, a (3.2) kezdeti feltételből csak y_0 értéke ismert. Ezeket a közelítéseket egy megfelelő rendben pontos egylépéses módszerrel határozzuk meg. (Tipikusan valamely Runge-Kutta típusú módszerrel.) Egy másik lehetőség, és ezt néhány programcsomagban alkalmazzák is, hogy y_l $(l = 1, 2, \ldots k-1)$ értékét egy megfelelően megválasztott $l-1$-lépéses módszerrel számolják. Itt ügyelnünk kell arra, hogy a kiválasztott módszer az alkalmazott lineáris többlépéses módszer rendjével egyezzék meg, hiszen ellenkező esetben a kezdeti közelítések alacsonyabb rendű meghatározásával elveszíthetjük a módszer eredeti pontosságát. Tehát, ha p-ed rendű lineáris többlépéses módszer-t alkalmazunk, akkor a kezdeti közelítéseknek is p-ed rendűeknek kell lenniük, azaz valamely p-ed rendű Runge-Kutta típusú módszert alkalmazunk.

Ha implicit módszert alkalmazunk, akkor lépésenként egy nemlineáris egyenlet vagy egyenletrendszer megoldása szükséges. Ezek megoldására valamely iterációs módszert, többnyire a Newton-módszert szokás használni. Ehhez viszont szükséges az iteráció kezdeti értékének meghatározása. Célravezető, ha ezt az értéket a kezdetiérték-feladatot megoldó valamely explicit módszerrel határozzuk meg.

5.1. Példa *Írjuk fel a negyedrendű Curtis-Hirschfeld retrográd differencia módszer algoritmusát!*

A korábbiaknak megfelelően az alábbi lépéseket tesszük.

- Mivel a Curtis-Hirschfeld retrográd differencia módszer (CH4) negyedrendű, ezért y_1, y_2, és y_3 értékét az RK4 Runge-Kutta típusú módszerrel számoljuk ki.
- Negyedrendű Adams-Bashforth-módszerrel meghatározzuk y_4 közelítését.
- A CH4 módszer felírásával nyert egyenletből meghatározzuk a $t = t_4$ pontbeli közelítést. Az egyenletet Newton-iterációval oldjuk meg, ahol kezdeti közelítésnek az explicit Adams-Bashforth-módszerrel nyert közelítést alkalmazzuk.
- Minden további $t = t_n$ $n \geq 5$ pontbeli közelítést az előző két lépés alapján határozunk meg.

Néhány többlépéses módszer ki is próbálható. A hozzá tartozó programot a az alábbi linkről tudja letölteni az Olvasó:
../programok/multistep.exe

5.4. Általános alakú lineáris többlépéses módszerek rendje

Ebben a szakaszban megvizsgáljuk az (5.5) általános alakú lineáris többlépéses módszer konzisztenciáját és rendjét. A módszer képlethibája

$$g_n(h) = \sum_{j=0}^{k} \left[\alpha_j u(t_{n-j}) - h\beta_j f(t_n - jh, u(t_n - jh))\right]. \tag{5.43}$$

Mivel a (3.1) alapján $u'(t_{n-j}) = f(t_{n-j}, u(t_{n-j}))$, ezért

$$g_n(h) = \sum_{j=0}^{k} \left[\alpha_j u(t_{n-j}) - h\beta_j u'(t_{n-j})\right]. \tag{5.44}$$

Fejtsük a $t = t_n$ pont körül Taylor-sorba p-ed illetve $p-1$-ed rendig az (5.44) jobb oldalán szereplő függvényeket! Ekkor

$$\begin{aligned} u(t_{n-j}) =& u(t_n - jh) = u(t_n) - jhu'(t_n) + \frac{1}{2!}j^2h^2u''(t_n) + \ldots + \\ & (-1)^p\frac{1}{p!}j^ph^pu^{(p)}(t_n) + \mathcal{O}(h^{p+1}), \end{aligned}$$

$$\begin{aligned} u'(t_{n-j}) =& u'(t_n - jh) = u'(t_n) - khu''(t_n) + \ldots + \\ & (-1)^{p-1}\frac{1}{(p-1)!}j^{p-1}h^{p-1}u^{(p)}(t_n) + \mathcal{O}(h^p). \end{aligned}$$

Ezt behelyettesítve az (5.44) összefüggésbe, a képlethibára a

$$\begin{aligned} g_n(h) =& d_0u(t_n) + hd_1u'(t_n) + h^2d_2u''(t_n) + \ldots + \\ & h^pd_pu^p(t_n) + \mathcal{O}(h^{p+1}) \end{aligned} \tag{5.45}$$

kifejezést kapjuk, ahol

$$\begin{aligned} d_0 &= \sum_{j=0}^{k} \alpha_j, \\ d_1 &= -\sum_{j=0}^{k} (j\alpha_j + \beta_j) \\ d_2 &= \sum_{j=0}^{k} \left(\frac{1}{2} j^2 \alpha_j + j\beta_j \right), \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \\ d_p &= (-1)^p \sum_{j=0}^{k} \left(\frac{1}{p!} j^p \alpha_j + \frac{1}{(p-1)!} j^{p-1} \beta_j \right). \end{aligned} \tag{5.46}$$

Egy lineáris többlépéses módszer pontosan akkor p-ed rendű, amikor $g_n(h) = \mathcal{O}(h^{p+1})$, azaz teljesülnek a $d_0 = d_1 = \ldots = d_p = 0$ feltételek. Ez az (5.46) alapján a következőt jelenti.

5.1. Tétel *Az (5.5) lineáris többlépéses módszer p-ed rendű, ha a módszert definiáló paraméterekre teljesülnek az alábbi feltételek:*

$$\begin{aligned} \alpha_0 = 1, \quad \sum_{j=0}^{k} \alpha_j = 0 \\ \frac{1}{s} \sum_{j=0}^{k} j^s \alpha_j + \sum_{j=0}^{k} j^{s-1} \beta_j = 0, \quad s = 1, 2, \ldots, p. \end{aligned} \tag{5.47}$$

Ezek alapján közvetlenül megfogalmazható a konzisztencia feltétele is.

5.2. Következmény Az (5.5) lineáris többlépéses módszer pontosan akkor konzisztens, amikor a módszert definiáló paraméterekre teljesülnek az

$$\begin{aligned} \alpha_0 = 1, \quad \sum_{j=0}^{k} \alpha_j = 0 \\ \sum_{j=0}^{k} j\alpha_j + \sum_{j=0}^{k} \beta_j = 0 \end{aligned} \tag{5.48}$$

feltételek.

5.3. Megjegyzés Az (5.48) feltétel kiírva a következőt jelenti: a módszer akkor konzisztens, amikor

$$\begin{aligned} 1 + \alpha_1 + \ldots + \alpha_k &= 0 \\ (\alpha_1 + 2\alpha_2 + \ldots + k\alpha_k) + (\beta_0 + \beta_1 + \ldots + \beta_k) &= 0. \end{aligned} \tag{5.49}$$

Továbbá, $p \geq 2$ rendben pontos, ha (5.49) mellett teljesülnek az

$$\frac{1}{s}\sum_{j=1}^{k} j^s \alpha_j + \sum_{j=1}^{k} j^{s-1}\beta_j = 0, \quad s = 2, 3, \ldots, p \tag{5.50}$$

feltételek. Ez például azt jelenti, hogy egy k-lépéses módszer másodrendűségéhez a konzisztencia mellett az

$$\frac{1}{2}\sum_{j=1}^{k} j^2 \alpha_j + \sum_{j=1}^{k} j\beta_j = 0 \tag{5.51}$$

feltétel teljesülése szükséges és elégséges. Könnyen ellenőrizhető, hogy az (5.2) numerikus módszer esetén (ahol $k = 2$, $\alpha_0 = 1$, $\alpha_1 = -4/3$, $\alpha_2 = 1/3$, és $\beta_0 = 2/3$, $\beta_1 = 0$, $\beta_2 = 0$) ezek az összefüggések érvényesek.

Milyen pontosságú lehet egy (5.5) alakú lineáris többlépéses módszer? Az általános alakban $2k+1$ paraméter ($\alpha_1, \alpha_2, \ldots, \alpha_k$ és $\beta_0, \beta_1, \ldots, \beta_k$) választható meg. Ugyanakkor ezeknek a paramétereknek a p-ed rendű pontossághoz $p + 1$ feltételt kell teljesíteniük. Így $p \leq 2k$. Ha a módszer explicit, azaz $\beta_0 = 0$, akkor egyel kevesebb a szabadon megválasztható paraméterek száma. Összefoglalóan, érvényes a következő állítás.

5.4. Tétel *Egy k-lépéses implicit lineáris többlépéses módszer maximális rendje $2k$, az explicit lineáris többlépéses módszeré pedig $2k - 1$.*

5.5. Megjegyzés A lineáris többlépéses módszer egyértelműségét biztosító (5.6) $\alpha_0 = 1$ feltétel helyett megadható más feltétel is. Gyakori a

$$\sum_{j=0}^{k} \beta_j = 1 \tag{5.52}$$

feltétel megadása. Ez azt biztosítja, hogy az (5.5)

$$\frac{\alpha_0 y_n + \alpha_1 y_{n-1} + \cdots + \alpha_k y_{n-k}}{h} = \beta_0 f_n + \beta_1 f_{n-1} + \ldots + \beta_k f_{n-k}$$

alakjában a jobb oldalon szereplő kifejezés pontosan approximálja az $f = $ állandó függvényt. Az (5.48) alakból könnyen látható, hogy az (5.52) feltétel mellett a konzisztencia feltétele

$$\sum_{j=0}^{k} \alpha_j = 0 \tag{5.53}$$

$$\sum_{j=1}^{k} j\alpha_j = -1. \tag{5.54}$$

A $p \geq 2$ rendűség feltétele az (5.50) alakból jól láthatóan a

$$\sum_{j=1}^{k} j^{s-1}\left(j\alpha_j + s\beta_j\right) = 0, \quad s = 2, 3, \ldots, p \tag{5.55}$$

feltételek. Mint látható, ebben a felírásban $2k+2$ ismeretlenünk van, és $k+2$ feltételt tűztünk ki rájuk. Így (természetesen) $p \leq 2k$, és a $p = 2k$ maximális rend eléréséhez az együtthatókat az alábbi módon határozhatjuk meg.

1. Megoldjuk az $\alpha_1, \alpha_2, \ldots, \alpha_k$ és $\beta_1, \beta_2, \ldots, \beta_k$ ismeretlenekre a $2k$ számú egyenletből álló (5.54)–(5.55) rendszert.

2. Ezután az (5.52) és az (5.53) feltételekből meghatározzuk az α_0 és β_0 együtthatókat az
$$\alpha_0 = -\sum_{j=1}^{k} \alpha_j = 0, \quad \beta_0 = 1 - \sum_{j=1}^{k} \beta_j \tag{5.56}$$
összefüggések alapján.

A lineáris többlépéses módszerek vizsgálatánál hasznos a következő két polinom bevezetése:

$$\varrho(\xi) = \sum_{j=0}^{k} \alpha_j \xi^{k-j} \tag{5.57}$$

$$\sigma(\xi) = \sum_{j=0}^{k} \beta_j \xi^{n-j}. \tag{5.58}$$

5.6. Definíció *Az (5.57) és (5.58) módon definiált $\varrho(\xi)$ és $\sigma(\xi)$ k-ad fokú polinomokat az (5.5) k-lépéses lineáris többlépéses módszer első illetve második karakterisztikus polinomjának nevezzük.*

Írjuk fel az (5.48) konzisztencia feltételeit a $\varrho(\xi)$ és a $\sigma(\xi)$ polinomok segítségével! Mivel $\varrho(1) = \sum_{j=0}^{k} \alpha_j$, ezért a konzisztencia első feltétele felírható a $\varrho(1) = 0$ alakban. Másrészt,

$$\varrho'(\xi) = \sum_{j=0}^{k-1} (k-j)\alpha_j \xi^{k-j-1},$$

ezért

$$\varrho'(1) = \sum_{j=0}^{k-1}(k-j)\alpha_j = k\sum_{j=0}^{k-1}\alpha_j - \sum_{j=0}^{k-1} j\alpha_j = k\sum_{j=0}^{k-1}\alpha_j - \sum_{j=1}^{k-1} j\alpha_j = \\ k\sum_{j=0}^{k}\alpha_j - \sum_{j=1}^{k} j\alpha_j = k\underbrace{\varrho(1)}_{=0} - \sum_{j=1}^{k} j\alpha_j = -\sum_{j=1}^{k} j\alpha_j. \tag{5.59}$$

Ezért a konzisztencia az (5.48) $\sum_{j=0}^{k} j\alpha_j + \sum_{j=0}^{k}\beta_j = 0$ második feltétele felírható

$$0 = -\varrho'(1) + \sum_{j=0}^{k}\beta_j = -\varrho'(1) + \sigma(1)$$

alakban.

Ezzel beláttuk az alábbi állítást.

5.7. Tétel *Az (5.7) lineáris többlépéses módszer pontosan akkor konzisztens, amikor a karakterisztikus polinomjaira érvényesek a*

$$\varrho(1) = 0, \quad \varrho'(1) = \sigma(1) \tag{5.60}$$

összefüggések.

5.8. Példa *Határozzuk meg az általános alakú Adams-módszerek konzisztenciájának feltételét!*

Az Adams-módszerek esetén $\alpha_0 = 1$, $\alpha_1 = -1$ és $\alpha_2 = \alpha_3 = \ldots = \alpha_k = 0$.

Ezért

$$\varrho(\xi) = \sum_{j=0}^{k}\alpha_j\xi^{k-j} = \xi^k - \xi^{k-1},$$

és így $\varrho(1) = 0$ és $\varrho'(1) = 1$. Az 5.7. tétel alapján tehát a módszer pontosan akkor konzisztens, amikor a

$$1 = \varrho'(1) = \sigma(1) = \sum_{j=0}^{k}\beta_j \tag{5.61}$$

feltétel teljesül.

5.9. Példa *Határozzuk meg az általános alakú Adams-módszerek maximális rendjének feltételét!*

Vizsgáljuk meg, hogy az (5.61) mellett milyen feltételt ró ki az (5.50) p-ed rendűséget biztosító feltétel! Figyelembe véve α_j értékeit, $\sum_{j=1}^{k} j^s \alpha_j = -1$, ezért a p-ed rendűség feltétele

$$s \sum_{j=1}^{k} j^{s-1} \beta_j = 1, \quad s = 2, 3, \ldots, p. \tag{5.62}$$

Mivel $s = 1$ esetén (5.62) az (5.61) feltételt jelenti, ezért összefoglalóan a p-ed rendűség feltétele

$$s \sum_{j=1}^{k} j^{s-1} \beta_j = 1, \quad s = 1, 2, 3, \ldots, p. \tag{5.63}$$

Az (5.61) feltételek p egyenletet jelentenek a $\beta_0, \beta_1, \ldots, \beta_k$ ismeretlenekre. Ezért a k-lépéses Adams-módszerek legfeljebb $k+1$ed rendűek lehetnek. Ha módszer explicit, azaz $\beta_0 = 0$, akkor pedig a rend legfeljebb p lehet.

5.10. Definíció *Azokat az implicit Adams-módszereket, amelyekre $p = k + 1$, illetve azokat az explicit Adams-módszereket, amelyekre $p = k$,* maximális rendű Adams-módszernek *nevezzük.*

Ezért az Adams-Bashforth- illetve az Adams-Moulton-módszerek maximális rendűek.

5.11. Megjegyzés Az első és második karakterisztikus polinomból meghatározható egy lineáris többlépéses módszer p-ed rendűsége is. Megmutatható [3], hogy egy módszer pontosan akkor p-ed rendű, amikor a $\xi = 1$ szám p-szeres gyöke a $\dfrac{\varrho(\xi)}{\ln \xi} - \sigma(\xi)$ komplex függvénynek.

5.5. Az általános alakú lineáris többlépéses módszerek konvergenciája

Áttérünk a lineáris többlépéses módszerek konvergenciájának vizsgálatára. Mint azt már az egylépéses módszereknél láttuk, a konzisztencia önmagában nem elégséges a konvergenciához. Ezen kérdéskör tárgyalása, mint látni fogjuk, meglehetősen összetett feladat. A könnyebb olvashatóság és érthetőség kedvéért először a matematikai alapok és háttér ismertetése nélkül összefoglaljuk a legfontosabb eredményeket, majd utána részletezzük azokat.

5.5.1. Bevezetés a lineáris többlépéses módszerek stabilitásába

A továbbiakban, bizonyítás nélkül, megadunk egy könnyen ellenőrizhető feltételt, amely a konvergenciát biztosítja.[6]
Mint láttuk, konzisztens módszerek esetén a $\xi = 1$ gyöke a $\varrho(\xi)$ karakterisztikus polinomnak. A következő definíció a többi gyök tulajdonságairól szól.

5.1. Definíció *Azt mondjuk, hogy egy lineáris többlépéses módszer* kielégíti a gyökkritériumot, *ha a* $\varrho(\xi) = 0$ *karakterisztikus egyenlet* $\xi_l \in \mathbb{C}$ $(l = 1, 2, \ldots, k)$ *gyökeire* $|\xi_l| \leq 1$, *és a* $|\xi_l| = 1$ *tulajdonságú gyökök egyszeresek.*

Mint azt a következő tétel mutatja, ez a feltétel a lineáris többlépéses módszer 0-stabilitását is jelenti.

5.2. Tétel *Ha egy lineáris többlépéses módszer konzisztens és érvényes rá a gyökkritérium, akkor konvergens is, azaz tetszőleges rögzített* $t^\star \in (0, T)$ *pontban* $h \to 0$ *esetén* $y_n \to u(t^\star)$, *ahol* $nh = t^\star$.

5.3. Megjegyzés A gyökkritérium teljesülésének szükségességét mutatja a következő példa. Tekintsük az

$$y_n + 4y_{n-1} - 5y_{n-2} = h(4f_{n-1} + 2f_{n-2}) \tag{5.64}$$

kétlépéses ($k = 2$) explicit módszert. Könnyen ellenőrizhetően ez a módszer maximálisan pontos, azaz konzisztenciarendje $p = 2k - 1 = 3$. Ugyanakkor első karakterisztikus polinomja $\varrho(\xi) = \xi^2 + 4\xi - 5 = (\xi - 1)(\xi + 5)$, tehát a gyökkritérium nem teljesül. Oldjuk meg az $u' = 0$ feladatot az $u(0) = 0$ kezdeti feltétellel. A feladat megoldása $u(t) = 0$. Legyen továbbá $y_0 = 0$ és $y_1 = \varepsilon$. (Ha valamilyen egylépéses módszerrel kiszámoljuk y_1 értékét, akkor az a módszer hibájának megfelelő rendben, de várhatóan csak kissé fog eltérni a pontos megoldástól az első időpontban, így ε egy nullától különböző kis számnak tekinthető. Mint láttuk, ez az eltérés például a Runge-Kutta típusú módszerek esetén jól becsülhető.) A fenti kétlépéses módszerrel számolva a kezdeti közelítésekre az alábbiakat kapjuk:

$$\begin{aligned} y_2 &= -4y_1 = -4\varepsilon \\ y_3 &= -4y_2 + 5y_1 = 21\varepsilon \\ y_4 &= -4y_3 + 5y_2 = -104\varepsilon \quad \text{stb.} \end{aligned} \tag{5.65}$$

Láthatóan a numerikus megoldás nem marad korlátos, így a konvergencia sem lehetséges.

[6] Tehát ez a szakasz csak egy általános összefoglalást ad. Az elméleti háttér iránt kevésbé érdeklődő Olvasó ezért a következő, a témához kapcsolódó szakaszokat ki is hagyhatja. Ugyanakkor az összefüggések és a bizonyítások iránt érdeklődő Olvasó a következő szakaszokat akár ezen rész elhagyásával is olvashatja.

Fontos megjegyeznünk, hogy a numerikus realizálások során a gyökkritérium általában nem elégséges a stabilitáshoz: vannak esetek, amikor a gyökkritérium teljesülése ellenére a számítások során kialakuló hibák miatt a módszerek nem szolgáltatnak megbízható eredményt. Az ilyen módszereknél az a probléma, hogy a karakterisztikus egyenletének több (bár csak egyszeres) egy abszolút értékű gyöke is van.

5.4. Definíció *Azt mondjuk, hogy egy lineáris többlépéses módszer* erősen stabil, *ha kielégíti a gyökkritériumot, és csak a* $\xi = 1$ *az egyetlen egy abszolút értékű gyöke.*

Például, az alábbi ún. Milne-módszerre

$$y_n - y_{n-2} = \frac{h}{3}(f_n + 4f_{n-1} + f_{n-2})$$

a gyökök $\xi_{1,2} = \pm 1$. Ezért a gyökkritérium érvényes, viszont nem lesz erősen stabil. Ezért ennek a módszernek a használata általában nem ajánlott. Mivel elsősorban az erősen stabil lineáris többlépéses módszerek használata a célszerű, ezért fontos megemlíteni G. Dahlquist eredményét, amely az ilyen típusú módszerek rendjéről szól [7].

5.5. Tétel *Egy erősen stabil* k*-lépéses lineáris többlépéses módszer legfeljebb* $k + 1$*-ed rendű lehet.*

Az egylépéses módszereknél már bemutattuk, hogy egy módszer konvergenciája nem garantálja a megoldás adekvát viselkedését valamely rögzített rácshálón. (Csupán azt biztosítja, hogy a numerikus megoldás a "megfelelően kicsi lépésközű rácshálón" közel van a pontos megoldáshoz. Ugyanakkor ez a lépésköz túlságosan kicsi is lehet.) Ennek elkerülése céljából definiáltuk az abszolút stabil módszereket. (Lásd a 3.3.3 fejezetet.) Felmerülhet a kérdés: mi a helyzet a lineáris többlépéses módszerek abszolút stabilitásával? Sajnálatosan a válasz azt mutatja, hogy ezekre a módszerekre az A-stabilitást nehéz biztosítani. Ugyanis az ún. első és második Dahlquist-korlátok szerint, amelyet egy későbbi szakaszban ismertetünk (lásd 5.13. megjegyzést), a következő korlátok léteznek:

- Az explicit lineáris többlépéses módszerek nem lehetnek A-stabilak.
- Az A-stabil lineáris többlépéses módszerek rendje nem lehet kettőnél nagyobb.

Ezek komoly korlátot jelentenek a lineáris többlépéses módszerek alkalmazhatóságára.

5.5.2. A lineáris többlépéses módszerek konvergenciája

Vizsgáljuk meg a lineáris többlépéses módszerek konvergenciáját.[7]

[7] Ismételten megemlítjük, hogy a fejezet további részeinek olvasásához az előző szakasz ismerete nem feltétlenül szükséges.

Először a tárgyaláshoz feltétlenül szükséges matematikai alapokat tárgyaljuk, nevezetesen, a differenciaegyenletek megoldásának szükséges összefüggéseit ismertetjük. Tekintsük a következő *lineáris, állandó együtthatós, k-ad rendű differenciaegyenletet*:

$$a_k y_{n-k} + a_{k-1} y_{n-k+1} + \ldots + a_0 y_n = b_n, \quad n = k, k+1, \ldots, \tag{5.66}$$

amely kiírva az alábbi végtelen rendszert jelenti:

$$\begin{aligned} a_k y_0 + a_{k-1} y_1 + \ldots + a_0 y_k &= b_k \\ a_k y_1 + a_{k-1} y_2 + \ldots + a_0 y_{k+1} &= b_{k+1} \\ a_k y_2 + a_{k-1} y_3 + \ldots + a_0 y_{k+2} &= b_{k+2} \\ \vdots \qquad\qquad \vdots \qquad\qquad & \vdots \end{aligned} \tag{5.67}$$

Ez a rendszer felírható $\mathbf{A}\mathbf{y} = \mathbf{b}$ alakban, ahol $\mathbf{A} \in \mathbb{R}^{\infty\times\infty}$ végtelen dimenziós mátrix, amelynek alakja

$$\mathbf{A} = \begin{pmatrix} a_k & a_{k-1} & a_{k-2} & \ldots & a_0 & 0 & 0 & 0 & \ldots \\ 0 & a_k & a_{k-1} & \ldots & a_1 & a_0 & 0 & 0 & \ldots \\ 0 & 0 & a_k & \ldots & a_2 & a_1 & a_0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix},$$

$\mathbf{y}, \mathbf{b} \in \mathbb{R}^{\infty}$ végtelen dimenziós vektorok, amelyekre

$$\mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ \vdots \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_k \\ b_{k+1} \\ \vdots \\ \vdots \end{pmatrix}.$$

feladatunk adott $\mathbf{A}$ és $\mathbf{b}$ mellett $\mathbf{y}$ meghatározása. Ez azt jelenti, hogy a feladat megoldása

$$\{y_n\} = (y_0, y_1, \ldots, y_k, \ldots)$$

végtelen sorozat. Az (5.66) egyenlet megoldására érvényesek a közönséges differenciálegyenletek elméletéből ismert állítások megfelelő átfogalmazásai. Nevezetesen, a következő tulajdonságok érvényesek.

- Ha $\{x_n\}$ sorozat jelöli az (5.66) feladat $\mathbf{b} = \mathbf{0}$ melletti (az ún. homogén feladat) általános megoldását, és $\{v_n\}$ sorozat az (5.66) feladat egy tetszőleges megoldását (az ún. partikuláris megoldást), akkor az (5.66) feladat általános megoldása $\{y_n\} = \{x_n\} + \{v_n\}$ alakú, azaz elemenként érvényes az $y_n = x_n + v_n$ egyenlőség.

- Az
$$a_k x_{n-k} + a_{k-1} x_{n-k+1} + \ldots + a_0 x_n = 0, \quad n = k, k+1, \ldots, \tag{5.68}$$
homogén feladatnak k darab lineárisan független megoldása van.

- A homogén feladat k darab lineárisan független megoldását úgy határozzuk meg, hogy a megoldást az $x_n = \xi^n$ alakú elemekből álló sorozat alakjában keressük. Ezt behelyettesítve az (5.68) egyenletbe, az
$$a_k \xi^{n-k} + a_{k-1} \xi^{n-k+1} + \ldots + a_0 \xi^n = 0, \quad n = k, k+1, \ldots, \tag{5.69}$$
egyenleteket kapjuk. Ezt leosztva ξ^{n-k}-vel, eredményül a
$$a_k + a_{k-1}\xi + \ldots + a_0 \xi^k = 0, \quad n = k, k+1, \ldots, \tag{5.70}$$
relációt nyerjük. Bevezetve a $\Phi(\xi) = \sum_{j=0}^{k} a_j \xi^{k-j}$ jelölést, (5.70) a
$$\Phi(\xi) = 0 \tag{5.71}$$
egyenletet jelenti, ahol $\Phi \in \mathcal{P}_k$ egy k-ad fokú polinom. Tegyük fel, hogy az (5.71) egyenletnek k darab valós, különböző gyöke van, ξ_1,ξ_2, …,ξ_k. Ekkor a homogén (5.66) feladat lineárisan független megoldásai az $\{x_n^{(l)}\} = \{\ (\xi_l)^n\}$ sorozatok, ahol $l = 1, 2, \ldots, k$.

- A homogén (5.66) feladat általános megoldása a fenti k darab lineárisan független megoldásainak lineáris kombinációi, azaz a
$$c_1\{\ (\xi_1)^n\} + c_2\{\ (\xi_2)^n\} + \ldots + c_k\{\ (\xi_k)^n\}$$
sorozat. Mindez azt jelenti, hogy a homogén (5.66) feladat általános megoldása az az $\{x_n\}$ sorozat, amelynek n-ik eleme
$$x_n = c_1(\xi_1)^n + c_2(\xi_2)^n + \ldots + c_k(\xi_k)^n. \tag{5.72}$$
Tehát az (5.66) inhomogén feladat általános megoldása egy $\{v_n\}$ partikuláris megoldás esetén:
$$y_n = c_1(\xi_1)^n + c_2(\xi_2)^n + \ldots c_k(\xi_k)^n + v_n = \sum_{j=1}^{k} c_j(\xi_j)^n + v_n. \tag{5.73}$$

- Ha az (5.71) egyenlet k darab valós gyökei közül az egyik gyök kétszeres, azaz $\xi_1 = \xi_2$, akkor az (5.66) inhomogén feladat általános megoldása
$$y_n = c_1(\xi_1)^n + c_2 n(\xi_1)^n + c_3(\xi_3)^n \ldots c_k(\xi_k)^n + v_n. \tag{5.74}$$

- A differenciaegyenlet stabilitása (azaz az $\{y_n\}$ sorozat korlátossága) a $\Phi(\xi)$ ún. karakterisztikus polinom gyökeitől függ. (Emlékeztetőül: stabilitás azt jelenti, hogy a c_i együtthatók kis perturbációja nem okoz nagy eltérést a megoldásban.)[8] Ezért (5.73) és (5.74) alapján a differenciaegyenlet stabil, ha a $\Phi(\xi)$ karakterisztikus polinom gyökeire teljesül a

$$|\xi_l| \leq 1, \quad l = 1, 2, \ldots, k, \text{ és a } |\xi| = 1 \text{ gyök egyszeres} \tag{5.75}$$

  feltétel. (Tehát megengedett a $\xi_1 = 1$ és $\xi_2 = -1$ eset is.) Ugyanakkor a differenciaegyenlet aszimptotikusan stabil (azaz a c_i együttható perturbációjából számított két megoldás különbsége nullához tart), ha

$$|\xi_l| < 1, \quad l = 1, 2, \ldots, k. \tag{5.76}$$

Most áttérünk az (5.5) alakú lineáris többlépéses módszerek stabilitásának (és így konvergenciájának) vizsgálatára. A konvergenciához szükséges 0-stabilitás vizsgálata előtt vizsgáljuk meg, hogy a módszer stabilitási tulajdonsága mivel áll kapcsolatban.

Alkalmazzuk az (5.5) módszert az

$$u' = \lambda u, \quad t > 0 \tag{5.77}$$

tesztegyenletre, ahol $\lambda \in \mathbb{C}$ adott állandó. Ekkor a

$$\sum_{j=0}^{k} (\alpha_j - h\lambda\beta_j) y_{n-j} = 0 \tag{5.78}$$

feladatot nyerjük. Ez egy (5.66) alakú differenciaegyenletet jelent, ahol $a_j = \alpha_j - h\lambda\beta_j$. Ezért tehát a lineáris többlépéses módszer stabilitását meghatározó $\Phi(\xi)$ polinom

$$\Phi(\xi) = \varrho(\xi) - h\lambda\sigma(\xi) \tag{5.79}$$

alakú. Ez is motiválja, hogy egy adott lineáris többlépéses módszer esetén bevezessük a

$$\Pi(\xi) = \varrho(\xi) - h\lambda\sigma(\xi) \tag{5.80}$$

polinomot.

5.6. Definíció *Az (5.80) alakú $\Pi(\xi) \in \mathcal{P}_k$ polinomot az (5.5) lineáris többlépéses módszer* karakterisztikus polinomjának *nevezzük.*

[8] A c_i paramétereket a differenciaegyenlet egyértelmű megoldásához szükséges k darab kiegészítő feltételekből határozzuk meg. Így valójában a stabilitás ezen bemenő paraméterekre vonatkozik: arra ad választ, hogy a bemenő adatok perturbációja hogyan változtatja meg a differenciaegyenlet megoldását.

Tehát az (5.77) tesztfeladatra $\Phi(\xi) = \Pi(\xi)$, és ennek gyökei határozzák meg a módszer stabilitását.

5.7. Megjegyzés Az egyszerűség kedvéért feltehető, hogy a tesztegyenlet kezdeti feltétele $u(0) = 1$. Ekkor a pontos megoldás a $t = t_n$ pontban

$$u(t_n) = e^{\lambda t_n} = e^{\lambda h n} = \left(e^{\lambda h}\right)^n.$$

Ha $\xi_1, \xi_2, \ldots, \xi_k$ a $\Pi(\xi) = 0$ karakterisztikus egyenlet megoldásai, akkor (5.73) alapján a numerikus megoldás

$$y_n = \sum_{j=1}^{k} c_j \xi_j^n \tag{5.81}$$

alakú, ahol a c_j együtthatók ismertek. Ez azt jelenti, hogy a $\xi_1, \xi_2, \ldots, \xi_k$ gyökök egyikének az $e^{\lambda h}$ értékét kell közelítenie, míg az összes többi gyöknek ki kell oltódnia (nullához kell tartania) az n növelésével.[9]

Legyen most a vizsgált differenciálegyenletünkben $f = 0$, azaz feladatunk az $u' = 0$ egyenlet megoldása adott kezdeti feltétellel. Alkalmazzuk ismét az (5.5) alakú lineáris többlépéses numerikus módszert. Ez nyilvánvalóan a

$$\sum_{j=0}^{k} \alpha_j y_{n-j} = 0, \quad n = k, k+1, \ldots$$

differenciaegyenletet eredményezi. Mivel ebben az esetben $\Phi(\xi) = \Pi(\xi) = \varrho(\xi)$, ezért a lineáris többlépéses módszer stabil, ha a $\varrho(\xi)$ első karakterisztikus polinom gyökeire teljesül az (5.75) szerinti gyökkritérium, azaz $|\xi_l| \leq 1$ minden $l = 1, 2, \ldots, k$ esetén, és a $|\xi| = 1$ tulajdonságú gyökök egyszeresek.

Áttérünk a konvergenciához szükséges stabilitásfogalom tárgyalására. Vegyük észre, hogy a 3.2.1 szakaszban leírtaknak megfelelően, ez a módszer is felírható operátoros alakban. Jelölje $\omega_h := \{0 = t_0 < t_1 < \ldots < t_{N-1} < t_N = T\}$, valamint $\omega_h^0 = \omega_h \backslash \{t_l,\ l = 0, 1, \ldots, k-1\}$ rácsokat, és $\mathbb{F}(\omega_h)$ a megfelelő rácsfüggvényeket. Legyen ismét $\mathcal{N}_h$ egy olyan operátor, amely az ω_h pontjaiban értelmezett rácsfüggvényekhez egy másik, szintén az ω_h pontjaiban értelmezett rácsfüggvényt rendel hozzá. Tehát $\mathcal{N}_h : \mathbb{F}(\omega_h) \to \mathbb{F}(\omega_h)$ adott leképezés, amely a lineáris többlépéses módszer esetén egy $w_h \in \mathbb{F}(\omega_h)$ rácsfüggvényhez a következő módon rendel hozzá:

$$(\mathcal{N}_h w_h)(t_n) = \begin{cases} h^{-1} \sum_{j=0}^{k} \alpha_j w_h(t_{n-j}) - \sum_{j=0}^{k} \beta_j f_{n-j}, & \text{ha } t_n \in \omega_h^0, \\ w_h(t_n), & \text{ha } t_n \in \{t_l :\ l = 0, 1, \ldots, k-1\} \end{cases} \tag{5.82}$$

[9] Az $e^{\lambda h}$ értékét közelítő gyököt fő gyöknek ("principal root"), a többi gyököt parazita gyöknek (extraneous root") szokásos nevezni.

Az előző fejezetben megvizsgáltuk a módszer konzisztenciáját, azaz megmutattuk, hogy a

$$d_n = |(\mathcal{N}_h \mathcal{P}_h u)(t_n)|, \quad t_n \in \omega_h^0 \tag{5.83}$$

lokális approximációs hiba milyen feltételek mellett lesz p-ed rendű.[10] A lineáris többlépéses módszerek 0-stabilitása közvetlenül definiálható az egylépéses módszerek (3.7.) definíciója alapján.

5.8. Definíció *Az $\mathcal{N}_h$ operátorral leírt, az (5.5) alakú lineáris többlépéses numerikus módszert 0-stabilnak (zéró-stabilnak) nevezzük, ha léteznek olyan h_0 és K pozitív állandók, hogy minden $h < h_0$ esetén két tetszőleges $\mathbb{F}(\omega_h)$-beli x_h és z_h rácsfüggvényre érvényes az*

$$\begin{aligned} &|x_h(t_n) - z_h(t_n)| \leq \\ &K\{\sum_{j=0}^{k-1} |x_h(t_j) - z_h(t_j)| + \max_{1\leq n\leq N} |h^{-1} \sum_{j=0}^{k} \alpha_j \left(x_h(t_{n-j}) - z_h(t_{n-j})\right) - \\ &\sum_{j=0}^{k} \beta_j \left(f(t_{n-j}, x_h(t_{n-j})) - f(t_{n-j}, z_h(t_{n-j}))\right) |\} \end{aligned} \tag{5.84}$$

egyenlőtlenség tetszőleges $n = k, k+1, \ldots, N$ indexre.

A 0-stabilitás tehát azt fejezi ki, hogy ha két $\mathbb{F}(\omega_h)$-beli rácsfüggvény olyan, hogy a

- a $t = t_l$ $(l = 0, 1, \ldots, k-1$ pontokban közel vannak egymáshoz,
- az $\mathcal{N}_h$-képük $\mathbb{F}(\omega_h^0)$-ben is közel vannak egymáshoz,

akkor maguk a függvények is közel vannak egymáshoz $\mathbb{F}(\omega_h)$-ben. Ez azt jelenti, hogy a numerikus módszert leíró adatokra (az $\mathcal{N}_h$ képzési szabályra és a rácsfüggvény t_l $(l = 0, 1, \ldots, k-1$) pontbeli értékeinek megválasztására nézve a numerikus módszer stabil: ezek kis megváltoztatása esetén a numerikus megoldás is csak korlátosan változik. Megjegyezzük, hogy a 3.9. tételhez hasonlóan a lineáris többlépéses módszerek esetén is érvényes a következő állítás.

5.9. Tétel *A*

$$\mathrm{dom}(\mathcal{N}_h) := \{w_h \in \mathbb{F}(\omega_h) : \ w_h(t_l) \text{ rögzített } l = 0, 1, \ldots, k-1 \text{ esetén}\}$$

halmazon definiált 0-stabil $\mathcal{N}_h$ operátor invertálható.

[10] Emlékeztetünk, hogy az adott fejezetben a csonkolási hibát vizsgáltuk. Ezért a konzisztencia rendje a csonkolási hiba rendjénél egyel kisebb.

A 0-stabilitás az 5.8. szerinti definíciójából, az $x_h = y_h$ és a $z_h = P_h u$ megválasztással, az egylépéses módszerrel megegyező módon kapjuk a lineáris többlépéses módszerre vonatkozó alaptételt.

5.10. Tétel *Tegyük fel, hogy*

- *a kezdetiérték-feladatnak létezik egyértelmű megoldása;*
- *az $\mathcal{N}_h : \mathbb{F}(\omega_h) \to \mathbb{F}(\omega_h)$ operátor p-ed rendben konzisztens,*
- *a kezdeti $y_0, y_1, \dots y_{k-1}$ közelítések szintén p-ed rendben pontosak,*
- *a lineáris többlépéses módszer 0-stabil.*

Ekkor

- *az lineáris többlépéses módszerrel definiált y_h közelítő megoldás létezik és egyértelmű,*
- *a numerikus megoldások sorozata konvergens, és a konvergencia rendje megegyezik a konzisztencia rendjével.*

A lineáris többlépéses módszer 0-stabilitásának belátása nagyon összetett feladat. Hosszú és bonyolult számolás után megmutatható(lásd például [14]), hogy szoros kapcsolatban áll a módszernek megfeleltetett differenciaegyenletek stabilitásával, mégpedig a következő módon.

5.11. Tétel *Egy lineáris többlépéses módszer pontosan akkor 0-stabil, ha az első karakterisztikus polinomjára teljesül az (5.75) gyökkritérium, azaz $\varrho(\xi) = 0$ egyenlet k darab gyökeire $|\xi_l| \leq 1$ minden $l = 1, 2, \dots, k$ esetén, és a $|\xi| = 1$ tulajdonságú gyökök egyszeresek.*

5.12. Megjegyzés A fenti tétel kicsit meglepő abban az értelemben, hogy a lineáris többlépéses módszer 0-stabilitása csak az első karakterisztikus polinomjának tulajdonságaitól függ, és független a második karakterisztikus polinomtól, azaz csak az α_j paraméterek értéke határozza meg a 0-stabilitást. Ennek az az oka, hogy a 0-stabilitás a módszer $h \to 0$ melletti viselkedését vizsgálja. Ugyanakkor, ha az (5.5) alakú lineáris többlépéses numerikus módszerben h-val nullához tartunk, akkor a

$$\sum_{j=0}^{k} \alpha_j y_{n-j} = 0, \quad n = k, k+1, \dots$$

sémát nyerjük, ami, mint azt láttuk, az $u' = 0$ esetre alkalmazott lineáris többlépéses módszer algoritmusa. Tehát ilyen értelemben a numerikus módszer ezen stabilitási

tulajdonsága csak az u' approximációjától függ. (Mindezt indokolttá teszi az is, hogy a folytonos feladat (3.86) szerinti $\mathcal{L}$ operátorban, ahol $(\mathcal{L}w)(t) = w'(t) - f(t, w(t))$, a w' a legmagasabb rendű tag, tehát az $\mathcal{N}_h$ operátor approximációjában elsődleges szerepet játszik.)

5.13. Megjegyzés Tegyünk néhány észrevételt.

- A konzisztens lineáris többlépéses módszerek esetén, mint azt az (5.60) feltételben megmutattuk, $\varrho(1) = 0$. Tehát konzisztens módszer esetén a $\xi = 1$ mindig gyöke az első karakterisztikus polinomnak. Mivel az egylépéses módszerek esetén nincs más gyök, innen is látszik, hogy a konzisztens egylépéses módszerek automatikusan 0-stabilak, és így konvergensek is.

- A 5.10. tételből, illetve az (5.12.) megjegyzésből következik, hogy β_j értékei nincsenek kihatással a 0-stabilitásra, viszont a konzisztenciára igen. (Lásd az (5.60) $\varrho'(1) = \sigma(1)$ feltételét.) Ez azt jelenti, hogy β_j értékei kihatnak a konzisztenciára, és ezen keresztül a konvergenciára, valamint annak rendjére is.

- Az 5.4. tételben megmutattuk, hogy a maximális rendben konzisztens k-lépéses lineáris többlépéses módszerek rendje $p = 2k$ illetve, explicit módszer esetén $p = 2k-1$. Vajon ez a rend elérhető a 0-stabilitás mellett is? A válasz nemleges, és erre vonatkozóan G. Dahlquist[11] bizonyította be a következőket, amelyet Dahlquist-féle első rendkorlátnak szokásos nevezni.

  1. Ha k páratlan, akkor $p \leq k + 1$;
  2. ha k páros, akkor $p \leq k + 2$;
  3. ha a lineáris többlépéses módszer explicit, akkor $p \leq k$.

  (A bizonyítás meglehetősen technikai, és megtalálható a [19] könyvben.)

- Ha nem teljesül a tétel feltétele, akkor a megoldás instabil. (Lásd az 5.3. megjegyzést.)

A 0-stabilitás megengedi a $\xi_1 = 1$ és a $\xi_2 = -1$ gyököket is. (Csak az a feltétel, hogy ezek egyszeresek legyenek.) Ugyanakkor ez a numerikus realizálás során problémát okozhat!

[11] Germund Dahlquist (1925 – 2005) svéd matematikus, aki a numerikus analízis (elsősorban a közönséges differenciálegyenletk numerikus megoldásában) elméletében ért el kiemelkedő eredményeket. További részletek illetve érdekességek az eredményeiről megtalálhatók a http://www.unige.ch/~wanner/DQsem.pdf és a http://www-history.mcs.st-and.ac.uk/Biographies/Dahlquist.html linkeken.

5.14. Példa *Alkalmazzuk az $u(0) = 1$ kezdeti feltételű (5.77) tesztegyenletre az*

$$y_n - y_{n-2} = \frac{1}{3}h(f_n + 4f_{n-1} + f_{n-2}) \tag{5.85}$$

ún. Milne-módszert!

Erre a módszerre

$$\alpha_0 = 1, \quad \alpha_1 = 0, \quad \alpha_2 = -1,$$
$$\beta_0 = \frac{1}{3}, \quad \beta_1 = \frac{4}{3}, \quad \beta_2 = \frac{1}{3}.$$

Könnyen ellenőrizhető, hogy a módszer másodrendű. (V.ö. az (5.49) és (5.50) feltételeket.) A fenti paraméterek mellett

$$\begin{aligned} \varrho(\xi) =& \xi^2 - 1 \\ \sigma(\xi) =& \frac{1}{3}\xi^2 + \frac{4}{3}\xi + \frac{1}{3}, \\ \Pi(\xi) =& \left(1 - \frac{1}{3}\lambda h\right)\xi^2 + \frac{4}{3}\lambda h\xi - \left(1 + \frac{1}{3}\lambda h\right) = 0. \end{aligned} \tag{5.86}$$

Ekkor a $\Pi(\xi) = 0$ karakterisztikus egyenlet megoldása

$$\xi_{1,2} = \frac{\frac{2}{3}\lambda h \pm \sqrt{1 + \frac{1}{3}(\lambda h)^2}}{1 - \frac{1}{3}\lambda h}.$$

Sorbafejtéssel[12] a két gyök alakja:

$$\xi_1 = e^{\lambda h} + \mathcal{O}\left((\lambda h)^5\right), \quad \xi_2 = -e^{-\left(\frac{\lambda h}{3}\right)} + \mathcal{O}\left((\lambda h)^3\right). \tag{5.87}$$

Ez azt jelenti, hogy $\lambda < 0$ esetén a ξ_2 parazita gyök dominál, és nem a ξ_1 főgyök.

Mi a probléma a fenti feladatban? A $\Pi(\xi)$ karakterisztikus polinom gyökei folytonosan függnek az együtthatóktól, azaz $\lambda h \to 0$ esetén a $\varrho(\xi)$ gyökeihez tart.[13] Ezért, ha $\varrho(\xi)$ valamely gyöke $\xi^{(\rho)}$, akkor $\Pi(\xi)$ megfelelő $\xi^{(\pi)}$ gyöke $\xi^{(\pi)} = \xi^{(\rho)} + \mathcal{O}(\lambda h)$. Így a megoldásban szereplő $\left(\xi^{(\pi)}\right)^n$ megfelelően kis h mellett csak akkor tart nullához $n \to \infty$ esetén, ha $|\xi^{(\rho)}| < 1$. Ha valamely gyökre $|\xi^{(\rho)}| = 1$, akkor viszont $|\xi^{(\pi)}| = 1 + \mathcal{O}(\lambda h)$, azaz $\left(\xi^{(\pi)}\right)^n$ nem tart nullához, azaz ez a gyök nem oltódik ki. Ha $\xi^{(\rho)} = 1$, akkor a karakterisztikus egyenlet megfelelő gyöke $\xi^{(\pi)} = 1 + \mathcal{O}(\lambda h)$, ami azt jelenti, hogy az $e^{\lambda h}$ értéket approximálja, azaz ez a főgyök. Viszont a többi gyöknek ki kell oltódnia, ezért tehát nem lehet az első karakterisztikus polinomnak több, $|\xi^{(\rho)}| = 1$ tulajdonságú gyöke.

Ez motiválja az alábbi fogalom bevezetését.

[12] Itt a $\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \ldots..$ és az $\frac{1}{1-x} = 1 + x + x^2 + \ldots...$ sorbafejtéseket alkalmazzuk.

[13] Az állítás belátása meglehetősen bonyolult, és a Rouché-tételen alapul [11].

5.15. Definíció *Azt mondjuk, hogy egy lineáris többlépéses módszer* erősen stabil, *ha a $\varrho(\xi) = 0$ egyenlet gyökei, a $\xi = 1$ főgyököt kivéve, az egységkörön belül vannak. Ha egy lineáris többlépéses módszer 0-stabil, de nem erősen stabil, akkor* gyengén stabilnak *nevezzük.*

5.16. Példa *Mutassuk meg, hogy az (5.85) alakú Milne-módszer gyengén stabil.*

Mint láttuk, erre a módszerre $\varrho(\xi) = \xi^2 - 1$. Ezért az első karakterisztikus polinom gyökei $\xi_{1,2} = \pm 1$, tehát a módszer 0-stabil, viszont nem erősen stabil.

Mi mondható el az erősen stabil és konzisztens lineáris többlépéses módszerek rendjéről? Erre vonatkozik G. Dahlquist alábbi tétele.[14]

5.17. Tétel *Egy erősen stabil k-lépéses lineáris többlépéses módszer legfeljebb $k+1$-ed rendű lehet.*

Vizsgáljuk meg az Adams-módszerek stabilitását! Mivel ezekre a módszerekre $\alpha_0 = 1$, $\alpha_1 = -1$, $\alpha_i = 0$ minden $i = 2, 3, \ldots, k$ esetén, ezért az első karakterisztikus polinomja $\varrho(\xi) = \xi^k - \xi^{k-1}$. Tehát a gyökök: $\xi_1 = 1$, és $\xi_i = 0$ minden $i = 2, 3, \ldots, k$ esetén. Ez azt jelenti, hogy az Adams-módszerek erősen stabilak. Mivel az Adams-Moulton-módszer konzisztenciarendje $k+1$, ezért ez a módszer erősen stabil és maximális rendben pontos. (Ez is indokolja a módszert széleskörű alkalmazását.)

A retrográd differencia módszerek bevezetésénél megemlítettük, hogy jól alkalmazhatók a merev feladatok numerikus megoldására. Ez nem jelenti azt, hogy ezek a módszerek automatikusan erősen stabilak is. Megmutatható, hogy a retrográd differencia módszerek csak $k = 1, 2 \ldots, 6$ esetén 0-stabilak, azaz $k > 7$ esetén a módszereket nem alkalmazzuk. Ugyanakkor a $k \leq 6$ értékekre a retrográd differencia módszerek erősen stabilak is.

Térjünk át a lineáris többlépéses módszerek abszolút stabilitásának vizsgálatára! A 3.3.3 fejezetben ezt a kérdést már tárgyaltuk az egylépéses módszerekre. A lényeg a következő volt: mivel az (5.77) tesztegyenlet $u(t) = e^{\lambda t} u_0$ ($u_0 = u(0) > 0$) pontos megoldása $\mathrm{Re}(\lambda) \leq 0$ esetén monoton csökkenő, ezért alapvető elvárás, hogy az alkalmazott numerikus módszer által származtatott numerikus megoldás is örökölje ezt a tulajdonságot, azaz a numerikus megoldásra teljesüljön az

$$|y_n| \leq |y_{n-1}| \tag{5.88}$$

egyenlőtlenség. Azon $z = \lambda h$ értékeket, amelyek mellett az (5.88) tulajdonság teljesül, a numerikus módszer abszolút stabilitási tartományának neveztük. Jelölje $\mathcal{S} \subset \mathbb{C}$ ezt

[14] Az eredeti bizonyítás elemei megtalálhatók G. Dahlquist "Stability and error bounds in the numerical integration of differential equaitions" (1958) doktori tézisében. A teljes bizonyítás megtalálható [7] könyvében (Theorem 3.9), ahol megmutatják, hogy a $k+2$ rendű módszerek szimmetrikusak, (azaz $\alpha_j = -\alpha_{k-j}$ és $\beta_j = \beta_{k-j}$), és ezért az első karakterisztikus polinomjának mindegyik gyöke az egységkörön van.

a abszolút stabilitási tartományt! Egy numerikus módszert A-stabil módszernek neveztünk, ha az abszolút stabilitási tartománya tartalmazta bal oldali komplex félsíkot, azaz $\mathbb{C}_0^- \subset \mathcal{S}$.

Ez azt jelenti, hogy az (5.5) alakú lineáris többlépéses módszer $\mathcal{S}$ abszolút stabilitási tartományát úgy tudjuk meghatározni, hogy alkalmazzuk az (5.77) tesztegyenletre, és megvizsgáljuk az (5.88) tulajdonság feltételét. Ez a

$$\sum_{j=0}^{k}(\alpha_j - h\lambda\beta_j)y_{n-j} = 0 \tag{5.89}$$

feladatot eredményezi, azaz egy (5.66) alakú differenciaegyenlet, amelyre

$$\Phi(\xi) = \Pi(\xi) = \varrho(\xi) - h\lambda\sigma(\xi). \tag{5.90}$$

Ennek feltétele (v.ö. (5.81)) pedig a következő.

5.18. Tétel *Az (5.5) alakú lineáris többlépéses módszer $\mathcal{S}$ abszolút stabilitási tartománya azon $z \in \mathbb{C}$ számok halmaza, amelyekre a $\Pi(\xi) = \varrho(\xi) - z\sigma(\xi)$ karakterisztikus polinom gyökeire teljesül az (5.75) gyökkritérium.*

5.19. Megjegyzés Az 5.18. tétel következménye, hogy az (5.5) alakú lineáris többlépéses módszer pontosan akkor 0-stabil, amikor a $z = 0$ pontban abszolút stabil, azaz $0 \in \mathcal{S}$.

5.20. Példa *Határozzuk meg explicit Euler-módszer abszolút stabilitási tartományát!*

Mivel az explicit Euler-módszer esetén $\Pi(\xi) = (\xi - 1) - z = \xi - (1 - z)$, és így a karakterisztikus egyenlet gyöke $\xi_1(z) = 1 + z$. Tehát az explicit Euler-módszer abszolút stabilitási tartománya a jól ismert

$$\mathcal{S}_{EE} = \{z \in \mathbb{C} : \quad |1 + z| \leq 1\}$$

halmaz.

5.21. Példa *Határozzuk meg implicit Euler-módszer abszolút stabilitási tartományát!*

Mivel az implicit Euler-módszer esetén $\Pi(\xi) = (\xi - 1) - z\xi = \xi(1 - z) - 1$, és így a karakterisztikus egyenlet gyöke $\xi_1(z) = \dfrac{1}{1 - z}$. Tehát az implicit Euler-módszer abszolút stabilitási tartománya a jól ismert

$$\mathcal{S}_{IE} = \{z \in \mathbb{C}, \quad \left|\frac{1}{1 - z}\right| \leq 1\} = \{z \in \mathbb{C}, \quad |1 - z| \geq 1\}$$

halmaz.

5.22. Példa *Határozzuk meg trapéz-módszer abszolút stabilitási tartományát!*

A trapéz-módszer esetén $\Pi(\xi) = \left(1 - \frac{1}{2}z\right)\xi - \left(1 + \frac{1}{2}z\right)$. Ezért a karakterisztikus egyenlet gyöke $\xi_1(z) = \dfrac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}$. Mivel a $z \mapsto \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$ komplex függvény bijektív módon képezi le a zárt egységkörlapot a $\mathbb{C}^-$ bal oldali komplex félsíkba, ezért a trapéz-módszer abszolút stabilitási tartománya a jól ismert

$$\mathcal{S}_{TR} = \{z \in \mathbb{C}, \quad \mathrm{Re}(z) \leq 0\}$$

halmaz.

5.23. Példa *Határozzuk meg az*

$$y_n - y_{n-2} = h f_{n-1} \tag{5.91}$$

ún. középponti szabály abszolút stabilitási tartományát!

(Vegyük észre, hogy ez módszer különbözik a középponti módszertől.)

Könnyen láthatóan a módszer másodrendű.[15] Erre a módszerre $\varrho(\xi) = \xi^2 - 1$, és $\sigma(\xi) = \xi$, azaz $\Pi(\xi) = \xi^2 - z\xi - 1$. Tehát a karakterisztikus egyenlet gyökei

$$\xi_{1,2}(z) = z \pm \sqrt{z^2 + 1}. \tag{5.92}$$

Először tegyük fel, hogy z tisztán komplex, azaz $z = ib$ alakú. Ekkor $\xi_{1,2}(ib) = \pm\sqrt{1 - b^2} + ib$.

- Ha $b \in (-1, 1)$, akkor $\xi_{1,2} = \pm\sqrt{1 - b^2} + ib$, azaz $\xi_1 \neq \xi_2$, és $|\xi_{1,2}|^2 = (1 - b^2) + b^2 = 1$. Tehát az ilyen z számok $\mathcal{S}_{TR}$-beliek.

- Ha $b = 1$, azaz $z = i$, akkor $\xi_{1,2} = i$. Mivel $|\xi_{1,2}| = 1$ és kétszeres gyök, ezért a $z = i$ nem $\mathcal{S}_{TR}$-beli.

- Ha $b = -1$, azaz $z = -i$, akkor $\xi_{1,2} = -i$. Mivel $|\xi_{1,2}| = 1$ és kétszeres gyök, ezért a $z = -i$ sem $\mathcal{S}_{TR}$-beli.

- Ha $|b| > 1$, akkor $\xi_{1,2} = ib + \pm i\sqrt{b^2 - 1} = i\left(b \pm \sqrt{b^2 - 1}\right)$. Ezért $|\xi_{1,2}| > 1$, és az ilyen tulajdonságú z komplex számok nem $\mathcal{S}_{TR}$-beliek.

[15] Az angol nyelvű irodalomban "leap-frog" módszernek szokásos nevezni.

Most tegyük fel, hogy $z = a + ib$ alakú, ahol $a \neq 0$. Ekkor $|\xi_{1,2}| > 1.$, és így az ilyen tulajdonságú z komplex számok sem $\mathcal{S}_{TR}$-beliek.

Tehát a trapéz-módszer abszolút stabilitási tartománya egy nyílt intervallum, amely a $\mathbb{C}$ komplex számsík $\mathrm{Im}(\mathbb{C})$ képzetes tengelyén helyezkedik el, azaz

$$\mathcal{S}_{TR} = \{z \in \mathbb{C}, \quad z = ib, \ b \in (-1, 1)\} \subset \mathrm{Im}(\mathbb{C}).$$

Tehát a középponti szabályt akkor érdemes alkalmazni, ha λ tisztán képzetes. (Például, az $\mathbf{u}' = \mathbf{A}\mathbf{u}$ lineáris rendszer esetén, amikor $\mathbf{A}$ ferdén szimmetrikus mátrix, azaz $\mathbf{A} = -\mathbf{A}^\top$.)

Általános esetben az abszolút stabilitási tartomány meghatározása meglehetősen bonyolult feladat. Több módszer is létezik, amelyek segítséget nyújtanak a tartomány meghatározásában. Ezek közül az ún. *boundary locus method* (BLM) a legelterjedtebb. A módszer lényege a következő. Mint láttuk, egy lineáris többlépéses módszer $\mathcal{S}$ abszolút stabilitási tartományát a $\Pi(\xi) = 0$ egyenlet z-től függő gyökeinek segítségével határozhatjuk meg. Mint láttuk, konzisztencia esetén van egy $\xi_1(z) \in \mathcal{S}$ főgyöke, amelyre $|\xi_1(z)| = 1$. Ezért valamely $\theta \in [0, 2\pi]$ mellett $\xi_1 = e^{i\theta}$. Tehát $\xi_1 = \xi_1(\theta)$, és $\Pi(e^{i\theta}) = 0$. Ezért a Π polinom definíciója alapján $\varrho(e^{i\theta}) - z\sigma(e^{i\theta}) = 0$, azaz

$$z(\theta) = \frac{\varrho(e^{i\theta})}{\sigma(e^{i\theta})}. \tag{5.93}$$

Tehát adott θ esetén, az (5.93) összefüggés alapján azon z értékek, amelyekre $|\xi(z)| = 1$, meghatározható. Ezt figyelembe véve, és felhasználva a polinomok gyökeinek paramétertől való folytonos függését, az $\mathcal{S}$ abszolút stabilitási tartomány határa nem más, mint a

$$\hat{z}(\theta) = \frac{\varrho(e^{i\theta})}{\sigma(e^{i\theta})}, \quad \theta \in [0, 2\pi] \tag{5.94}$$

komplex zárt görbe által határolt $\mathbb{C}$-beli ponthalmaz.

5.24. Megjegyzés Hogyan lehet meghatározni, hogy az (5.94) zárt görbe belseje vagy külseje képezi az $\mathcal{S}$ abszolút stabilitási tartományt? Ha egy lineáris többlépéses módszer konzisztens, akkor a$\varrho(\xi) = 0$ egyenletnek az $\xi = 1$ érték a gyöke, emellett, ha p-ed rendben konzisztens, akkor a $\Pi(\xi) = 0$ karakterisztikus egyenlet $\xi_1(z)$ főgyöke a $z = 0$ pont körül p-ed rendben approximálja az e^z függvényt. (Pl., az explicit Euler-módszer esetén $\xi_1(z) = 1 + z$, és $1 + z = e^z + \mathcal{O}(z^2)$.) Kis z (azaz kis h) esetén a numerikus megoldás viselkedése $y_n \approx e^z y_{n-1}$. A $z = 0$ pont körül egy kis sugarú kört meghatározva, ezen belül a $y_n = R(z)y_{n-1}$ numerikus módszerben $R(z)$ az e^z függvényhez hasonlóan viselkedik. Mivel $|e^z| < 1$ a $\mathrm{Re}(z) < 0$ értékekre, és $|e^z| > 1$ a $\mathrm{Re}(z) > 0$ értékekre, ezért a kör bal oldalán ($\mathrm{Re}(z) < 0$) az $\mathcal{S}$ abszolút stabilitási tartományban maradunk, míg a jobb oldalán ($\mathrm{Re}(z) > 0$) kilépünk belőle.

5.25. Példa *Határozzuk meg a BLM módszerrel az explicit Euler-módszer abszolút stabilitási tartományát!*

Az explicit Euler-módszer esetén $\varrho(\xi) = \xi - 1$ és $\sigma(\xi) = 1$. Így az $\mathcal{S}_{EE}$ abszolút stabilitási tartomány határa a

$$\hat{z}(\theta) = e^{i\theta} - 1, \quad \theta \in [0, 2\pi]$$

görbe. Ezt a görbét könnyű ábrázolni az (x, y) síkon, mivel

$$e^{i\theta} - 1 = (\cos\theta - 1) + i \sin\theta.$$

Tehát a zárt görbe az

$$x(\theta) = \cos\theta - 1, \quad y(\theta) = \sin\theta, \quad \theta \in [0, 2\pi]$$

paraméteres görbe, amely a szokásos kört jelenti. (Mivel a $z = 0$ halmaz bal oldali része a kör belsejébe esik, ezért $\mathcal{S}$ a zárt kör belsejét jelenti.)

5.26. Példa *Határozzuk meg a BLM módszerrel az implicit Euler-módszer abszolút stabilitási tartományát!*

Az implicit Euler-módszer esetén $\varrho(\xi) = \xi - 1$ és $\sigma(\xi) = \xi$. Így az $\mathcal{S}_{EE}$ abszolút stabilitási tartomány határa a

$$\hat{z}(\theta) = \frac{e^{i\theta} - 1}{e^{i\theta}}, \quad \theta \in [0, 2\pi]$$

görbe. Ezt a görbét könnyű ábrázolni az (x, y) síkon, mivel ebben az esetben

$$x(\theta) = 1 - \cos\theta, \quad y(\theta) = \sin\theta, \quad \theta \in [0, 2\pi]$$

a paraméteres zárt görbe, amely a szokásos kört jelenti. (Mivel a $z = 0$ halmaz jobb oldala szintén a kör belsejébe esik, ezért $\mathcal{S}$ a nyílt kör külseje.)

A már tárgyalt lineáris többlépéses módszerek stabilitási tartománya is felrajzolható a LBM segítségével. Ezek alakjai megtalálhatók a

http://www.scholarpedia.org/article/Linear_multistep_method

linken.

5.27. Megjegyzés A LBM mellett más módszerrel is meghatározható valamely lineáris többlépéses módszer $\mathcal{S}$ abszolút stabilitási tartománya. Ezek közül a leggyakoribb a $\Pi(z)$ karakterisztikus polinom gyökeinek vizsgálata *algebrai módszerekkel*. Tipikusan a *Schur-kritériumot*, illetve a *Routh-Hurwitz-kritériumot* szokásos alkalmazni [17].

Egy adott módszer A-stabilitásához azt kell ellenőrizni, hogy a módszer $\mathcal{S}$ abszolút stabilitási tartománya tartalmazza-e a $\mathbb{C}^-$ félsíkot. Az alábbi eredmény, amelyet *Dahlquist második rendkorlátjaként* ismerünk, a következő választ adja a lineáris többlépéses módszerek A-stabilitására vonatkozóan.

- Az explicit lineáris többlépéses módszerek nem A-stabilak.
- Az A-stabil lineáris többlépéses módszerek maximális rendje $p = 2$.

Így a retrográd differencia módszerek csak $k = 1, 2$ esetén A-stabilak. A fenti Dahlquist-féle rendkorlát azt jelenti, hogy a másodrendnél magasabb lineáris többlépéses módszerek nem A-stabilak. Ugyanakkor számos feladat esetén nem szükséges az A-stabilitás, mivel elegendő az abszolút stabilitás csak azon $\lambda \in \mathbb{C}$ komplex számokra, melyek viszonylag távol esnek a komplex számsík képzetes tengelyétől. Mint ismeretes, minden komplex szám felírható $z = re^{i\theta}$ alakban, ahol $\theta \in (-\pi, \pi]$ és $\arg(z) = \theta$.

5.28. Definíció *Egy numerikus módszert $A(\alpha)$-stabilnak nevezünk, ha az abszolút stabilitási tartományára érvényes az*

$$\mathcal{S} \supset \{z \in \mathbb{C} : \quad z = 0 \cup |\arg(z)| \leq \alpha\}$$

tartalmazás.

Következményként, az $A(\pi/2)$ stabilitás az A-stabilitást, az $A(0)$-stabilitás pedig a valós $\mathbb{R}^-$ tengelyen való stabilitást jelenti.

A retrográd differencia módszer mint említettük, $k = 6$ értékig erősen stabil, és a $k = 1, 2$ értékekre A-stabil A $k = 3, 4, 5, 6, 7$ értékekre $A(\alpha)$ stabil egyre csökkenő szögekkel. Az alábbi táblázatban megadjuk az egyes retrográd differencia módszerek abszolút stabilitási tartományát.

k	1	2	3	4	5	6	7
α	90°	90°	86°	73°	51°	17°	–

A szakasz befejezéseként foglalkozzunk az implicit lineáris többlépéses módszerek realizálásának kérdésével. Ilyenkor általában a létrejövő nemlineáris egyenletet (vagy egyenletrendszert) valamilyen iterációs módszerrel (egyszerű iteráció, avagy merev rendszerek esetén Newton-iteráció) segítségével oldjuk meg. (Az egylépéses módszerek esetén ezt a kérdést a 3.3.2 fejezetben tárgyaltuk részletesen.)

Az implicit lineáris többlépéses módszerek esetén az egyszerű iteráció alakja az n-dik időpontban a következő:

$$y_n^{(s+1)} = h\beta_0 f(t_n, y_n^{(s)}) - \sum_{j=1}^{k} \alpha_j y_{n-j} + h \sum_{j=1}^{k} \beta_j f_{n-j}, \quad s = 0, 1, \ldots, \tag{5.95}$$

ahol az (5.95) iteráció akkor konvergens, amikor a

$$h \left\| \beta_0 \frac{\partial f}{\partial y} \right\| \leq r < 1$$

feltétel teljesül. Mint azt már ebben a fejezetben megjegyeztük, hogy kezdőértéknek egy explicit módszerrel meghatározott értéket használunk. Az iterációt általában akkor állítjuk le, amikor két egymást követő iterált érték eltérése egy előre megadott ε értéknél kisebb. (A gyakorlatban $2-3$ lépést szokásos tenni, és ha ezen belül nem éri el a kívánt pontosságot, akkor h értékét csökkentve (tipikusan $h/2$-t véve) újraszámoljuk a feladatot.

Merev feladatok esetén általában a módosított Newton-iterációt alkalmazzuk a nemlineáris egyenletrendszer megoldására. Ennek alakja a lineáris többlépéses módszerek esetén

$$y_n^{(s+1)} = y_n^{(s)} - \left(\mathbf{I} - h\beta_0 \frac{\partial f}{\partial y} \right)^{-1} \left[\sum_{j=1}^{k} \alpha_j y_{n-j} - h \sum_{j=1}^{k} \beta_j f_{n-j} \right], \quad s = 0, 1, \ldots,$$

ahol $\mathbf{I}$ az egységmátrix, $\dfrac{\partial f}{\partial y}$ a Jacobi mátrix a $(t_n, y_n^{(0)})$ pontban.

Egy p-ed rendű lineáris többlépéses módszer nemlineáris feladatának elkerülésének egy másik módja, általában nem-merev feladatok esetén, amikor előre meghatározott számú iterációs lépést szándékozunk tenni a linearizálás céljából.

A módszer a következő lépésekből áll.

- Egy explicit, szintén p-ed rendű módszerrel "megjósoljuk" (predict) a korábbi időpontbeli közelítésekből a t_n pontbeli közelítést. Tipikusan valamely Adams-Bashforth-módszert szokásos alkalmazni. Legyen ez a módszer a következő:

$$(\mathcal{P}) \quad y_n^{(0)} + \hat{\alpha}_1 y_{n-1} + \ldots + \hat{\alpha}_k y_{n-k} = h \left(\hat{\beta}_1 f_{n-1} + \ldots + \hat{\beta}_k f_{n-k} \right).$$

- Ezután ezzel az értékkel kiértékeljük (evaluate) az f függvényt a t_n pontban, azaz meghatározzuk a

$$(\mathcal{E}) \quad f_n^{(0)} = f(t_n, y_n^{(0)})$$

értéket.

- Ezt az értéket behelyettesítve az implicit lineáris többlépéses módszerbe, korrigáljuk (correct) a megoldást az

$$(\mathcal{C}) \quad y_n^{(1)} + \alpha_1 y_{n-1} + \ldots + \alpha_k y_{n-k} = h \left(\beta_0 f_n^{(0)} + \beta_1 f_{n-1} + \ldots + \beta_k f_{n-k} \right)$$

képlet alapján.

Ha ezzel befejezzük a számításunkat, akkor a módszert PEC módszernek nevezzük.

Ugyanakkor a második ($\mathcal{E}$) és a harmadik ($\mathcal{C}$) lépés megismételhető, mindig az utolsó kiszámolt értékkel. Ha ezt m-szer hajtjuk végre, akkor a módszert $P(EC)^m$ módszernek nevezzük, és a t_n pontbeli közelítés az $y_n = y_n^{(m)}$ érték lesz.

5.29. Definíció *A $P(EC)^m$ típusú módszereket prediktor-korrektor (avagy "jósló-javító") módszernek nevezzük.*

Megjegyezzük, hogy általában még egy kiértékelést (($\mathcal{E}$) lépést) szokásos megtenni, azaz

- $$(\mathcal{E}) \quad f_n^{(m)} = f(t_n, y_n^{(m)})$$

  értéket is meghatározzuk. Ezt azért számoljuk ki, mert amikor áttérünk a következő időpontra (azaz a t_{n+1} pontban határozzuk meg a közelítést a $P(EC)^m$ módszerrel, akkor a jósló ($\mathcal{C}$) formulában a jobb oldalon $\hat{\beta}_1 f_{n-1}$ helyett $\hat{\beta}_1 f_n$ szerepel majd, és ehhez az $f_n = f_n^{(m)}$ értéket használjuk.

Az így nyert teljes módszert $P(EC)^m E$ típusú módszernek nevezzük.

5.30. Definíció *A $P(EC)^m$ és $P(EC)^m E$ típusú módszereket prediktor-korrektor (avagy "jósló-javító") módszereknek nevezzük.*

Befejezésül néhány megjegyzés.

- Általában a kiértékelés (($\mathcal{E}$) lépés) eléggé költséges, ezért a $P(EC)^m E$ módszer helyett alkalmazhatjuk a $P(EC)^m$ módszert, és ebben az esetben az $f_n = f_n^{(m)} = f(t_n, y_n^{(m)})$ helyett az $f_n = f_n^{(m-1)} = f(t_n, y_n^{(m-1)})$ értékkel számolunk. (Ezt már kiszámoltuk a $P(EC)^m$ algoritmusban.) Ugyanakkor megmutatható, hogy a $P(EC)^m E$ módszer abszolút stabilitási tartománya lényegesen nagyobb, mint a $P(EC)^m$ módszeré.

- A korrektor formulát általában csak egyszer alkalmazzuk, azaz a $PECE$ formula alkalmazása a tipikus. Ezért (is) a $P(EC)^m$ és $P(EC)^m E$ módszerek abszolút stabilitási tartománya eltér a javító ($\mathcal{C}$) lépésben alkalmazott implicit lineáris többlépéses módszer stabilitási tulajdonságától.

- A prediktor-korrektor módszert úgy is szokásos alkalmazni, hogy a prediktor módszer $p-1$-ed rendű, a javító viszont p-ed rendű. Megmutatható, hogy ilyenkor a $P(EC)^m$ és a $P(EC)^m E$ módszerek rendje egyaránt p.

6. fejezet

Közönséges differenciálegyenletek kétpontos peremérték-feladata

Láttuk, hogy a közönséges differenciálegyenletek megoldásának egyértelműségéhez kiegészítő feltételek megadása szükséges. Az előző fejezetben ezek a feltételek a megoldás valamely kezdeti ($t = 0$) időpontban való tulajdonságai voltak. Például, az

$$u'' = f(t, u, u') \tag{6.1}$$

másodrendű differenciálegyenlethez az

$$u(0) = u_0; \quad u'(0) = u'_0 \tag{6.2}$$

feltételeket adtuk meg, ahol u_0, u'_0 adott számok. Ugyanakkor gyakori eset, amikor a megoldást a $[0, T]$ korlátos időintervallumon vizsgáljuk, és a megoldás értékét ismerjük ezen időintervallum mindkét végpontjában, vagyis a (6.1) feladat megoldására az

$$u(0) = u_0, \quad u(T) = u_1 \tag{6.3}$$

kiegészítő feltételeket adjuk meg. Ez a feladatosztály lényegesen eltér az eddig tárgyalt feladatosztálytól, ezért vizsgálatuk is különbözik. A következőkben megvizsgáljuk a folytonos feladat megoldhatóságát, illetve megadunk numerikus módszereket ezen feladatok megoldására.

6.1. Bevezetés, motiváció

Kezdjük a peremérték-feladatok tárgyalását egy példával!

6.1. Példa *Tegyük fel, hogy egy rögzített pontból valamely irányba kilövünk egy ágyúgolyót. Jelölje $y(t)$ a kilőtt golyó magasságát, $x(t)$ pedig a kilövési ponttól mért vízszintes távolságát a $t \geq 0$ időpontban. Feltesszük, hogy*

- *a vízszintes (x) irányban állandó sebességgel halad a golyó;*
- *a függőleges (y) irányú mozgására csak a gravitáció hat.*

Határozzuk meg, hogy milyen szögben kell kilőni a golyót ahhoz, hogy egy előre rögzített $x = L$ pontba érkezzen!

A fenti feladat megoldása a következő. Jelölje $v > 0$ az állandó vízszintes irányú sebességet. Ekkor a mozgást leíró egyenletek:

$$\begin{aligned} \dot{x}(t) &= v \\ \ddot{y}(t) &= -g. \end{aligned} \tag{6.4}$$

Emellett $x(0) = 0$ és $y(0) = 0$. (A $t = 0$ kezdeti időpontban sem vízszintesen, sem függőlegesen nem távolodott el a golyó a kezdeti helyzetből.) Ezért a kezdeti feltétel figyelembevételével az első egyenlet megoldása $x(t) = vt$, azaz $t = x/v$. Bevezetve az $y(t) = y(x/v) =: Y(x)$ függvényt, az összetett függvény deriválási szabálya alapján

$$\begin{aligned} \dot{y}(t) &= \frac{dY}{dx}\frac{dx}{dt} = \frac{dY}{dx}v, \\ \ddot{y}(t) &= v\frac{d^2Y}{dx^2}v = v^2\frac{d^2Y}{dx^2}. \end{aligned} \tag{6.5}$$

Ekkor a feltételeink alapján az ismeretlen új függvény az alábbi tulajdonságokkal rendelkezik:

$$\begin{aligned} Y''(x) &= -\frac{g}{v^2}, \quad x \in (0, L) \\ Y(0) &= 0, \quad Y(L) = 0. \end{aligned} \tag{6.6}$$

Mivel ez a differenciálegyenlet könnyen kiintegrálható, ezért a (6.6) feladat megoldása közvetlenül kiszámítható:

$$Y(x) = \frac{gx}{2v^2}(L - x).$$

Ezért a kilövés α szögét a

$$\operatorname{tg}\alpha = Y'(0) = \frac{gL}{2v^2}$$

összefüggésből határozhatjuk meg.

Megjegyezzük, hogy a fenti példában a t változóról x változóra való áttérést azt motiválta, hogy a (6.1) (6.3) peremérték-feladat valamely korlátos *térbeli tartományon* lett kitűzve. Ez eléggé tipikus a peremérték-feladatokra, és mi a továbbiakban ezt a jelölést használjuk.

6.2. Definíció *Az $u = u(x)$ ismeretlen függvényre kitűzött*

$$\begin{aligned} u'' &= f(x, u, u'), \quad x \in (a, b), \\ u(a) &= \alpha, \quad u(b) = \beta \end{aligned} \tag{6.7}$$

feladatot kétpontos peremérték-feladatnak *nevezzük.*

Vegyük észre, hogy a (6.7) peremérték-feladatban a peremfeltétel megadása általánosabban is megfogalmazható. Legyen $g : \mathbb{R}^2 \to \mathbb{R}^2$ egy adott függvény.

6.3. Definíció *Az*

$$\begin{aligned} &u'' = f(x, u, u'), \quad x \in (a, b), \\ &g(u(a), u(b)) = \boldsymbol{0} \end{aligned} \tag{6.8}$$

feladatot általános alakú kétpontos peremérték-feladatnak nevezzük.

6.4. Példa *Adjuk meg a g függvény alakját a (6.7) feladat peremfeltételéhez!*

Legyen $g : \mathbb{R}^2 \to \mathbb{R}^2$ a következő függvény:

$$g(s_1, s_2) = \begin{pmatrix} s_1 - \alpha \\ s_2 - \beta \end{pmatrix}.$$

Ekkor a $g(u(a), u(b)) = \mathbf{0}$ feltétel a szükséges $u(a) = \alpha$ és $u(b) = \beta$ feltétellel ekvivalens.

6.2. Közönséges differenciálegyenletek peremérték-feladatának megoldhatósága

A továbbiakban azt vizsgáljuk meg, hogy a (6.7) peremérték-feladatnak milyen feltételek mellett létezik egyértelmű megoldása. Emlékeztetünk, hogy az előző részben ezt a kérdést a kezdetiérték-feladatokra vizsgáltuk, és megmutattuk, hogy a differenciálegyenlet jobb oldalán szereplő f függvény tulajdonsága (nevezetesen, a második változó szerinti lipschitzessége) határozza meg a megoldhatóságot és annak egyértelműségét, függetlenül a kezdeti feltétel (avagy magasabb rendű differenciálegyenletek esetén, a kezdeti feltételek) megválasztásától. Az alábbi példa megmutatja, hogy a peremérték-feladatok esetén ez megváltozik, az f függvény mellett bizonyos esetekben a peremfeltételek is kihatnak a megoldás létezésére és annak egyértelműségére.

6.1. Példa *Legyen a (6.3) feladatban $f(x, u, u') = -u$, tehát vizsgáljuk az*

$$u'' = -u \tag{6.9}$$

egyenletet. Mint ismeretes, (6.9) általános megoldása $u(x) = C_1 \sin x + C_2 \cos x$, ahol C_1 és C_2 állandók. Ez utóbbiak meghatározására szolgálnak a (6.7) feladatban a $x = a$ és $x = b$ pontokban megadott peremfeltételek.

- *Legyen először $a = 0$, $b = \pi/2$, $\alpha = 3$ és $\beta = 7$. Könnyen láthatóan ekkor az egyértelmű megoldás az $u(x) = 7 \sin x + 3 \cos x$ függvény.*

- *Legyen most $a = 0$, $b = \pi$, $\alpha = 3$ és $\beta = 7$. (Tehát csak b értékét változtattuk meg.) Behelyettesítve az általános megoldásba láthatóan nincs olyan C_1 és C_2 értékpár, amely mellett ez a peremfeltétel teljesül.*

- *Legyen $a = 0$, $b = \pi$, $\alpha = \beta = 0$. Az általános megoldásba behelyettesítve a peremfeltételeket láthatjuk, hogy a $C_1 =$ tetszőleges és $C_2 = 0$ megválasztással megoldást kapunk, azaz az $u(x) = C_1 \sin x$ függvény, ahol C_1 tetszőleges, megoldása a peremérték-feladatnak.*

Megjegyezzük, hogy van olyan példa is, amelyben létezik ugyan a peremérték-feladatnak megoldása, de az nem egyértelmű. Például, az

$$\begin{aligned} u'' &= -\exp(u+1), \quad x \in (a,b), \\ u(0) &= 0, \quad u(1) = 0 \end{aligned} \tag{6.10}$$

feladatnak megoldása az

$$u(x) = -2 \ln \frac{\cosh[(x-0.5)\theta/2]}{\cosh(\theta/4)}$$

függvény, ahol θ a $\theta = \sqrt{2e}\cos(\theta/4)$ egyenlet megoldása. Mivel ez utóbbi egyenletnek két megoldása van, ezért a (6.10) feladatnak két megoldása létezik.

A következő tétel egy elégséges feltételt ad egyértelmű megoldás létezésére. (Lásd [17])

6.2. Tétel *Tegyük fel, hogy a $T := \{(x, s_1, s_2) : x \in [a,b],\ s_1, s_2 \in \mathbb{R}\}$ jelöléssel a (6.7) feladat $f : \mathbb{R}^3 \to \mathbb{R}$ függvényére teljesülnek a következők:*

1. *$f \in C(T)$,*
2. *$\partial_2 f, \partial_3 f \in C(T)$,*
3. *$\partial_2 f > 0$ T-n,*
4. *létezik olyan $M \geq 0$, amelyre $|\partial_3 f| \leq M$ T-n.*

Ekkor a (6.7) peremérték-feladatnak létezik egyértelmű megoldása.

A 6.2. tétel fontos következménye az alábbi állítás.

6.3. Következmény Legyen f lineáris, azaz tekintsük az

$$\begin{aligned} u'' &= f(x,u,u') \equiv p(x)u' + q(x)u + r(x), \quad x \in [a,b], \\ u(a) &= \alpha, \quad u(b) = \beta \end{aligned} \tag{6.11}$$

feladatot, ahol $p, q, r \in C[a,b]$ adott folytonos függvények. Ha $q(x) > 0$ az $[a,b]$ intervallumon, akkor a (6.11) lineáris peremérték-feladatnak létezik egyértelmű megoldása.

Mint azt a 6.1. példa is mutatja, a $q(x) > 0$ feltétel nem hagyható el, hiszen ezt a feltételt kivéve a példában szereplő $f(x, s_1, s_2) = -s_1$ függvényre a 6.3. következményben szereplő valamennyi feltétel teljesül. Ugyanakkor, mint azt beláttuk, a második esetben nem létezik megoldás.

A második fejezetben megmutattuk, hogy a magasabbrendű differenciálegyenletek átírhatók elsőrendű rendszerek alakjában. (Lásd a 2.7. megjegyzést.) Alkalmazva ezt az átviteli elvet, a (6.7) peremérték-feladat egyenlete is átírható kétismeretlenes rendszerré. Vezessük be az $\mathbf{u} : [a, b] \to \mathbb{R}^2$, $\mathbf{u}(x) = (u_1(x), u_2(x))$ függvényt a következő módon: $u_1(x) = u(x)$ és $u_2(x) = u'(x)$. Ekkor a feladatunk felírható

$$\begin{aligned} &u_1' = u_2, \quad u_2' = f(x, u_1, u_2), \\ &u_1(a) = \alpha, \quad u_1(b) = \beta \end{aligned} \tag{6.12}$$

alakban. Természetesen a (6.12) feladat ebben a formában nem oldható meg, hiszen csak az u_1 függvényre ismerünk kiegészítő feltételeket.

Megjegyezzük, hogy a (6.12) egyenlet speciális alakja a következő általános alakban felírt egyenletnek:

$$\mathbf{u}' = \mathbf{f}(x, \mathbf{u}), \quad x \in [a, b] \tag{6.13}$$

ahol $\mathbf{f} : \mathbb{R}^3 \to \mathbb{R}^2$ adott függvény. Ugyanis, ha a (6.13) feladatban

$$\mathbf{f}(x, \mathbf{u}) = \mathbf{f}(x, u_1, u_2) = \begin{pmatrix} u_2 \\ f(x, u_1, u_2) \end{pmatrix}, \quad x \in [a, b] \tag{6.14}$$

alakban választjuk meg az $\mathbf{f}$ függvényt, akkor éppen a (6.12) egyenletrendszerét kapjuk.

6.3. A lineáris peremérték-feladat megoldhatósága

Írjuk fel

a (6.11) lineáris peremérték-feladatot elsőrendű rendszer alakjában! Könnyen láthatóan az egyenlet

$$\mathbf{u}' = \mathbf{A}(x)\mathbf{u} + \mathbf{r}(x) \tag{6.15}$$

alakú, ahol

$$\mathbf{A}(x) = \begin{pmatrix} 0 & 1 \\ q(x) & p(x) \end{pmatrix}, \quad \mathbf{r}(x) = \begin{pmatrix} 0 \\ r(x) \end{pmatrix}. \tag{6.16}$$

A peremfeltételek felírásához vezessük be a

$$\mathbf{B}_a = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{B}_b = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{6.17}$$

jelöléseket. Ekkor a (6.11) feladat peremfeltétele

$$\mathbf{B}_a\mathbf{u}(a) + \mathbf{B}_b\mathbf{u}(b) = \mathbf{v} \tag{6.18}$$

alakban írható fel.

Ha egy m-ed rendű ($m \geq 2$) lineáris differenciálegyenlet peremérték-feladatát vizsgáljuk, akkor az átviteli elv segítségével az is felírható (6.15)-(6.18) alakban, ahol $\mathbf{A}(x)$, $\mathbf{B}_a$ és $\mathbf{B}_b$ megfelelő, $\mathbb{R}^{m\times m}$ méretű mátrixok.

6.1. Példa *Írjuk fel az*

$$\begin{aligned} u'''(x) = -2\lambda^3 u(x) + \lambda^2 u'(x) + 2\lambda u''(x), \quad x \in (0,1), \\ u(0) = b_1, \quad u(1) = b_2, \quad u'(1) = b_3 \end{aligned} \tag{6.19}$$

lineáris differenciálegyenletet a (6.15)–(6.18) mátrixos alakban!

Legyen $\mathbf{u}(x) = (u(x), u'(x), u''(x))$ egy $[0,1] \to \mathbb{R}^3$ függvény! Ekkor az átviteli elv felhasználásával (6.19) differenciálegyenlete felírható $\mathbf{u}' = \mathbf{A}\mathbf{u}$ alakban, ahol $\mathbf{A} \in \mathbb{R}^{3\times 3}$

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2\lambda^3 & \lambda^2 & 2\lambda \end{pmatrix}$$

alakú mátrix. A peremfeltételek a $\mathbf{B}_a, \mathbf{B}_b \in \mathbb{R}^{3\times 3}$,

$$\mathbf{B}_a = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{B}_b = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{6.20}$$

alakú mátrixok segítségével felírható a (6.18) alakban.

A továbbiakban előállítjuk a (6.15)–(6.18) feladat megoldását.

A (6.15) egyenlet általános megoldása felírható a következő módon. Legyen $\mathbf{Y}(x) \in \mathbb{R}^{m\times m}$ az egyenlet alapmegoldása (más néven fundamentális mátrixa), vagyis az

$$\begin{aligned} \mathbf{Y}'(x) &= \mathbf{A}(x)\mathbf{Y}(x), \quad x \in [a,b] \\ \mathbf{Y}(a) &= \mathbf{I} \end{aligned} \tag{6.21}$$

Cauchy-feladat megoldása, ahol $\mathbf{I} \in \mathbb{R}^{m\times m}$ az egységmátrixot jelöli. Ekkor a (6.15) egyenlet általános megoldása

$$\mathbf{u}(x) = \mathbf{Y}(x)\left(\mathbf{c} + \int_a^x \mathbf{Y}^{-1}(s)\mathbf{r}(s)ds\right), \tag{6.22}$$

ahol $\mathbf{c} \in \mathbb{R}^m$ egy tetszőleges vektor.[1] Célunk $\mathbf{c}$ olyan megválasztása, amely mellett a (6.22) szerint definiált $\mathbf{u}(x)$ függvény kielégíti a (6.18) peremfeltételeket. Behelyettesítve a (6.22) képletet a (6.18) feltételbe, a

$$\mathbf{B}_a\mathbf{u}(a) + \mathbf{B}_b\mathbf{u}(b) = \mathbf{v} = \mathbf{B}_a\mathbf{Y}(a)\mathbf{c} + \mathbf{B}_b\mathbf{Y}(b)\left(\mathbf{c} + \int_a^b \mathbf{Y}^{-1}(s)\mathbf{r}(s)ds\right). \tag{6.23}$$

feltételt kapjuk. Ebből a $\mathbf{c}$ vektorra rendezve és az $\mathbf{Y}(a) = \mathbf{I}$ feltételt figyelembe véve a

$$\left(\mathbf{B}_a + \mathbf{B}_b\mathbf{Y}(b)\right)\mathbf{c} = \mathbf{v} - \mathbf{B}_b\mathbf{Y}(b)\int_a^b \mathbf{Y}^{-1}(s)\mathbf{r}(s)ds \tag{6.24}$$

egyenletet nyerjük. Ezért a

$$\mathbf{Q} = \mathbf{B}_a + \mathbf{B}_b\mathbf{Y}(b) \tag{6.25}$$

jelöléssel a $\mathbf{c}$ vektorra a

$$\mathbf{Q}\mathbf{c} = \mathbf{v} - \mathbf{B}_b\mathbf{Y}(b)\int_a^b \mathbf{Y}^{-1}(s)\mathbf{r}(s)ds \tag{6.26}$$

feladatot kapjuk. Ezért érvényes az alábbi állítás.

6.2. Tétel *A (6.15)–(6.18) lineáris peremérték-feladatnak pontosan akkor létezik egyértelmű megoldása, amikor a (6.25) alakú $\boldsymbol{Q}$ mátrix reguláris. Emellett a megoldás (6.22) alakú, ahol*

$$\boldsymbol{c} = \boldsymbol{Q}^{-1}\left(\boldsymbol{v} - \boldsymbol{B}_b\,\boldsymbol{Y}(b)\int_a^b \boldsymbol{Y}^{-1}(s)\boldsymbol{r}(s)ds\right). \tag{6.27}$$

6.3. Példa *Vizsgáljuk meg az*

$$\begin{aligned} u'' &= -u, \quad x \in (0,b) \\ u(0) &= \alpha, \quad u(b) = \beta \end{aligned} \tag{6.28}$$

feladat megoldhatóságát!

Mivel erre a feladatra

$$\mathbf{A}(x) = \mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{6.29}$$

ezért az alapmegoldása az

$$\mathbf{Y}(x) = \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix} \tag{6.30}$$

[1] A (6.15)–(6.18) feladatban nyilvánvalóan $m = 2$. Ugyanakkor, ha egy tetszőleges $m \geq$ 2-ed rendű közönséges differenciálegyenletet tekintenénk, akkor az átviteli elv alkalmazása után nyert feladatra is érvényesek a továbbiak.

alakú mátrix. Ezért, a $\mathbf{B}_a$ és $\mathbf{B}_b$ mátrixok (6.17) definíciója alapján a (6.25) szerinti $\mathbf{Q}$ mátrix

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 \\ \cos b & \sin b \end{pmatrix} \tag{6.31}$$

alakú. Ez a $\mathbf{Q}$ mátrix pontosan akkor szinguláris, amikor $b = j\pi$, ahol $j \in \mathbb{N}$ tetszőleges pozitív egész szám. Tehát pontosan akkor létezik a (6.28) feladatnak egyértelmű megoldása, amikor a b végpont a π nem egész számú többszöröse. (Ez az eredmény egyben megválaszolja a 6.1. példa eredményét, hogy miért volt egyértelmű megoldás $b = \pi/2$ esetén, és miért nem létezik, vagy éppenséggel miért létezik végtelen sok megoldás a $b = \pi$ megválasztás mellett.)

6.4. Megjegyzés Ha $\mathbf{Q}$ reguláris, akkor vezessük be a

$$\mathbf{\Phi}(x) = \mathbf{Y}(x)\mathbf{Q}^{-1} \tag{6.32}$$

függvényt! Ekkor

$$\mathbf{\Phi}'(x) = \mathbf{Y}'(x)\mathbf{Q}^{-1} = \mathbf{A}(x)\mathbf{Y}(x)\mathbf{Q}^{-1} = \mathbf{A}(x)\mathbf{\Phi}(x).$$

Másrészt

$$\mathbf{B}_a\mathbf{\Phi}(a) + \mathbf{B}_b\mathbf{\Phi}(b) = \left(\mathbf{B}_a\mathbf{Y}(a) + \mathbf{B}_b\mathbf{Y}(b)\right)\mathbf{Q}^{-1} = \mathbf{I}.$$

Ezért a (6.15)–(6.18) lineáris peremérték-feladat esetén a (6.32) alapján definiált $\mathbf{\Phi}(x)$ függvény az alapmegoldás, és segítségével a megoldás a bemenő függvényekből közvetlenül felírható:

$$\mathbf{u}(x) = \mathbf{\Phi}(x)\mathbf{v} + \int_a^t \mathbf{G}(x,s)\mathbf{r}(s)ds, \tag{6.33}$$

ahol[2]

$$\mathbf{G}(x,s) = \begin{cases} \mathbf{\Phi}(x)\mathbf{B}_a\mathbf{\Phi}(a)\mathbf{\Phi}^{-1}(s), & \text{ha } s \le t; \\ -\mathbf{\Phi}(x)\mathbf{B}_b\mathbf{\Phi}(b)\mathbf{\Phi}^{-1}(s), & \text{ha s >t.} \end{cases} \tag{6.34}$$

[2]Ezt a $\mathbf{G}(x,s)$ függvényt a feladat Green-függvényének nevezzük.

7. fejezet

Közönséges differenciálegyenletek kétpontos peremérték-feladatának numerikus megoldása

Bár az előző pontban megmutattuk, hogy egy peremérték-feladat megoldása képletek segítségével előállítható (vö.(6.33)), ez a felírás formális: konkrét feladatok esetén általában sem az alapmegoldás, sem a $\mathbf{G}(x,s)$ függvény nem határozható meg. Ezért a Cauchy-feladatokhoz hasonlóan a peremérték-feladatok esetén is valamely numerikus megoldás alkalmazása szükséges.

Az ismertetésre kerülő numerikus módszereket két csoportba sorolhatjuk.

- A peremérték-feladat megoldását visszavezetjük Cauchy-féle kezdetiérték-feladatokra, és ezek megoldására a korábban ismertetett numerikus módszerek valamelyikét alkalmazzuk.
- Közvetlenül diszkretizáljuk a peremérték-feladatot.

Ezek leírásával és vizsgálatával foglalkozunk a továbbiakban.

7.1. Peremérték-feladat numerikus megoldása Cauchy-feladatra való visszavezetéssel

Ebben a pontban az első módszerrel foglalkozunk, azaz megvizsgáljuk, hogy megfelelő kezdetiérték-feladatra való visszavezetéssel hogyan oldhatók meg a peremérték-feladatok.

7.1.1. A belövéses módszer

Tekintsük az

$$\begin{aligned} \mathbf{u}' &= \mathbf{f}(x, \mathbf{u}), \quad x \in [a, b] \\ u_1(a) &= \alpha, \quad u_1(b) = \beta \end{aligned} \tag{7.1}$$

előző szakaszban már felírt általános alakú peremérték-feladatot! A *belövéses módszer*[1] lényege, hogy a (7.1) feladat helyett az

$$\begin{aligned} \mathbf{u}' &= \mathbf{f}(x, \mathbf{u}), \quad x \in [a, b] \\ \mathbf{u}(a) &= \mathbf{c}, \end{aligned} \tag{7.2}$$

kezdetiérték-feladatot oldjuk meg. Mivel

$$\mathbf{u}(a) = \begin{pmatrix} u_1(a) \\ u_2(a) \end{pmatrix}, \tag{7.3}$$

ezért a $x = a$ pontbeli kezdeti állapotot leíró $\mathbf{c} \in \mathbb{R}^2$ vektor első komponense a (7.1) peremfeltételéből ismert, viszont a második komponensét ($u_2(a) = u'(a)$) nem ismerjük. Jelölje $cn = u_2(a)$. A belövéses módszer lényege, hogy a

$$\mathbf{c} = \begin{pmatrix} \alpha \\ c \end{pmatrix} \tag{7.4}$$

kezdetiérték megválasztásában a c állandót úgy határozzuk meg, hogy a (7.2) feladat megoldása a $x = b$ pontban az előírtaknak megfelelően a β értéket vegye fel. Ez a következőt jelenti. Jelölje $\mathbf{u}(x, c)$ a (7.1) feladatnak a (c paraméter megválasztásától függő) megoldását! Feladatunk a c paraméter olyan megválasztása, amely mellett

$$u_1(b, c) = \beta. \tag{7.5}$$

Ezt a

$$\varphi(c) = u_1(b, c) - \beta \tag{7.6}$$

függvény jelölése mellett a

$$\varphi(c) = 0 \tag{7.7}$$

egyenlet megoldásával kaphatjuk meg. A (7.7) egyenlet egy nemlineáris egyenlet, amelyben $\varphi : \mathbb{R} \to \mathbb{R}$ típusú függvény. Megoldására több ismert módszer is alkalmazható.

Az egyik legegyszerűbb módszer az *intervallum-felező módszer*, azaz keresünk olyan c_1 és c_2 értékeket, amelyekre $\varphi(c_1)\varphi(c_2) < 0$, és ekkor a (c_1, c_2) intervallumon lévő gyököt az intervallum folyamatos felezésével határozhatjuk meg.

A belövéses módszer intervallum-felező módszeres algoritmusa tehát a következő.

[1] A módszer angol elnevezése: *shooting method.*

1. Rögzítünk valamely c értéket.

2. A (7.4) kezdeti vektorral megoldjuk az $[a, b]$ intervallumon a (7.1) kezdetiérték-feladatot, alkalmazva az előző fejezetben ismertetett valamely numerikus módszert. (Például egy Runge-Kutta típusú módszert.)

3. Az 1. lépésben két olyan c értéket keresünk, amelyekre a 2. lépésben kiszámolt $x = b$ pontbeli közelítések ellentétes előjelűek lesznek. Legyenek ezek c_1 és c_2.

4. A $c = 0.5(c_1 + c_2)$ értékkel újra számoljuk a 2. lépést.

5. Az így nyert megoldás $x = b$ pontbeli értékének előjele alapján újra megválasztjuk c_1 vagy c_2 értékét, és folytatjuk az eljárást.

6. Amikor a két végpont távolsága kisebb, mint egy előre megadott $\varepsilon > 0$ érték, akkor befejezzük az iterációt.

A belövéses módszer során a (7.7) egyenlet megoldása szükséges, amihez az előzőekben a felező eljárást alkalmaztuk. Ennél hatékonyabb (bár még mindig csak elsőrendű) a *húrmódszer*. Az előzőek alapján a megoldási algoritmus viszonylag könnyen felírható [5]. Ugyanakkor a felezőmódszer és a húrmódszer közös hátránya a kezdeti ellentétes előjelű alappontok meghatározása. Ha a *szelőmódszert* alkalmazzuk, akkor erre nincs szükség. Ezzel a módszerrel algoritmusunk a következő:

1. Legyenek $c^{(0)}$ és $c^{(1)}$ tetszőleges értékek.

2. A (7.4) kezdeti vektor képletébe először $c^{(0)}$-t, majd $c^{(1)}$-t helyettesítve, megoldjuk az $[a, b]$ intervallumon a (7.1) kezdetiérték-feladatot, alkalmazva az előző fejezetben ismertetett valamely numerikus módszert.

3. Ekkor $\varphi(c^{(0)}) = u_1(b, c^{(0)}) - \beta$ és $\varphi(c^{(1)}) = u_1(b, c^{(1)}) - \beta$.

4. Lineáris közelítést alkalmazva a $(c^{(0)}, \varphi(c^{(0)}))$ és a $(c^{(1)}, \varphi(c^{(1)}))$ pontok között, meghatározzuk azt a $c^{(2)}$ pontot, ahol ez a közelítés nulla értéket vesz fel:

$$c^{(2)} = c^{(1)} - \left(\frac{c^{(1)} - c^{(0)}}{\varphi(c^{(1)}) - \varphi(c^{(0)})} \right) \varphi(c^{(1)}). \tag{7.8}$$

5. A (7.8) képlettel kiszámolt $c^{(2)}$ értékkel meghatározzuk a (7.4) alakú kezdeti vektort, és a Cauchy-feladat numerikus megoldásával meghatározzuk a $\varphi(c^{(2)})$ értéket.

6. A továbbiakban egy kétlépéses iterációt építünk fel a fenti lépések analógiájaként: $k = 2, 3, \ldots$ értékekre a már kiszámolt $(c^{(k-2)}, \varphi(c^{(k-2)}))$ és a $(c^{(k-1)}, \varphi(c^{(k-1)}))$ pontok ismeretében meghatározzuk a

$$c^{(k)} = c^{(k-1)} - \left(\frac{c^{(k-1)} - c^{(k-2)}}{\varphi(c^{(k-1)}) - \varphi(c^{(k-2)})} \right) \varphi(c^{(k-1)}) \tag{7.9}$$

közelítést, majd ennek ismeretében a $\varphi(c^{(k)})$ értéket.

7. Amikor $|\varphi(c^{(k)})|$ értéke kisebb, mint egy előre megadott $\varepsilon > 0$ érték, akkor befejezzük az iterációt.

7.1. Megjegyzés A felezőmódszerrel és a húrmódszerrel összehasonlítva a szelőmódszer előnye tehát, hogy az iteráció elindításához nem szükséges két olyan kezdeti közelítés megkeresése, amely mellett a φ ellentétes előjelet vesz fel ezekben a pontokban, hanem indulhatunk két tetszőleges $c^{(0)}$ és $c^{(1)}$ közelítésből. Ugyanakkor tetszőleges feladatok esetén a módszer konvergenciája a'priori nem biztosítható. (Lásd [5].)

A fenti algoritmus utolsó lépése után szokásos még egy javító lépést tenni. (Ennek oka, hogy ha $\varphi(c^{(k-1)})$ és $\varphi(c^{(k)})$ nagyon közel vannak a nullához, akkor az algoritmusunk instabil. Ezért a 7. lépésben megadott ε értékét nem célszerű túlságosan kicsinek választani.) Ezt a következő módon hajtjuk végre. Elkészítjük a

$$\begin{array}{|c|c|c|c|} \hline \varphi(c^{(0)}) & \varphi(c^{(1)}) & \ldots & \varphi(c^{(n)}) \\ \hline c^{(0)} & c^{(1)} & \ldots & c^{(n)} \\ \hline \end{array} \tag{7.10}$$

táblázatot és egy n-ed fokú interpolációs polinomot fektetünk ezekre az adatokra, azaz meghatározzuk azt a $p(x)$ n-ed fokú polinomot, amelyre $p(\varphi(c^{(k)})) = c^{(k)}$ minden $k = 0, 1, \ldots n$ esetén. Tehát a p polinom a φ^{-1} függvényt (a φ függvény inverzét) interpolálja. Ekkor a φ függvény gyökéhez való $c^{(n+1)}$ közelítést a $p(0) = c^{(n+1)}$ összefüggés alapján határozhatjuk meg.

Az egyenlet megoldására lényegesen hatékonyabb módszer a másodrendű *Newton-módszer*, amelynek során a

$$c^{(s+1)} = c^{(s)} - \frac{\varphi(c^{(s)})}{\varphi'(c^{(s)})}, \quad s = 0, 1, \ldots \tag{7.11}$$

iterációval felépített sorozattal közelítjük a keresett gyököt. (A (7.11) iterációban a $c^{(0)}$ megfelelő megválasztása valamely elsőrendű módszerrel nyerhető.) A (7.11) képlet alkalmazásánál alapvető probléma a $\varphi'(c^{(s)})$ értékének meghatározása.

Ehhez térjünk vissza az eredeti (6.7) feladatra! A belövéses módszer alkalmazása azt jelenti, hogy keressük valamely $c \in \mathbb{R}$ mellett az

$$\begin{aligned} u'' &= f(x, u, u'), \quad x \in (a, b), \\ u(a) &= \alpha, \quad u'(a) = c \end{aligned} \tag{7.12}$$

feladat megoldását. Legyen ez az $u(x,c)$ függvény és ekkor

$$\begin{aligned} u''(x,c) =& f(x,u(x,c),u'(x,c)), \quad x\in(a,b),\\ u(a,c)=\alpha,& \quad u'(a,c)=c. \end{aligned} \tag{7.13}$$

Deriváljuk a (7.13) azonosságot a c paraméter szerint![2] Ekkor

$$\begin{aligned} \frac{\partial u''}{\partial c}(x,c) = \partial_2 f(x,u(t,c),u'(x,c))\frac{\partial u}{\partial c}(x,c)+\partial_3 f(x,u(t,c),u'(x,c))\frac{\partial u'}{\partial c}(x,c),\\ \frac{\partial u(a,c)}{\partial c}=0, \quad \frac{\partial u'(a,c)}{\partial c}=1. \end{aligned} \tag{7.14}$$

Vezessük be a

$$w(x) := \frac{\partial u}{\partial c}(x,c)$$

új függvényt! A (7.14) összefüggések alapján erre a függvényre:

$$\begin{aligned} w''(x) = \partial_2 f(x,u(x,c),u'(x,c))w(x) + \partial_3 f(x,u(x,c),u'(x,c))w'(x),\\ w(a)=0, \quad w'(a)=1. \end{aligned} \tag{7.15}$$

A (7.15) reláció látszólag az ismeretlen $w(t)$ függvényre egy Cauchy-feladat[3]. Ugyanakkor valójában nem az, hiszen a $\partial_2 f$ és $\partial_3 f$ függvényeket az ismeretlen $(x,u(x,c),u'(x,c))$ pontban nem tudjuk kiértékelni. Ha hozzávesszük a (7.15) relációhoz a (7.13) egyenlőséget, akkor egy négyismeretlenes elsőrendű közönséges differenciálegyenlet-rendszer Cauchy-feladatát kapjuk! Nevezetesen, bevezetve a

$$v_1(x)=u(x,c), \quad v_2(x)=u'(x,c), \quad v_3(x)=w(x), \quad v_4(x)=w'(x) \tag{7.16}$$

függvényeket, ezekre a függvényekre a fenti összefüggések alapján a következő kapcsolat áll fenn:

$$\begin{aligned} v_1'(x) =& v_2(x),\\ v_2'(x) =& f(x,v_1(x),v_2(x)),\\ v_3'(x) =& v_4(x),\\ v_4'(x) =& \partial_2 f(x,v_1(x),v_2(x))v_3(x) + \partial_3 f(x,v_1(x),v_2(x))v_4(x). \end{aligned} \tag{7.17}$$

Emellett,

$$v_1(a)=\alpha, \quad v_2(a)=c, \quad v_3(a)=0, \quad v_4(a)=1. \tag{7.18}$$

[2]Ehhez felhasználjuk az alábbi állítást. (Lásd[17].) Jelölje $\Phi(x;x_0,y_0)$ az $y'=f(x,y)$ differenciálegyenlet megoldását az $y(x_0)=y_0$ kezdeti feltétel mellett. Ha az f függvény k-szor folytonosan differenciálható, akkor $\Phi(x;x_0,y_0)$ is k-szor folytonosan differenciálható a harmadik változójában, és minden q esetén az $u(x)=(\partial_3\Phi)(x;x_0,y_0)q$ függvény megoldása az $u'(x)=(\partial_2 f)(x;\Phi(x;x_0,y_0))u(x)$, $u(x_0)=q$ kezdetiérték-feladatnak.

[3]Ezt a feladatot az irodalomban *első variációs egyenletnek* szokás nevezni.

Ezért az $[a,b]$ intervallumon megoldva a

$$\begin{aligned} v_1' =& v_2, \\ v_2' =& f(x, v_1, v_2) \\ v_3' =& v_4, \\ v_4' =& \partial_2 f(x, v_1, v_2) v_3 + \partial_3 f(x, v_1, v_2) v_4, \\ v_1(a) = \alpha, \quad & v_2(a) = c, \quad v_3(a) = 0, \quad v_4(a) = 1 \end{aligned} \tag{7.19}$$

Cauchy-feladatot, meghatározhatjuk az ismeretlen függvényeinket. Mivel

$$v_3(b) = w(b) = \frac{\partial u}{\partial c}(b, c), \tag{7.20}$$

ezért (7.6) miatt $h'(c) = v_3(b)$, azaz a (7.20) szerinti Newton-iteráció jobb oldalának nevezője kiszámítható.

A belövéses módszer Newton-módszeres algoritmusa tehát a következő.

1. Rögzítünk valamely $c^{(0)}$ értéket.
2. A $c = c^{(0)}$ megválasztással az $[a,b]$ intervallumon megoldjuk a (7.19) rendszerre kitűzött kezdetiérték-feladatot az előző fejezetben ismertetett valamely numerikus módszerrel.
3. A megoldások ismeretében kiszámoljuk a $\varphi(c^{(0)}) = v_1(b)$ és $\varphi'(c^{(0)}) = v_3(b)$ értékeket.
4. A (7.11) képlet alapján meghatározzuk a $c^{(1)}$ közelítést.
5. Folytatjuk az eljárást a 2. lépéstől kezdve.
6. Amikor $|\varphi(c^{(k)})|$ értéke kisebb, mint egy előre megadott $\varepsilon > 0$ érték, akkor befejezzük az iterációt.

7.1.2. Lineáris peremérték-feladatok numerikus megoldása

Ebben a szakaszban a (6.11) módon bevezetett

$$\begin{aligned} u'' =& p(x)u' + q(x)u + r(x), \quad x \in [a,b], \\ & u(a) = \alpha, \quad u(b) = \beta \end{aligned} \tag{7.21}$$

lineáris peremérték-feladat numerikus megoldásával foglalkozunk. Megmutatjuk, hogy ha az általános alakú (6.7) feladatban az f függvényt a (7.21) szerinti speciális módon adjuk meg, akkor a numerikus tárgyalás is lényegesen egyszerűbbé válik.

Tegyük fel, hogy a (7.21) feladatot a belövéses módszer segítségével megoldjuk a $c = c_1$ és a $c = c_2$ megválasztással. Jelölje $u_1(x) = u(x, c_1)$ és $u_2(x) = u(x, c_2)$ a két Cauchy-feladat megoldását. Ekkor tehát

$$\begin{aligned} u_1''(x) =& p(x)u_1'(x) + q(x)u_1(x) + r(x), \quad x \in [a, b], \\ & u_1(a) = \alpha, \quad u_1'(a) = c_1, \end{aligned} \tag{7.22}$$

$$\begin{aligned} u_2''(x) =& p(x)u_2'(x) + q(x)u_2(x) + r(x), \quad x \in [a, b], \\ & u_2(a) = \alpha, \quad u_2'(a) = c_2. \end{aligned} \tag{7.23}$$

Legyen továbbá

$$w(x) = \lambda u_1(x) + (1 - \lambda)u_2(x) \tag{7.24}$$

egy új függvény, ahol $\lambda \in \mathbb{R}$ valamely, egyelőre tetszőleges paraméter. Mivel a (7.21) feladat lineáris, ezért az egyenlet két megoldásának (7.24) szerinti lineáris kombinációja is megoldás, azaz

$$w''(x) = p(x)w'(x) + q(x)w(x) + r(x) \quad x \in [a, b]. \tag{7.25}$$

Másrészt $w(a) = \alpha$ és $w(b) = \lambda u_1(b) + (1 - \lambda)u_2(b)$. Ezért válasszuk meg a λ paramétert úgy, hogy teljesüljön a $w(b) = \beta$ feltétel, azaz $\lambda u_1(b) + (1 - \lambda)u_2(b) = \beta$. Innen λ értékét a

$$\lambda = \frac{\beta - u_2(b)}{u_1(b) - u_2(b)} \tag{7.26}$$

egyenlőségből határozhatjuk meg. Ezért a (7.26) értékű λ melletti (7.24) alakú $w(t)$ függvény lesz az eredeti (7.21) lineáris peremérték-feladat megoldása.

A fenti megoldás előállításának számítógépes realizálása a következő módon valósítható meg. Tekintsünk a következő két Cauchy-feladatot:

$$\begin{aligned} u'' =& p(x)u' + q(x)u + r(x), \quad x \in [a, b], \\ & u(a) = \alpha, \quad u'(a) = 0, \end{aligned} \tag{7.27}$$

valamint

$$\begin{aligned} u'' =& p(x)u' + q(x)u + r(x), \quad x \in [a, b], \\ & u(a) = \alpha, \quad u'(a) = 1. \end{aligned} \tag{7.28}$$

Jelölje a (7.27) megoldását $u_1(x)$, a (7.28) feladat megoldását pedig $u_2(x)$. (Ezeket a (7.27) és a (7.28) feladatok elsőrendű rendszerre való visszavezetésével megoldhatjuk az $[a, b]$ intervallumon, az előző fejezetben ismertetett numerikus módszerek valamelyikével.) Ezután a (7.26) képlet alapján meghatározzuk a λ paraméter értékét, és végezetül a (7.21) feladat megoldását az $u_1(x)$ és $u_2(x)$ előzőekben már kiszámolt közelítéseiből a (7.24) összefüggés alapján határozzuk meg.

7.2. Megjegyzés Felmerülhet a kérdés: vajon mindig kifejezhető-e λ a (7.26) képlettel? Érvényes a következő állítás [8].

7.3. Tétel *Tegyük fel, hogy a (7.21) lineáris peremérték-feladatnak létezik egyértelmű megoldása. Ekkor*

- *vagy a (7.22) tulajdonságú $u_1(x)$ függvény a megoldás,*
- *vagy $u_1(b) - u_2(b) \neq 0$, és így a (7.26) képletből λ meghatározható.*

Tehát korrekt kitűzésű feladatok esetén a numerikus megoldás is meghatározható.

A (7.21) lineáris peremérték-feladat egyértelműen létező megoldását elő tudjuk állítani nemcsak a (7.27) és a (7.28) Cauchy-feladatok lineáris kombinációjaként, hanem más Cauchy-feladatok megoldásainak kombinációjaként is. Tekintsünk a következő két Cauchy-feladatot:

$$\begin{aligned} u'' =& p(x)u' + q(x)u + r(x), \quad x \in [a,b], \\ & u(a) = \alpha, \quad u'(a) = 0, \end{aligned} \tag{7.29}$$

valamint

$$\begin{aligned} u'' =& p(x)u' + q(x)u, \quad x \in [a,b], \\ & u(a) = 0, \quad u'(a) = 1. \end{aligned} \tag{7.30}$$

Jelölje ismét a (7.29) megoldását $u_1(x)$, a (7.30) feladat megoldását pedig $u_2(x)$. Könnyen láthatóan ekkor a

$$w(x) = u_1(x) + \frac{\beta - u_1(b)}{u_2(b)} u_2(x)$$

függvény megoldása a (7.21) feladatnak. Így elegendő numerikusan megoldani a (7.29) és a (7.30) Cauchy-feladatokat.

Összefoglalóan tehát megállapíthatjuk, hogy a lineáris peremérték-feladatok numerikus megoldása során ténylegesen nem alkalmazzuk a belövéses módszert, mivel a megoldást elő tudjuk állítani két, egyszerű struktúrájú, másodrendű közönséges differenciálegyenlet Cauchy-feladatának megoldásával. Ez, összehasonlítva az általános esettel, a számítási költségek tekintetében jelentős megtakarítást jelent: míg a lineáris esetben összesen kétszer kell valamely egylépéses kezdetiérték-feladatot megoldó numerikus módszert alkalmazni, addig az általános esetben annyiszor, ahány iterációs lépést hajtunk végre a (7.7) egyenlet közelítő megoldására.

7.2. A peremérték-feladat numerikus megoldása véges differenciák módszerével

Az előző szakaszban láttuk, hogy a belövéses módszer alkalmazása meglehetősen munkaigényes: minden egyes függvénykiértékelés egy Cauchy-feladat numerikus megoldását

igényli. Ebben a szakaszban más megközelítést alkalmazunk: a belövéses módszertől eltérően, a véges differenciás módszer alkalmazása során nem egy kezdetiérték-feladatra való visszavezetéssel oldjuk meg az

$$\begin{aligned} u'' =& f(x,u,u'), \quad x \in (a,b), \\ & u(a)=\alpha, \quad u(b)=\beta \end{aligned} \tag{7.31}$$

feladatot, hanem a megoldási tartományon rácshálót határozunk meg, és a feladatban szereplő deriváltakat ezen a rácshálón approximáljuk. Ezután ennek segítségével állítjuk elő a közelítő megoldást.

7.2.1. Véges differenciás approximáció

A továbbiakban felhasználjuk a megfelelően sima függvényekre jól ismert alábbi összefüggéseket:

$$u'(x) = \frac{u(x+h)-u(x)}{h} - \frac{1}{2}hu''(\zeta_1) \tag{7.32}$$

$$u'(x) = \frac{u(x)-u(x-h)}{h} + \frac{1}{2}hu''(\zeta_2) \tag{7.33}$$

$$u'(x) = \frac{u(x+h)-u(x-h)}{2h} - \frac{1}{6}h^2u'''(\zeta_3) \tag{7.34}$$

$$u''(x) = \frac{u(x+h)-2u(x)+u(x-h)}{h^2} - \frac{1}{12}h^2u^{(4)}(\zeta_4), \tag{7.35}$$

ahol ζ_i az x pont h sugarú környezetéből vett érték, azaz az $i = 1,2,3,4$ indexekre $\zeta_i \in (x-h, x+h)$.[4] Ez adja azt az ötletet egy (megfelelően sima) $u(x)$ függvény egy rögzített x_i pontbeli deriváltjainak közelítésére.

Jelöljünk ki az $[a,b]$ intervallumon egy rácshálót! Az egyszerűség kedvéért legyen ez az

$$\overline{\omega}_h = \{x_i = a+ih: \quad i = 0,1\ldots,N+1,\ h=(b-a)/(N+1)\} \tag{7.36}$$

ekvidisztáns rácsháló, ahol h a *rácsháló lépésköze*. Jelölje ω_h a *rácsháló belső pontjait*, azaz

$$\omega_h = \{x_i = a+ih: \quad i = 1,2,\ldots,N,\ h=(b-a)/(N+1)\}, \tag{7.37}$$

továbbá γ_h a rácsháló határpontjait, azaz

$$\gamma_h = \{x_0 = a,\ x_{N+1} = b\}. \tag{7.38}$$

[4] A (7.32)–(7.35) összefüggések az $u(x+h)$ és az $u(x-h)$ függvények x pontbeli Taylor-polinomjával és a Lagrange-féle maradéktaggal közvetlenül beláthatók.

Tegyük fel, hogy a (7.31) feladatnak létezik egyértelmű $u(x)$ megoldása, azaz $u(x) \in C^2[a,b]$ olyan függvény[5], amelyre

$$\begin{aligned} u''(x) =& f(x, u(x), u'(x)), \quad x \in (a,b), \\ & u(a) = \alpha, \quad u(b) = \beta. \end{aligned} \tag{7.39}$$

Ezért az $\overline{\omega}_h$ rácsháló pontjaiban felírva a fenti összefüggést, az

$$\begin{aligned} u''(x_i) =& f(t_i, u(x_i), u'(x_i)), \quad i = 1, 2, \ldots, N \\ & u(x_0) = \alpha, \quad u(x_{N+1}) = \beta \end{aligned} \tag{7.40}$$

összefüggések állnak fenn. Alkalmazva a (7.40) azonosságban a deriváltakra nyert (7.32)–(7.35) szerinti közelítéseket, az ismeretlen y_i értékekre (amelyek szándékaink szerint az $u(x_i)$ közelítései), az alábbi egyenleteket nyerjük:

$$\begin{aligned} \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} =& f(x_i, y_i, \frac{y_{i+1} - y_i}{h}), \quad i = 1, 2, \ldots, N \\ & y_0 = \alpha, \quad y_{N+1} = \beta. \end{aligned} \tag{7.41}$$

Megoldva a fenti $N+2$ ismeretlenes egyenletrendszert az $y_0, y_1, \ldots, y_{N+1}$ ismeretlenekre, meghatározhatjuk a véges differenciás közelítéseket. Ugyanakkor, valójában nem tudjuk a választ a következő kérdésekre.

- A (7.41) rendszernek létezik-e egyértelmű megoldása?
- Ha igen, hogyan oldható meg a rendszer hatékonyan?
- Közel van-e y_i értéke az $u(x_i)$ értékéhez?
- Ha sűrítjük a rácsházlót (azaz, ha h nullához tart), akkor az $[a,b]$ intervallumon a közelítő megoldások sorozata valamely értelemben tart-e a pontos megoldáshoz?

7.1. Megjegyzés A (7.41) diszkretizációban az első deriváltat a (7.32) formulájával helyettesítettük. Ha a (7.33) formulával helyettesítjük, akkor az

$$\begin{aligned} \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} =& f(x_i, y_i, \frac{y_i - y_{i-1}}{h}), \quad i = 1, 2, \ldots, N \\ & y_0 = \alpha, \quad y_{N+1} = \beta, \end{aligned} \tag{7.42}$$

ha pedig a (7.34) formulával, akkor az

$$\begin{aligned} \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} =& f(t_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h}), \quad i = 1, 2, \ldots, N \\ & y_0 = \alpha, \quad y_{N+1} = \beta \end{aligned} \tag{7.43}$$

feladatot kapjuk.

[5]Mint az a (7.39) összefüggésből látszik, elegendő az $u(x) \in C^2(a,b) \cap C[a,b]$ simasági feltétel, de a gyakorlati esetekben ez az egyszerűbben ellenőrizhető feltétel nem jelent tényleges megszorítást.

Vezessük be a következő jelöléseket:

$$y_{x,i} = \frac{y_{i+1} - y_i}{h}; \quad y_{\bar{x},i} = \frac{y_i - y_{i-1}}{h}; \quad y_{x^\circ,i} = \frac{y_{i+1} - y_{i-1}}{2h}, \tag{7.44}$$

amelyeket rendre jobb, bal és középponti differenciáknak nevezünk. A (7.32) és a (7.33) alapján nyilvánvalóan a jobb és bal oldali differenciák elsőrendben, míg a (7.34) következtében a középponti differencia másodrendben approximálja az első deriváltat az $x = x_i$ pontban.
Könnyen látható, hogy

$$y_{\bar{x}x,i} = \frac{y_{x,i} - y_{\bar{x},i}}{h} = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}, \tag{7.45}$$

és ezért a (7.35) összefüggés miatt $y_{\bar{x}x,i}$ másodrendben approximálja az $u''(x)$ függvényt a $x = x_i$ pontban. Ezen jelölésekkel az egyes sémák a következő alakot öltik:

- A (7.41) séma:

$$\begin{aligned} y_{\bar{x}x,i} =& f(x_i, y_i, y_{x,i}), \quad i = 1, 2, \ldots, N \\ & y_0 = \alpha, \quad y_{N+1} = \beta. \end{aligned} \tag{7.46}$$

- A (7.42) séma:

$$\begin{aligned} y_{\bar{x}x,i} =& f(x_i, y_i, y_{\bar{x},i}), \quad i = 1, 2, \ldots, N \\ & y_0 = \alpha, \quad y_{N+1} = \beta. \end{aligned} \tag{7.47}$$

- A (7.43) séma:

$$\begin{aligned} y_{\bar{x}x,i} =& f(x_i, y_i, y_{y_{x^\circ,i}}), \quad i = 1, 2, \ldots, N \\ & y_0 = \alpha, \quad y_{N+1} = \beta. \end{aligned} \tag{7.48}$$

7.2.2. Lineáris peremérték-feladatok approximációja véges differenciák módszerével

Ebben a szakaszban választ adunk az előzőekben megfogalmazott kérdésekre a

$$\begin{aligned} u'' =& p(x)u' + q(x)u + r(x), \quad x \in [a, b], \\ & u(a) = \alpha, \quad u(b) = \beta \end{aligned} \tag{7.49}$$

lineáris peremérték-feladat véges differenciás numerikus megoldására vonatkozóan. Mint azt megmutattuk, az egyértelmű megoldhatósághoz feltételezzük, hogy $p, q, r \in C[a, b]$ és $q(x) > 0$ az $[a, b]$ intervallumon, azaz $\min_{[a,b]} q := q_{\min} > 0$. Vezessük be a $p_i = p(x_i), q_i = q(x_i), r_i = r(x_i)$ jelöléseket.

7.2. Tétel *A (7.49) lineáris peremérték-feladat (7.46)-(7.48) véges differenciás diszkretizációja*

$$\begin{gathered} a_i y_{i-1} + d_i y_i + c_i y_{i+1} = -r_i, \quad i = 1, 2, \ldots, N \\ y_0 = \alpha, \quad y_{N+1} = \beta \end{gathered} \tag{7.50}$$

alakú lineáris algebrai egyenletrendszert eredményez, ahol megfelelően kis h esetén mindhárom diszkretizációra érvényes a

$$|d_i| - |a_i| - |c_i| = q_i \tag{7.51}$$

egyenlőség.

Bizonyítás. A (7.49) feladatra a (7.46) diszkretizáció az

$$\begin{gathered} \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = p_i \frac{y_{i+1} - y_i}{h} + q_i y_i + r_i, \quad i = 1, 2, \ldots, N, \\ y_0 = \alpha, \quad y_{N+1} = \beta \end{gathered} \tag{7.52}$$

lineáris algebrai egyenletrendszert jelenti, ami az

$$a_i = -\frac{1}{h^2}, \; d_i = \frac{2}{h^2} + q_i - \frac{1}{h} p_i, \; c_i = -\frac{1}{h^2} + \frac{1}{h} p_i \tag{7.53}$$

megválasztással a (7.50) feladatot jelenti. Megfelelően kis h esetén ezért

$$|d_i| - |a_i| - |c_i| = \frac{1}{h^2} \left((2 + h^2 q_i - h p_i) - 1 - (1 - h p_i) \right) = q_i. \tag{7.54}$$

Könnyen látható, hogy a (7.47) diszkretizáció esetén

$$a_i = -\frac{1}{h^2} - \frac{1}{h} p_i, \; d_i = \frac{2}{h^2} + q_i + \frac{1}{h} p_i, \; c_i = -\frac{1}{h^2}, \tag{7.55}$$

míg a (7.48) diszkretizáció esetén

$$a_i = -\frac{1}{h^2} - \frac{1}{2h} p_i, \; d_i = \frac{2}{h^2} + q_i, \; c_i = -\frac{1}{h^2} + \frac{1}{2h} p_i, \tag{7.56}$$

és mindkét esetben alkalmasan megválasztott kis h értékek mellett (7.51) érvényes. □

7.3. Következmény Mivel a (7.50) rendszer felírható

$$\begin{aligned} d_1 y_1 + c_1 y_2 &= -r_1 - a_1 \alpha, \\ a_i y_{i-1} + d_i y_i + c_i y_{i+1} &= -r_i, \quad i = 2, 3, \ldots, N-1, \\ a_N y_{N-1} + d_N y_N &= -r_N - c_N \beta \end{aligned} \tag{7.57}$$

alakban, ezért a diszkretizált feladat valójában egy N ismeretlenes lineáris algebrai egyenletrendszert jelent, amelynek együtthatómátrixa tridiagonális, szigorúan diagonálisan domináns mátrix.

Jelölje e_i az $x = x_i$ pontban a pontos és a közelítő megoldás eltérését (azaz $e_i = u_i - y_i$), és $\mathbf{e}_h$ az ω_h rácsháló pontjaiban értelmezett hibafüggvényt (azaz $\mathbf{e}_h(x_i) = e_i$).

7.4. Definíció *Valamely numerikus módszer* konvergenciája *azt jelenti, hogy az általa generált numerikus megoldásra a rácsháló finomításával a hibafüggvény a rácshálón értelmezett függvények terének valamely normájában nullához tart, azaz*

$$\lim_{h\to 0} \boldsymbol{e}_h = 0. \tag{7.58}$$

Ha $\boldsymbol{e}_h = \mathcal{O}(h^p)$, *akkor a numerikus módszert* p*-ed rendű módszernek nevezzük.*

Megjegyezzük, hogy mi a továbbiakban a maximumnormában fogjuk vizsgálni a konvergenciát.

A következő állításban a fenti véges differenciás közelítések konvergenciájával és a konvergencia rendjével foglalkozunk.

7.5. Tétel *A (7.53), (7.55) és (7.56) megválasztású (7.50) alakú véges differenciás diszkretizációk lineáris peremérték-feladat esetén konvergensek, emellett*

1. *a (7.56) megválasztás esetén a séma másodrendben,*
2. *a (7.53) és a (7.55) megválasztás esetén pedig a sémák elsőrendben*

konvergálnak a (7.49) feladat megoldásához.

Bizonyítás. Mivel a két állítás bizonyítása lényegében megegyezik, a továbbiakban csak az elsőt mutatjuk meg.
Felhasználva a (7.34) és a (7.35) approximációs tulajdonságokat, a (7.49) feladat a $t = t_i$ pontban így írható fel:

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - \frac{1}{12}h^2 u^{(4)}(\zeta_4^{(i)}) = p_i \left(\frac{u_{i+1} - u_{i-1}}{2h} - \frac{1}{6}h^2 u'''(\zeta_3^{(i)}) \right) + q_i u_i + r_i. \tag{7.59}$$

Átrendezve a (7.59) egyenlőséget a következő alakot kapjuk:

$$\left(-\frac{1}{h^2} + \frac{p_i}{2h} \right) u_{i+1} + \left(\frac{2}{h^2} + q_i \right) u_i + \left(-\frac{1}{h^2} - \frac{p_i}{2h} \right) u_{i-1} = -r_i - h^2 g_i, \tag{7.60}$$

ahol

$$g_i = \frac{1}{12}u^{(4)}(\zeta_4^{(i)}) - \frac{p_i}{6}u'''(\zeta_3^{(i)}).$$

Ezért a (7.56) jelöléseivel (7.60) felírható az

$$a_i u_{i-1} + d_i u_i + c_i u_{i+1} = -r_i - h^2 g_i, \tag{7.61}$$

alakban. Mivel a numerikus megoldásra érvényes a (7.50) összefüggés, (ahol az együtthatók (7.56) szerintiek), a (7.61) egyenlőségből kivonva a (7.50) egyenleteket, az $\mathbf{e}_h$ hibavektor koordinátái kielégítik az

$$\begin{gathered} a_i e_{i-1} + d_i e_i + c_i e_{i+1} = -h^2 g_i, \quad i = 1, 2, \ldots, N \\ e_0 = e_{N+1} = 0 \end{gathered} \tag{7.62}$$

alakú lineáris algebrai egyenletrendszert. Az i-edik egyenletet átrendezve:

$$d_i e_i = -a_i e_{i-1} - c_i e_{i+1} - h^2 g_i, \tag{7.63}$$

azaz abszolút értékben becsülve a

$$|d_i||e_i| \leq |a_i||e_{i-1}| + |c_i||e_{i+1}| + h^2|g_i| \leq (|a_i| + |c_i|)\,\|\mathbf{e}_h\|_\infty + h^2\|\mathbf{g}\|_\infty \tag{7.64}$$

egyenlőtlenséget nyerjük, ahol $\mathbf{g}$ jelöli a g_i koordinátájú vektort. Jelölje i_0 azt az indexet, amelyre $\|\mathbf{e}_h\|_\infty = |e_{i_0}|$. (Nyilvánvalóan $i_0 \neq 0$ és $i_0 \neq N+1$, mivel $e_0 = e_{N+1} = 0$.) Mivel (7.64) minden $i = 1, 2, \ldots N$ értékre érvényes, ezért az $i = i_0$ indexre is, azaz

$$|d_{i_0}|\|\mathbf{e}_h\|_\infty \leq (|a_{i_0}| + |c_{i_0}|)\,\|\mathbf{e}_h\|_\infty + h^2\|\mathbf{g}\|_\infty. \tag{7.65}$$

Átrendezve a (7.65) egyenlőtlenséget, a

$$(|d_{i_0}| - |a_{i_0}| - |c_{i_0}|)\,\|\mathbf{e}_h\|_\infty \leq h^2\|\mathbf{g}\|_\infty \tag{7.66}$$

egyenlőtlenséghez jutunk. Ez a (7.51) tulajdonság figyelembevételével a

$$q_{i_0}\|\mathbf{e}_h\|_\infty \leq h^2\|\mathbf{g}\|_\infty \tag{7.67}$$

becslést jelenti. A $\min_{[a,b]} q := q_{\min} > 0$ feltétel következtében

$$\|\mathbf{e}_h\|_\infty \leq h^2 \frac{\|\mathbf{g}\|_\infty}{q_{\min}} \leq \tilde{C}h^2, \tag{7.68}$$

ahol

$$\tilde{C} = \left(\frac{M_4}{12} + \frac{p_{\max} M_3}{6}\right)/q_{\min}, \quad M_j = \max_{[a,b]} |u^{(j)}|, \quad p_{\max} = \max_{[a,b]} |p|.$$

Mindez azt jelenti, hogy $h \to 0$ esetén a hiba másodrendben tart nullához. □

7.6. Példa *Mutassuk meg a fenti tétel második állítását a fenti bizonyítás felhasználásával!*

A 7.5. tétel arra az esetre vonatkozik, amikor a (7.49) feladatban $q(x) > 0$ az $[a, b]$ intervallumon. Így nem ad választ a (6.6) feladat $p = q = 0$ speciális esetére.

Tekintsük tehát az

$$\begin{aligned} u'' &= r(x), \quad x \in (0, l), \\ u(0) &= \alpha, \quad u(l) = \beta \end{aligned} \tag{7.69}$$

peremérték-feladatot. (Az egyszerűség kedvéért az intervallumot l hosszúságúnak tekintjük, és a bal oldali végpontot az origóba helyezzük.) Ebben az esetben tehát az $\overline{\omega}_h$ rácsháló

$$\overline{\omega}_h = \{x_i = ih : \quad i = 0, 1, \ldots, N+1, \ h = l/(N+1)\} \tag{7.70}$$

alakú. Ekkor a (7.50) rendszer mindegyik diszkretizáció esetén a (7.57) feladatot eredményezi, ahol

$$a_i = -\frac{1}{h^2}, \ d_i = \frac{2}{h^2}, \ c_i = -\frac{1}{h^2}. \tag{7.71}$$

Ezért a diszkrét feladat az

$$\mathbf{A}_h \mathbf{y}_h = \mathbf{b}_h \tag{7.72}$$

alakú lineáris algebrai egyenletrendszert jelenti, ahol az $\mathbf{A}_h \in \mathbb{R}^{N \times N}$ mátrix

$$\mathbf{A}_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & 0 & -1 & 2 \end{pmatrix} \tag{7.73}$$

alakú, és a $\mathbf{b}_h \in \mathbb{R}^N$ vektorra pedig

$$\mathbf{b}_h = \begin{pmatrix} -r(x_1) + \alpha/h^2 \\ -r(x_i), \ i = 2, 3, \ldots, N-1 \\ -r(x_N) + \beta/h^2 \end{pmatrix}. \tag{7.74}$$

Először megmutatjuk, hogy a fenti diszkretizáció korrekt kitűzésű feladatot eredményez.

7.7. Tétel *A (7.72) feladatnak tetszőleges $h > 0$ esetén létezik egyértelmű megoldása.*

Bizonyítás. Azt kell megmutatnunk, hogy bármely $h > 0$ mellett az $\mathbf{A}_h$ mátrix reguláris, azaz a $\lambda = 0$ nem sajátértéke. Felhasználva a sajátértékekre vonatkozó (2.165) formulát, az $\mathbf{A}_h$ mátrix sajátértékei a következők:

$$\lambda_k = \frac{4}{h^2} \sin^2 \frac{k\pi h}{2l}, \quad k = 1, 2, \ldots, N. \tag{7.75}$$

A legkisebb sajátérték a $k = 1$ indexhez tartozó érték, amelyre érvényes

$$\inf_{h>0} \lambda_1(h) = \lambda_1(l/2) = \frac{16}{l^2} \sin^2 \frac{\pi}{4} = \frac{8}{l^2} \tag{7.76}$$

egyenlőség. (V.ö. a (2.167) képlettel.) Így minden $h > 0$ esetén

$$\lambda_1(h) \geq \frac{8}{l^2}, \tag{7.77}$$

vagyis az $\mathbf{A}_h$ mátrix legkisebb sajátértéke minden $h > 0$ esetén a $\delta := 8/l^2 > 0$ számnál nagyobb. Ezzel beláttuk állításunkat. □

7.8. Következmény A fenti tételből és a bizonyításából közvetlenül következnek az alábbi állítások.

1. Az $\mathbf{A}_h$ mátrix invertálható, és $\|\mathbf{A}_h^{-1}\|_2 \leq 1/\delta = l^2/8$.

2. Az $\mathbf{A}_h$ mátrix szimmetrikus, szigorúan pozitív definit.

3. A (2.166) formulából, azaz a k-ik sajátvektor

$$\mathbf{v}_k = \left(\sin \frac{k\pi i h}{l} \right)_{i=1}^{N}, \quad k = 1, 2, \ldots, N \tag{7.78}$$

alakjából látható, hogy $\mathbf{v}_1 > 0$. Mivel $\lambda_1(h) > 0$ is igaz, ezért az $\mathbf{A}_h \mathbf{v}_1 = \lambda_1(h)\mathbf{v}_1$ összefüggés alapján igaz, hogy a $\mathbf{v}_1 > 0$ vektorral $\mathbf{A}_h \mathbf{v}_1 > 0$, azaz $\mathbf{A}_h$ M-mátrix, és így egyben reguláris is [2].

Eredményünket felhasználva közvetlenül belátható a (7.57), (7.71) numerikus módszer konvergenciája az $\overline{\omega}_h := \{x_1, x_2, \ldots, x_N\}$, h lépésközű ekvidisztáns rácshálón értelmezett w_h rácsfüggvények

$$\|w_h\|_{2,h}^2 := h \sum_{i=1}^{N} u_i^2 \tag{7.79}$$

alakú normájában.[6]

Megjegyezzük, hogy ekkor a megfelelő mátrixnormára

$$\|\mathbf{A}\|_{2,h} = \sup_{\mathbf{v} \in \mathbb{R}^N} \frac{\|\mathbf{A}\mathbf{v}\|_{2,h}}{\|\mathbf{v}\|_{2,h}} = \sup_{\mathbf{v} \in \mathbb{R}^N} \frac{h\|\mathbf{A}\mathbf{v}\|_2}{h\|\mathbf{v}\|_2} = \|\mathbf{A}\|_2,$$

[6]Könnyen látható, hogy az $a = x_1$ és $b = x_N$ rögzített értékek esetén ez a norma az $L_2[a, b]$ tér normájának diszkrét analógja, azaz ha w az $[a, b]$ intervallumon értelmezett, négyzetesen integrálható függvény, és $u_h(x_i) = u(x_i)$, akkor $\lim_{h \to 0} \|u_h\|_{2,h}^2 = \|u\|_{L_2[a,b]}^2 \equiv \int_a^b |u|^2(x)dx$.

tehát a szokásos spektrálnormával számolható.

Tekintsük tehát az

$$\begin{gathered}\frac{y_{i-1}-2y_i+y_{i+1}}{h^2}=r_i,\quad i=1,2,\dots,N\\ y_0=\alpha,\quad y_{N+1}=\beta\end{gathered}\tag{7.80}$$

feladatot, és mutassuk meg, hogy a megoldása a $\|\cdot\|_{2,h}$ normában $h\to 0$ esetén tart a (7.69) peremérték-feladat megoldásához. Jelölje ismét $e_i=u(t_i)-y_i$ a hibafüggvényt. Ekkor $y_i=-e_i+u(x_i)$, és ezt behelyettesítve a (7.80) sémába a hibafüggvényre a

$$\begin{gathered}\frac{-e_{i-1}+2e_i-e_{i+1}}{h^2}=r_i-\frac{u(x_{i-1})-2u(x_i)+u(x_{i+1})}{h^2}\equiv\psi_i,\quad i=1,2,\dots,N\\ e_0=0,\quad e_{N+1}=0\end{gathered}\tag{7.81}$$

feladatot kapjuk. A (7.81) feladat felírható

$$\mathbf{A}_h\mathbf{e}_h=\boldsymbol{\psi}_h\tag{7.82}$$

alakú lineáris algebrai egyenletrendszerként, ahol $\mathbf{A}_h\in\mathbb{R}^{N\times N}$ mátrix (7.73) alakú, a $\boldsymbol{\psi}_h\in\mathbb{R}^N$ vektorra pedig $(\boldsymbol{\psi}_h)_i=\psi_i$, $i=1,2,\dots,N$. A (7.82) egyenletből

$$\mathbf{e}_h=\mathbf{A}_h^{-1}\boldsymbol{\psi}_h,\tag{7.83}$$

azaz

$$\|\mathbf{e}_h\|_{2,h}\le\|\mathbf{A}_h^{-1}\|_{2,h}\|\boldsymbol{\psi}_h\|_{2,h}=\|\mathbf{A}_h^{-1}\|_2\|\boldsymbol{\psi}_h\|_{2,h}.\tag{7.84}$$

Mivel (7.35) következtében $\psi_i=\mathcal{O}(h^2)$, ezért az $hN<l$ következtében

$$\|\boldsymbol{\psi}_h\|_{2,h}^2=h\sum_{i=1}^N\psi_i^2=h\sum_{i=1}^N\mathcal{O}(h^4)=hN\mathcal{O}(h^4)=\mathcal{O}(h^4).$$

Ezért tehát $\|\boldsymbol{\psi}_h\|_{2,h}=\mathcal{O}(h^2)$. Másrészt, a 7.8. következmény szerint $\|\mathbf{A}_h^{-1}\|_2\le l^2/8$. Így (7.84) alapján

$$\|\mathbf{e}_h\|_{2,h}=\mathcal{O}(h^2).\tag{7.85}$$

Ezzel beláttuk az alábbi állítást.

7.9. Tétel *A (7.80) séma megoldása $h\to 0$ esetén másodrendben tart a $\|\cdot\|_{2,h}$ normában a (7.69) peremérték-feladat megoldásához.*

A fenti séma konvergenciája az alábbi animáción látható:
../animaciok/pef.gif

Sőt, az egyenletet megoldó program le is tölthető, az alábbi linkről:
../programok/fdm.exe

7.3. A lineáris peremérték-feladat numerikus megoldásának általános vizsgálata

A korábbiakban megvizsgáltuk a lineáris peremérték-feladatok véges differenciák módszeres megoldását adott feladatra. Ebben a szakaszban ismételten a lineáris peremérték-feladatok numerikus megoldásával és a közelítések konvergenciájával foglalkozunk, de most az eredményeinket általánosan fogalmazzuk meg. Megmutatjuk, hogy ezen általános megközelítésből speciális esetként az előző szakaszban leírt eredményeink hogyan nyerhetők.[7]

Jelölje L_0 azt az operátort, amely a $C^2[a,b]$-beli függvényeken van értelmezve, és

$$L_0 u = -u'' + p(x)u' + q(x)u, \tag{7.86}$$

valamint B_1, B_2 azokat a szintén $C^2[a,b]$-beli függvényeken értelmezett operátorokat, amelyekre

$$B_1 u = u(a), \quad B_2 u = u(b). \tag{7.87}$$

Ekkor $L_0 : C^2[a,b] \to C[a,b]$, $B_1, B_2 : C^2[a,b] \to \mathbb{R}$ típusú lineáris operátorok. Jelölje

$$\mathbb{F}_1[a,b] = C^2[a,b], \quad \mathbb{F}_2[a,b] = C[a,b] \times \mathbb{R} \times \mathbb{R} \tag{7.88}$$

tereket, valamint L azt az operátort, amely az $\mathbb{F}_1[a,b]$ halmazon van értelmezve, és

$$Lv = \begin{pmatrix} L_0 v \\ B_1 v \\ B_2 v \end{pmatrix}. \tag{7.89}$$

Tehát L egy $\mathbb{F}_1[a,b] \to \mathbb{F}_2[a,b]$ lineáris operátor, amely egy $v \in C^2[a,b]$ függvényhez az

$$Lv = \begin{pmatrix} -v'' + p(x)v' + q(x)v \\ v(a) \\ v(b) \end{pmatrix} \in \mathbb{F}_2[a,b] \tag{7.90}$$

hármast rendeli hozzá.
Ekkor a (7.49) feladat felírható a következő módon. Az

$$f := \begin{pmatrix} -r(x) \\ \alpha \\ \beta \end{pmatrix} \in \mathbb{F}_2[a,b] \tag{7.91}$$

jelölés mellett keressük azon $u \in F_1[a,b]$ elemet (azaz $C^2[a,b]$-beli függvényt), amelyre

$$Lu = f. \tag{7.92}$$

[7] Ezt a részt elsősorban az érdeklődő Olvasónak szánjuk.

Jelölje $\mathbb{F}(\overline{\omega}_h)$ illetve $\mathbb{F}(\omega_h)$ az $\overline{\omega}_h$ illetve ω_h rácsháló\-kon értelmezett rácsfüggvények vektorterét, $L_{0,h} : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\omega_h)$ pedig azt az operátort, amely valamely $v_h \in \mathbb{F}(\overline{\omega}_h)$ függvény esetén a következő módon hat:

$$\left(L_{0,h}^{(1)} v_h\right)(x) = -\frac{v_h(x+h) - 2v_h(x) + v_h(x-h)}{h^2} + p_i \frac{v_h(x+h) - v_h(x)}{h} + q_i v_h(x), \quad x \in \omega_h. \tag{7.93}$$

Jelölje továbbá $B_{1,h}, B_{2,h}$ azokat az $\mathbb{F}(\overline{\omega}_h)$-beli rácsfüggvényeken értelmezett operátorokat, amelyekre

$$B_{1,h} v_h = v_h(x_0) = a, \quad B_{1,h} v_h = v_h(x_{N+1}) = b. \tag{7.94}$$

Ekkor tehát $L_{0,h}^{(1)} : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\omega_h)$, $B_{1,h}, B_{2,h} : \mathbb{F}(\overline{\omega}_h) \to \mathbb{R}$ típusú lineáris operátorok. Jelölje $L_h^{(1)}$ azt az operátort, amely az $\mathbb{F}(\overline{\omega}_h)$ halmazon van értelmezve és

$$L_h^{(1)} v_h = \begin{pmatrix} L_{0,h}^{(1)} v_h \\ B_{1,h} v_h \\ B_{2,h} v_h \end{pmatrix}. \tag{7.95}$$

Tehát $L_h^{(1)}$ egy $\mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\omega_h) \times \mathbb{R} \times \mathbb{R} \equiv \mathbb{F}(\overline{\omega}_h)$ típusú lineáris operátor, amely a v_h ($\overline{\omega}_h$ rácsháló pontjaiban definiált) rácsfüggvényhez az

$$L_h^{(1)} v_h = \begin{pmatrix} -\dfrac{v_h(x+h) - 2v_h(x) + v_h(x-h)}{h^2} + p_i \dfrac{v_h(x+h) - v_h(x)}{h} + q_i v_h(x), \\ v_h(a) \\ v_h(b) \end{pmatrix}$$

(ahol $x \in \omega_h$) $\mathbb{F}(\overline{\omega}_h)$-beli, ugyancsak a $\overline{\omega}_h$ rácsháló pontjaiban definiált rácsfüggvényt rendeli hozzá.

Legyen r_h az az $\mathbb{F}(\omega_h)$-beli rácsfüggvény, amelyre $r_h(x) = r(x)$ mindegyik $x \in \omega_h$ pontban. Ekkor a (7.52) feladat felírható a következő módon. Az

$$f_h := \begin{pmatrix} -r_h \\ \alpha \\ \beta \end{pmatrix} \in \mathbb{F}(\overline{\omega}_h) \tag{7.96}$$

jelölés mellett keressük azon $v_h \in \mathbb{F}(\overline{\omega}_h)$ elemet (vagyis az $\overline{\omega}_h$ rácson értelmezett rácsfüggvényt), amelyre

$$L_h^{(1)} v_h = f_h. \tag{7.97}$$

Vegyük észre, hogy (7.97) – a jelöléseinket figyelembe véve – egy lineáris algebrai egyenletrendszert jelent, amelynek mérete megegyezik az $\overline{\omega}_h$ rácsháló pontjainak a számával, azaz

$N+2$ ismeretlent tartalmaz. Tehát az ismeretlen v_h illetve az adott f_h rácsfüggvényeket $N+2$ dimenziós vektorokkal, míg az $L_h^{(1)}$ lineáris operátort egy $(N+2)\times(N+2)$ dimenziós mátrix segítségével adhatjuk meg. Ez azt jelenti, hogy a (7.97) feladat felírható

$$\overline{\mathbf{A}}_h^{(1)}\overline{\mathbf{y}}_h = \bar{\mathbf{f}}_h \tag{7.98}$$

alakban, ahol $\overline{\mathbf{A}}_h^{(1)} \in \mathbb{R}^{(N+2)\times(N+2)}$ és $\overline{\mathbf{y}}_h, \bar{\mathbf{f}}_h \in \mathbb{R}^{N+2}$. Mivel a vektorok i-edik koordinátája a $x_i \in \overline{\omega}_h$ rácsponthoz tartozó értékeket jelenti, ezért a (7.98) egyenletben

$$(\overline{\mathbf{y}}_h)_i = v_h(t_i), \quad (\bar{\mathbf{f}}_h)_i = f_h(t_i), \quad i = 0, 1, \ldots, N+1, \tag{7.99}$$

és (7.91) alapján az $\overline{\mathbf{A}}_h^{(1)}$ mátrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ a_1 & d_1 & c_1 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & a_2 & d_2 & c_2 & \ldots & 0 & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & 0 & a_N & d_N & c_N \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 1 \end{pmatrix} \tag{7.100}$$

alakú, ahol a_i, d_i, c_i értékei a (7.53) képlet szerintiek.

7.1. Példa *Jelölje $L_{0,h}^{(2)}, L_{0,h}^{(3)} : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\omega_h)$ azokat az operátorokat, amelyek egy $v_h \in \mathbb{F}(\overline{\omega}_h)$ függvényre a következő módon hatnak:*

$$\begin{aligned}\left(L_{0,h}^{(2)} v_h\right)(x) = &-\frac{v_h(x+h) - 2v_h(x) + v_h(x-h)}{h^2} + \\ &p_i \frac{v_h(x) - v_h(x-h)}{h} + q_i v_h(x), \quad x \in \omega_h.\end{aligned} \tag{7.101}$$

$$\begin{aligned}\left(L_{0,h}^{(3)} v_h\right)(x) = &-\frac{v_h(x+h) - 2v_h(x) + v_h(x-h)}{h^2} + \\ &p_i \frac{v_h(x+h) - v_h(x-h)}{2h} + q_i v_h(x), \quad x \in \omega_h.\end{aligned} \tag{7.102}$$

Jelölje

$$L_h^{(k)} v_h = \begin{pmatrix} L_{0,h}^{(k)} v_h \\ B_{1,h} v_h \\ B_{2,h} v_h \end{pmatrix}, \quad k = 2, 3. \tag{7.103}$$

Mutassuk meg, hogy $L_h^{(k)}$ ebben az esetben is felírható (a továbbiakban $\overline{\boldsymbol{A}}_h^{(k)}$-val jelölt) (7.100) alakú mátrix alakjában, ahol a_i, d_i, c_i értékei a (7.55) illetve a (7.56) képlet szerintiek.

Mivel $(\overline{\mathbf{y}}_h)_0$ és $(\overline{\mathbf{y}}_h)_{N+1}$ értékei ismertek (a peremfeltételek miatt ezek értéke α és β), ezért a (7.98) egyenlet ekvivalens az

$$\mathbf{A}_h^{(1)}\mathbf{y}_h = \mathbf{f}_h \tag{7.104}$$

feladattal, ahol $\mathbf{A}_h^{(1)} \in \mathbb{R}^{N\times N}$, $\mathbf{y}_h, \mathbf{f}_h \in \mathbb{R}^N$ és

$$\begin{gathered}(\mathbf{f}_h)_1 = (\bar{\mathbf{f}}_h)_1 - a_1\alpha, \quad (\mathbf{f}_h)_i = (\bar{\mathbf{f}}_h)_i,\ i = 2,3,\dots,N-1, \quad (\mathbf{f}_h)_N = (\bar{\mathbf{f}}_h)_N - c_N\beta,\\ (\mathbf{y}_h)_i = (\overline{\mathbf{y}}_h)_i,\ i = 1,2,\dots,N,\\ \mathbf{A}_h^{(1)} = \begin{pmatrix} d_1 & c_1 & 0 & \dots & 0 & 0 & 0\\ a_2 & d_2 & c_2 & \dots & 0 & 0 & 0\\ \dots & \dots & \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & \dots & 0 & a_N & d_N \end{pmatrix}.\end{gathered} \tag{7.105}$$

A (7.104)-(7.105) feladat megegyezik a (7.57) feladattal.

7.2. Megjegyzés A másik két diszkretizációra is az értelemszerűen definiált $\mathbf{A}_h^{(2)}$ és $\mathbf{A}_h^{(3)}$ mátrixokkal (7.104) érvényes.

A továbbiakban arra vagyunk kíváncsiak, hogy a fenti sémák valamelyikével definiált numerikus megoldások használhatók-e az eredeti peremérték-feladat megoldásának közelítésére.

Ezen szakasz elején megmutattuk a következőket.

- A 7.2. tétel alapján az alkalmasan megválasztott h értékek mellett az $\overline{\mathbf{A}}_h^{(k)}$ mátrixok szigorúan diagonálisan dominánsak, ezért regulárisak. Tehát megfelelően kis h esetén az $L_h^{(k)}v_h = f_h$ $(k = 1,2,3)$ feladatok mindegyike egyértelműen megoldható.
- A 7.5. tétel alapján az e_h hibafüggvény nullához tart, azaz a közelítő megoldások sorozata $h \to 0$ esetén tart a pontos megoldáshoz. Ennek belátása két lépésben történt:
  1. Felírtuk a (7.62) hibaegyenletet.
  2. Megmutattuk, hogy ezen rendszer megoldásai ($h \to 0$ esetén) nullához tartanak.

7.3. Megjegyzés Ezzel kapcsolatosan két dologra hívjuk fel a figyelmet.

1. A (7.62) hibaegyenlet jobb oldalán szereplő kifejezés a lokális approximációs hiba. Tehát a hibaegyenlet jobb oldalán egy olyan vektor áll, amelynek i-edik koordinátája azt mutatja meg, hogy a $t_i \in \omega_h$ pontban a rácsháló pontjaiban értelmezett megoldás $L_h^{(k)}$-képe milyen közel van a megoldás L-képének értékéhez.

2. A hibavektort a (7.62) rendszer megoldásával állítjuk elő. Ezért a globális hiba viselkedése (nevezetesen, nullához tartása) azon múlik, hogy a rendszer jobb oldalán szereplő vektor hogyan viselkednek a rendszer együtthatómátrixának inverzén, azaz $h \to 0$ esetén az $\left(\overline{\mathbf{A}}_h^{(k)}\right)^{-1}$ mátrixok hogyan hatnak a lokális approximációs hibavektorra.

A továbbiakban *általánosan tárgyaljuk* a diszkretizációs módszereket és azok konvergenciáját. Legyen $L_{0,h} : \mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\omega_h)$ egy adott lineáris operátor, azaz egy olyan leképezés, amely egy $\overline{\omega}_h$ rácson értelmezett rácsfüggvényhez valamilyen lineáris leképezési szabály szerint megfeleltet egy ω_h-n értelmezett másik rácsfüggvényt. Legyen L_h az ezen $L_{0,h}$ operátor segítségével definiált

$$L_h v_h = \begin{pmatrix} L_{0,h} v_h \\ B_{1,h} v_h \\ B_{2,h} v_h \end{pmatrix} \tag{7.106}$$

$\mathbb{F}(\overline{\omega}_h) \to \mathbb{F}(\overline{\omega}_h)$ típusú lineáris operátor. (Tehát L_h jelentheti az $L_h^{(k)}$ $(k = 1, 2, 3)$ operátorok valamelyikét, de jelenthet valamely más lineáris operátort is.) A továbbiakban az L_h operátornak megfeleltetett mátrixokat $\overline{\mathbf{A}}_h$ illetve $\mathbf{A}_h$-val jelöljük. Legyen $\mathcal{P}_h^{(2)}$ egy olyan leképezés, amely egy $\mathbb{F}_2[a,b] = C[a,b] \times \mathbb{R} \times \mathbb{R}$ elemnek megfeleltet egy $\mathbb{F}(\overline{\omega}_h)$-beli rácsfüggvényt. Jelölje

$$f_h := \mathcal{P}_h^{(2)} f \in \mathbb{F}(\overline{\omega}_h) \tag{7.107}$$

rácsfüggvényt, ahol $f \in \mathbb{F}_2[a,b]$ a (7.92) feladat jobb oldala.
A diszkretizációs eljárások általános tárgyalása során a következő kérdések megválaszolása szükséges.

1. Létezik-e az
$$L_h v_h = f_h \tag{7.108}$$
feladatnak egyértelmű megoldása?

2. Hogyan értelmezhető egy $\mathbb{F}(\overline{\omega}_h)$-beli rácsfüggvény és egy $F_1[a,b]$-beli függvény távolsága?

3. Milyen közel van a $v_h \in \mathbb{F}(\overline{\omega}_h)$ függvény a (7.92) egyenlet $u \in F_1[a,b]$ megoldásához?

4. A $h \to 0$ esetben milyen feltételek mellett tart a közelítő és a pontos megoldás távolsága nullához?

A továbbiakban ezekkel a kérdésekkel foglalkozunk.

1. Az általános tárgyalás során *mindig feltesszük*, hogy a (7.108) feladatnak létezik egyértelmű megoldása. Ugyanakkor minden konkrét L_h megválasztás esetén ezt *be kell bizonyítani*, azaz azt kell megmutatni, hogy az $\overline{\mathbf{A}}_h$ (avagy az $\mathbf{A}_h$) mátrix reguláris.

2. Ennél a kérdésnél az az alapvető probléma, hogy az $[a,b]$ intervallumon értelmezett $w \in F_1[a,b]$ és az $\overline{\omega}_h$ rácspontokban értelmezett $g_h \in \mathbb{F}(\overline{\omega}_h)$-beli rácsfüggvény *különböző alaphalmazokon* vannak értelmezve. Ezért jelöljön $\mathcal{P}_h^{(1)}$ egy $F_1[a,b] \to \mathbb{F}(\overline{\omega}_h)$ típusú lineáris leképezést. (Tehát $\mathcal{P}_h^{(1)}$ egy olyan leképezés, amely az intervallumon értelmezett függvényt leképezi egy, a rácspontokban értelmezett függvényre.) Ekkor a két függvény távolságán $\|\mathcal{P}_h^{(1)} w - g_h\|_h$ számot értjük, ahol $\|\cdot\|_h$ egy $\mathbb{F}(\overline{\omega}_h) = \mathbb{R}^{N+2}$ térbeli rögzített norma.

3. Legyen a $v_h \in \mathbb{F}(\overline{\omega}_h)$ függvény a (7.108) feladat megoldása. Jelölje u_h azt az $\mathbb{F}(\overline{\omega}_h)$-beli rácsfüggvényt, amelyet a (7.92) egyenlet $u \in F_1[a,b]$ megoldásának $\mathcal{P}_h^{(1)}$-képeként kapunk, azaz
$$u_h = \mathcal{P}_h^{(1)} u, \tag{7.109}$$
továbbá $e_h \in \mathbb{F}(\overline{\omega}_h)$ azt a rácsfüggvényt, amelyre
$$e_h = u_h - v_h. \tag{7.110}$$
Ekkor az e_h rácsfüggvényt a v_h közelítő megoldáshoz tartozó *globális hibának*, és (a 2. pontnak megfelelően) a $\|e_h\|_h$ értéket a közelítő megoldás hibájának nevezzük. A vizsgált módszert akkor nevezzük *konvergensnek*, amikor $\lim_{h\to 0} \|e_h\|_h = 0$. Ha $\|e_h\|_h = \mathcal{O}(h^p)$, akkor a konvergenciát p-ed rendűnek nevezzük.

   7.4. Megjegyzés A numerikus módszer alkalmazásával az a célunk, hogy a numerikus megoldások a (7.92) folytonos feladat megoldásának $u_h = \mathcal{P}_h^{(1)} u$ képéhez kerüljenek közel, és ne egy esetleges más elem $\mathcal{P}_h^{(1)}$ képéhez. Ezért a $\|\cdot\|_h$ normára kikötjük a következő ún. *kompatibilitási* feltételt. Legyen $L : \mathbb{F}_1[a,b] \to \mathbb{F}_2[a,b]$, és jelölje $\|\cdot\|$ az $\mathbb{F}_1[a,b]$ térbeli normát. Azt mondjuk, hogy a $\|\cdot\|_h$ norma *kompatibilis* a $\|\cdot\|$ normával, ha $\lim_{h\to 0} \|\mathcal{P}_h^{(1)} w\|_h = \|w\|$ minden $w \in \mathbb{F}_1[a,b]$ esetén. Ekkor megmutatható a kívánt egyértelműség.[8]

4. Az előző pontban definiált konvergencia biztosítására alkalmazzuk az $L_h = L_h^{(k)}$ megválasztás esetén alkalmazott módszerünket. (Lásd a 7.3. megjegyzést.) Jelölje

[8] Ugyanis, indirekt módon tegyük fel, hogy léteznek olyan $w_1, w_2 \in \mathbb{F}_1[a,b]$, $(w_1 \neq w_2)$ elemek, amelyekre $\lim_{h\to 0} \|w_h - \mathcal{P}_h^{(1)} w_1\|_h = 0$ és $\lim_{h\to 0} \|w_h - \mathcal{P}_h^{(1)} w_2\|_h = 0$. Ekkor $\|\mathcal{P}_h^{(1)} w_1 - \mathcal{P}_h^{(1)} w_2\|_h = \|\mathcal{P}_h^{(1)} w_1 - w_h + w_h - \mathcal{P}_h^{(1)} w_2\|_h \leq \|\mathcal{P}_h^{(1)} w_1 - w_h\|_h + \|w_h - \mathcal{P}_h^{(1)} w_2\|_h \to 0$. Mivel a kompatibilitási feltételt felhasználva ekkor $\|w_1 - w_2\| = \lim_{h\to 0} \|\mathcal{P}_h^{(1)} w_1 - \mathcal{P}_h^{(1)} w_2\|_h = 0$, ezért $w_1 = w_2$, ami állításunkat igazolja. □

tetszőleges $w \in F_1[a,b]$ esetén

$$d_h(w) = L_h(\mathcal{P}_h^{(1)} w) - \mathcal{P}_h^{(2)}(Lw) \in \mathbb{F}(\overline{\omega}_h). \tag{7.111}$$

Azt mondjuk, hogy az L_h *operátor konzisztens az* L *operátorral*, ha valamely $p > 0$ mellett $d_h(w) = \mathcal{O}(h^p)$ minden $w \in F_2[a,b]$ függvényre, és a p számot a konzisztencia rendjének nevezzük.

A differenciaséma egy másik fontos tulajdonsága (a differenciálegyenletek elméletéhez hasonlóan) a *stabilitás*, amely azt fejezi ki, hogy a megoldás folytonosan függ a feladatot meghatározó adatoktól. (Más kifejezéssel: két közeli adatokhoz tartozó megoldás is közel van egymáshoz.) Ezt a fogalmat definiáljuk a következőkben.

Azt mondjuk, hogy a (7.108) feladat stabil kitűzésű (a továbbiakban: *az* L_h *operátor stabil*), ha minden $f_h \in \mathbb{F}(\overline{\omega}_h)$ esetén létezik a feladatnak egyértelmű megoldása, és a megoldásokra teljesül a

$$\|v_h\|_h \leq C\|f_h\|_h \tag{7.112}$$

egyenlőtlenség, ahol $C > 0$ egy, a h-tól független állandó.

A stabilitás (7.112) tulajdonságának ellenőrzése elég nehéz. Megmutatható[9], hogy ez a fogalom ekvivalens az L_h operátorsereg következő, gyakran könnyebben ellenőrizhető tulajdonságával: az invertálható L_h operátorokhoz létezik olyan $C \geq 0$, a h értékétől független állandó, amely mellett

$$\|L_h^{-1}\|_h \leq C. \tag{7.113}$$

7.5. Tétel *Tegyük fel, hogy*

1. L_h *p-ed rendben konzisztens az* L *operátorral,*
2. *az* L_h *operátor stabil.*

Ekkor a numerikus módszer p-ed rendben konvergens, azaz

$$\|e_h\|_h = \mathcal{O}(h^p). \tag{7.114}$$

[9] A (7.112) és a (7.113) tulajdonság ekvivalenciáját a következők módon láthatjuk be. Mivel $v_h = L_h^{-1} f_h$, a norma tulajdonságai miatt $\|v_h\|_h = \|L_h^{-1} f_h\|_h \leq \|L_h^{-1}\|_h \|f_h\|_h$, és így a (7.113) $\Rightarrow$ (7.112) implikáció nyilvánvaló. Másrészt, (7.112) esetén (ismételten a $v_h = L_h^{-1} f_h$ egyenlőség miatt) $\|L_h^{-1} f_h\|_h \leq C\|f_h\|_h$, azaz $\|L_h^{-1} f_h\|_h / \|f_h\|_h \leq C$, ami a norma definíciója miatt a (7.112) $\Rightarrow$ (7.113) implikációt bizonyítja.

Bizonyítás. Felhasználva az e_h globális hiba (7.110) definícióját, illetve a (7.109) definíciót, a következő egyenlőség érvényes:

$$\begin{aligned} L_h e_h =& L_h u_h - L_h v_h = L_h(\mathcal{P}_h^{(1)} u) - L_h v_h = \\ & \underbrace{L_h(\mathcal{P}_h^{(1)} u) - \mathcal{P}_h^{(2)}(Lu)}_{=:d_h(u)} + \underbrace{\mathcal{P}_h^{(2)}(Lu) - L_h v_h}_{=:d_h(f)}. \end{aligned} \tag{7.115}$$

A (7.92) egyenlet mindkét oldalának $\mathcal{P}_h^{(2)}$-képét véve és figyelembe véve a (7.107) összefüggést,

$$\mathcal{P}_h^{(2)}(Lu) = \mathcal{P}_h^{(2)} f = f_h. \tag{7.116}$$

Másrészt, a (7.108) egyenlet alapján $L_h v_h = f_h$. Így a (7.115) relációban $d_h(f) = 0$, tehát

$$L_h e_h = d_h(u). \tag{7.117}$$

Innen a stabilitás (7.112) tulajdonsága miatt érvényes az

$$\|e_h\|_h \leq C\|d_h(u)\|_h = \mathcal{O}(h^p) \tag{7.118}$$

egyenlőtlenség, amely a tétel állítását bizonyítja. □

7.6. Megjegyzés A (7.113) feltételt a *numerikus módszer stabilitásának* is szokásos nevezni. Tehát leegyszerűsítve a 7.5. tétel így fogalmazható meg: "konzisztencia + stabilitás = konvergencia".

A fejezet befejezéseként vizsgáljuk meg, hogy mit jelent a (7.112) stabilitásfogalom, pontosabban, van-e kapcsolata a kezdetiérték-feladatokra definiált 0-stabilitással? Nyilvánvalóan a 0-stabilitás 3.7. definiciója azt fejezi ki, hogy a kezdetiérték-feladatot leíró adatoktól (tehát a kezdeti feltételtől és az diszkretizált egyenlet jobb oldalától) a megoldás egyenletesen folytonosan függ. Más szóval, ha azt szeretnénk garantálni, hogy a két különböző bemenő adatú feladat megoldásai egy előre megadott értéknél kevésbé térjenek el egymástól, akkor előre meg tudjuk mondani, hogy a bemenő adatoknak ehhez milyen közel kell lenniük egymáshoz. Ez azt jelenti, hogy ha operátoregyenletként tekintjük a problémát, akkor az $L_h v_h^1 = f_h^1$ és $L_h v_h^2 = f_h^2$ egyenletekre (feltételezve L_h invertálhatóságát) azt jelenti, hogy ha $\|f_h^1 - f_h^2\|$ kellően kicsi, akkor ez biztosítja $\|v_h^1 - v_h^2\|$ kicsinységét. Ez azt jelenti, hogy léteznie kell egy $C > 0$ állandónak, amely mellett

$$\|v_h^1 - v_h^2\| \leq C\|f_h^1 - f_h^2\|$$

teljesül tetszőleges f_h^1 és f_h^2 esetén. Mivel

$$\|v_h^1 - v_h^2\| = \|L_h^{-1} f_h^1 - L_h^{-1} f_h^2\| = \|L_h^{-1}(f_h^1 - f_h^2)\|,$$

ezért az $f_h = f_h^1 - f_h^2$ jelöléssel ez a

$$\frac{\|L_h^{-1} f_h\|}{\|f_h\|} \leq C$$

feltételt jelenti. Ezért az indukált operátornorma definiciója alapján a 0-stabilitási feltétel a

$$\|L_h^{-1}\| = \sup_{f_h} \frac{\|L_h^{-1} f_h\|}{\|f_h\|} \leq C$$

feltételt, azaz éppen a peremérték-feladatokra kimondott (7.113) stabilitást jelenti.

8. fejezet

Közönséges differenciálegyenletek numerikus megoldása Matlab programmal

Ebben a szakaszban a korábban tárgyalt és elsősorban az elmélet szempontjából megvizsgált numerikus módszerek Matlab programjaival foglalkozunk. Bevezetést adunk a Matlab programozás alapjaiba, majd a kezdetiérték-feladatokra és a peremérték-feladatra való programok elkészítéséről, illetve a programcsomagban már eleve szereplő programok lehetőségeiről adunk ismertetést.

8.1. A Matlab alapjai és a kezdetiérték feladatok megoldása

A Matlab segítségével viszonylag egyszerűen realizálhatók a numerikus eljárásaink.[1] A programokat vagy saját magunk elkészítjük, vagy a Matlab programrendszerben szereplő, "gyárilag elkészített és beégetett" programmal dolgozhatunk. Bár az utóbbi tűnik az egyszerűbbnek, de, mint azt látni fogjuk, az egyszerűbb algoritmusok programjainak önálló elkészítése sem nehéz. Például, az explicit Euler-módszer megírása egy m-fájlban és a továbbiakban önálló függvényként való használata igen egyszerű.

Tekintsük azokat a lépéseket, amelyek ennek megvalósításához szükségesek.

[1] A Matlab egy széles körben elterjedt, több funkcióval is rendelkező matematikai programcsomag. Alapvetően a mérnöki és műszaki számításokat egyszerűsíti le, és elsősorban numerikus és mátrixalgebrai feladatokra dolgozták ki. (A Matlab neve a MATrix és a LABoratory szavakból ered.) A Matlab nyelve egy magas szintű, BASIC-szerű programozási nyelv, éppen ezért könnyű vele dolgozni. Számos kiegészítő csomaggal (ún. toolbox-szal) rendelkezik, amelyek az egyes feladatosztályhoz tartozó eljárásokat tartalmazza.

- Első lépésben indítsuk el a Matlabot, majd az Editorba írjuk a következőket:

```
function[t,y] = expeuler(diffegy, t0, y0, h, N)
t=zeros(N+1,1);
y = zeros(N+1,1);
t(1) = t0;
y(1) = y0;
for i=1:N
t(i+1) = t(i) + h;
y(i+1) = y(i) + h * diffegy(t(i),y(i));
end
```

- A fenti programban az első sorban azt adtuk meg, hogyan tudjuk meghívni a módszerünket. (Mi most expeuler-nek neveztük el.) Ezután felsoroljuk a feladatot illetve a bemenő paramétereket, a bal oldalon pedig a kimenő paramétereket. (Ezek tipikusan a kiszámított és további célra felhasználni kívánt eredmények.) A mi esetünkben öt bemenő és kettő kimenő paraméter van. A két kimenő paraméter a diszkretizált idő vektor (t) és az ezen pontokban kiszámolt numerikus közelítő megoldások vektora (y). Az első bemenő paraméter egy alfüggvény, aminek jelen esetben diffegy a neve, és ebben a függvényben írjuk meg a differenciálegyenletet specifikáló f függvényt. A második paraméter t0, ez a kezdeti időpontot jelöli, a harmadik a t0 pontbeli y0 kezdeti értéket jelenti. A következő paraméter h, amely az időintervallum diszkretizációs lépéstávolságát jelenti, majd végül N jelöli a megtett időlépések számát. A második és harmadik sorban vesszük fel t és y vektorokat, amikben az értékeket először lenullázzuk, és a következő két sorban beállítjuk a kezdőértékeket. Utána következik lényegében a módszer algoritmusa: egy ciklus keretében először beállítjuk a t_i értékeket, majd kiszámoljuk a meredekséget, végül az y_i értékeket, az explicit Euler-módszer képletének megfelelően.

- A fenti programmal még nem tudjuk közvetlenül az adott differenciálegyenlet numerikus megoldását előállítani, ehhez szükségünk van az f függvényt megadó "diffegy" alfüggvény megadására. Ha az

$$u'(t) = -u(t) + t + 1,$$

ahol $t \in [0, 1]$ és $u(0) = 1$ feladat[2] megoldását szeretnénk előállítani, akkor a diffegy nevű alfüggvény elkészítéséhez nyissunk egy új m-fájlt, amibe írjuk a következőket:

```
function dydt = diffegy(t,y)
dydt = -y + t + 1;
```

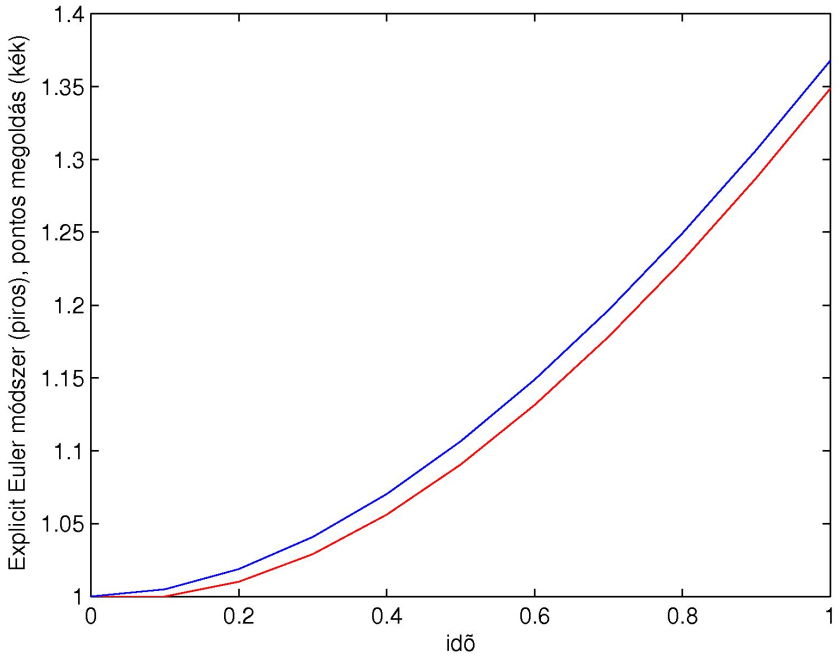
[2]Emlékeztetünk rá, hogy ezen a feladaton teszteltük a numerikus módszereinket a 3.1.1 és a 4.1 fejezetekben. Ebben a részben a Runge-Kutta típusú módszerrel illetve lineáris többlépéses módszerrel is megoldjuk ezt a feladatot.

- Ha már megírtuk mindkét függvényt, ugyanabba a könyvtárba mentve futtassuk le a programunkat $h = 0.1$, $h = 0.01$ és $h = 0.001$ lépéstávolságokkal a $[0, 1]$ intervallumon. A Command ablakba írjuk a következőt: `[T1,Ye] = expeuler(@diffegy, 0, 1, 0.1, 10)`. Enter után megkapjuk a T1 és Ye vektorokat, amelyek a közelítések időbeli helyét és értékét tartalmazzák. Ha ki is szeretnénk rajzoltatni, akkor a `plot(T1,Ye)` paranccsal egy külön ablakban megjelenik a függvényünk. (A $h = 0.01$ és $h = 0.001$ lépésközökre értelemszerűen a `[T1,Ye] = expeuler(@diffegy, 0, 1, 0.01, 100)` illetve a `[T1,Ye] = expeuler(@diffegy, 0, 1, 0.001, 1000)` parancsokat kell beírni.)

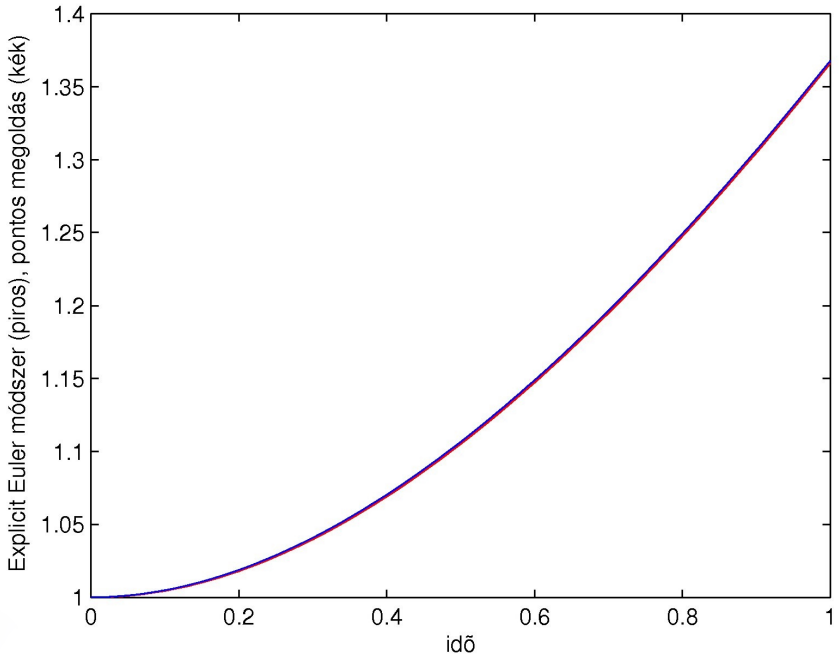
Ha az explicit Euler-módszer hibájára vagyunk kíváncsiak, akkor a pontos megoldással szükséges a numerikus megoldásunkat összehasonlítani. A példánk pontos megoldása az

$$u(t) = e^{-t} + t$$

függvény. A 8.1-8.3 ábrákon láthatjuk a módszer pontosságát a $h = 0.1$, $h = 0.01$ és $h = 0.001$ lépéstávolságokkal. Az elvártaknak megfelelően h csökkenésével a numerikus megoldás grafikonja közeledik a pontos megoldás grafikonjához.



8.1. ábra. A $h = 0.1$ lépésközű explicit Euler-módszer



8.2. ábra. A $h = 0.01$ lépésközű explicit Euler-módszer

A rendszerekre készített programot egy egyszerű populációdinamikai modellen, az ún. Lotka-Volterra-modellen mutatjuk be. Ez az ún. ragadozó-préda modell, amely a két faj számának időbeli alakulását (fejlődését) írja le.

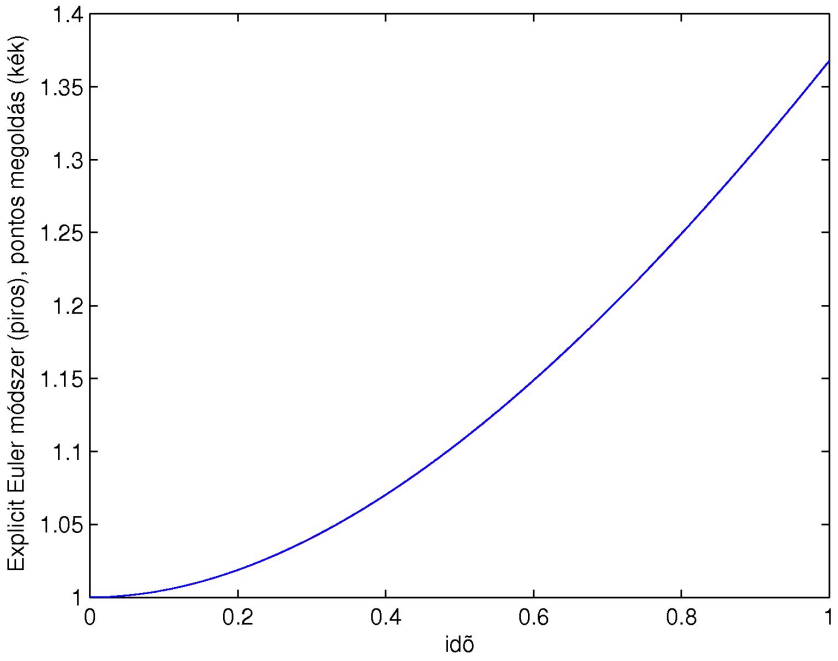
8.1. Példa *Jelölje $x(t)$ egy nyúlpopuláció méretét a t időpontban, míg $y(t)$ egy rókapopuláció méretét. Feltesszük, hogy ismerjük a populációk kezdeti méretét, azaz az $x(0)$ és $y(0)$ értékeket. Ekkor egy egyszerűsített matematikai modellünk felírható a következő differenciálegyenlet-rendszer segítségével:*

$$x'(t) = ax(t) - bx(t)y(t)$$

$$y'(t) = cx(t)y(t) - dy(t),$$

ahol a a nyulak növekedési rátája, d a rókapopuláció halálozási aránya. Mikor egy róka és egy nyúl találkozik, akkor a nyulak számának csökkennie kell, és a rókák számának nőnie, ezt fejezik ki a b illetve c együtthatók.

A modell kapcsán elvárásaink a következők. Ha nincs ragadozó, akkor feltesszük, hogy a préda populáció exponenciális, azaz $x' = ax$ dinamikájú, ahol a a préda populáció növekedési rátája. Feltesszük továbbá, hogy préda hiányában a ragadozók kihalnak



8.3. ábra. A $h = 0.001$ lépésközű explicit Euler-módszer

az $y' = -dy$ egyenlet szerint, ahol d a ragadozó populáció halálozási aránya. Ha a préda ragadozóval találkozik, akkor a prédapopulációnak csökkennie, és a ragadozópopulációnak pedig növekednie kell.

A modell hiányossága, hogy ragadozó hiányában a nyulak száma nem korlátos, azaz kinő a végtelenbe. Ezért módosítsuk a modellünket a következő módon. Az első egyenlet jobb oldalához vegyünk hozzá egy négyzetes elnyelő tagot, azaz valamely γ pozitív állandóval az

$$x'(t) = ax(t) - bx(t)y(t) - \gamma x^2(t)$$

egyenletet tekintsük. Ekkor rókák hiányában a nyulak számának alakulását az $x'(t) = ax(t) - \gamma x^2(t)$ egyenlet írja le, amely már időben korlátos megoldást biztosítja.

A továbbiakban a Matlab segítségével numerikusan oldjuk meg ezen utóbbi, komplexebb modellt. Numerikus módszerként ismét az explicit Euler-módszert választjuk.

Tekintsük a következő példát:

$$x' = x - 2x^2 - xy$$

$$y' = -2y + 6xy,$$

legyen $x(0) = 1$, $y(0) = 0.1$.

Készítsük el az alábbi m-fájlt.

```
function LVexpeuler(x,y,T,N)
h=T/N;
for n=1:N
u=f(x,y); v=g(x,y);
x=x+h*u; y=y+h*v;
xhistory=[xhistory,x]; yhistory=[yhistory,y];
end
t=0:h:T;
plot(t,xhistory,'red', t, yhistory,'blue')
xlabel('idö'), ylabel('préda (piros), ragadozó (kék)')
function U=f(x,y)
U=x-2*x.*x-x.*y;
function V=g(x,y)
V=-2*y+6*x.*y;
```

Itt bemenő paraméterként a következőket adjuk meg: x: a prédák kezdeti viszonyszáma, y: a ragadozók kezdeti viszonyszáma, T: a vizsgált időintervallum és N: az osztásrészek száma.

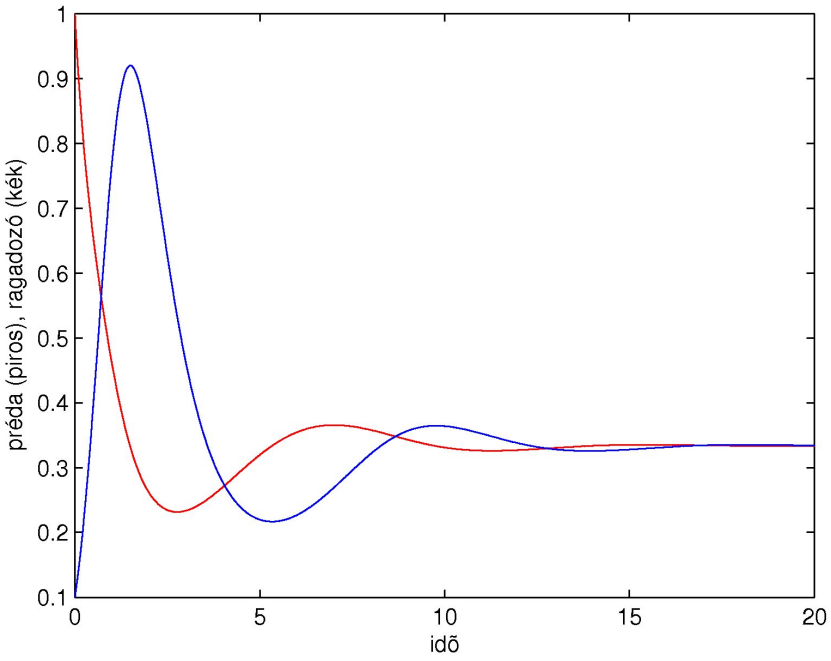
Mentsük el, és az `LVexpeuler(1,0.1,20,1000)` utasítással számoljuk ki az eredményt. Futtatás után a 8.4 grafikont kapjuk.

Ahogy az elvárható, a populáció vagy kihal, vagy egyensúlyba kerül. A grafikonon is jól látszik, hogy a kezdeti időpontokban sok ragadozó van és nincsenek préda állatok, ezután viszont a prédaállatok száma nő, és az idő múlásával egyensúlyba kerülnek a populáció méretei.

Megjegyezzük, hogy a többi egylépéses módszerek Matlab programja hasonlóan elkészíthető, de az implicit módszerek esetén lépésenként egy (általában nemlineáris) egyenlet megoldási algoritmusát is be kell építenünk.

Mint azt már említettük, léteznek a Matlab programrendszerben beépített programok, amelyek alkalmasak akár nagy pontossággal és hatékonyan megoldani a kezdetiérték-feladatokat. Ilyen például a gyakran alkalmazott ODE45 rutin, ami a beágyazott Dormand-Prince-módszeren alapul. Az ODE45 rutint ugyanolyan módon hívjuk meg, mint a korábban leírt, saját magunk által írt programokat. Nevezetesen, `[T1, Y45] = ode45(@diffegy, T1, 1)`. Itt is két kimenő paraméter van, az idő és a közelítő értékek vektora. A bemenő paraméterek rendre: az alfüggvény, ami leírja a differenciálegyenletünket, az idő vektor, hogy melyik időpontokban számoljuk ki a megoldást, a kezdeti-érték vektor.[3]

[3]Megjegyezzük, hogy megadható egy negyedik (opcionális) paraméter is, amelynek segítségével beállíthatjuk az integrálási paramétereket. Ugyanakkor a gyakorlatban elegendő az alapbeállítás, és ezért az alapértelmezést csak indokolt esetekben célszerű megváltoztatni. Ezt az `odeset` rutinnal hajthatjuk végre, amelynek részletei a Matlab help leírásában megtalálhatók.



8.4. ábra. A Lotka-Volterra-modell megoldása explicit Euler-módszerrel

Az ODE45 módszer pontosságának jellemzésére szolgáló 8.1 és 8.2 táblázatok a szokásos tesztfeladatunkon a $h = 0.1$ és $h = 0.01$ lépéstávolság melletti eredményeket tartalmazzák. Láthatóan itt a módszer pontossága a lépésköz csökkentésével érdemben nem javul. Ennek több oka is van. A lényeg, hogy a beépített módszerek adaptív módon választják meg a lépésközt (tehát nem a mi általunk megadott h lépésközű, ekvidisztáns rácshálón számol!), és ehhez az általunk megadott pontosságot használja kritériumként. A teszt során mi meglehetősen nagy pontosságot követeltünk meg, amit azt a táblázatban szereplő hiba nagysága is jelzi. A h paramétert azért adjuk meg, hogy megadjuk azokat a pontokat, ahol a numerikus megoldást kiértékeljük, azaz a h lépésközű, ekvidisztáns rácsháló csomópontjaiban. Így h csökkentése lényegében csupán a kiértékelendő pontok számát növelte meg.[4]

[4]Megjegyezzük, hogy ha a kiértékelendő csomópontban nem számol közvetlenül a módszer numerikus megoldást, akkor erre a pontra a többi pontbeli numerikus értékből interpolációval határozza meg a közelítést. Ezért h csökkentésével várható az interpoláció javulása, azaz bizonyos értelemben javulhat a numerikus eredmény, ha h túlságosan nagy.

t_i	pontos megoldás	numerikus megoldás	hiba
0	1.0000	1.0000	0
0.1000	1.0048	1.0048	$2.9737e-010$
0.2000	1.0187	1.0187	$5.3815e-010$
0.3000	1.0408	1.0408	$7.3041e-010$
0.4000	1.0703	1.0703	$8.8120e-010$
0.5000	1.1065	1.1065	$9.9668e-010$
0.6000	1.1488	1.1488	$1.0822e-009$
0.7000	1.1966	1.1966	$1.1424e-009$
0.8000	1.2493	1.2493	$1.1814e-009$
0.9000	1.3066	1.3066	$1.2026e-009$
1.0000	1.3679	1.3679	$1.2090e-009$

8.1. táblázat. Az ODE45 eredményei a $h = 0.1$ lépésközű rácshálón

t_i	pontos megoldás	numerikus megoldás	hiba
0	1.0000	1.0000	0
0.0100	1.0000	1.0000	$2.2204e-016$
0.0200	1.0002	1.0002	$8.8940e-011$
⋮	⋮	⋮	⋮
0.1000	1.0048	1.0048	$5.4080e-009$
⋮	⋮	⋮	⋮
0.5000	1.1065	1.1065	$2.8278e-009$
⋮	⋮	⋮	⋮
0.8000	1.2493	1.2493	$1.6519e-009$
⋮	⋮	⋮	⋮
0.9000	1.3066	1.3066	$1.3610e-009$
⋮	⋮	⋮	⋮
0.9900	1.3616	1.3616	$1.0920e-009$

8.2. táblázat. Az ODE45 eredményei a $h = 0.01$ lépésközű rácshálón

t_i	a pontos megoldás	a numerikus megoldás	a hiba
0	1.0000	1.0000	0
0.1000	1.0048	1.0048	$6.2300e-006$
0.2000	1.0187	1.0187	$1.8714e-005$
0.3000	1.0408	1.0408	$2.7885e-005$
0.4000	1.0703	1.0703	$2.1933e-005$
0.5000	1.1065	1.1065	$1.8889e-005$
0.6000	1.1488	1.1488	$1.7254e-005$
0.7000	1.1966	1.1966	$1.5668e-005$
0.8000	1.2493	1.2493	$1.4228e-005$
0.9000	1.3066	1.3066	$1.2872e-005$
1.0000	1.3679	1.3679	$1.1643e-005$

8.3. táblázat. Az ODE113 eredményei a $h = 0.1$ lépésközű rácshálón

A többlépéses módszerek algoritmusai is megtalálhatók a Matlabban. Ilyen az ODE113, amelynek a rendje 1-től 13-ig változhat és az Adams-Bashforth-Moulton-módszeren alapszik. Összehasonlítva az ODE45 módszerrel, kisebb számítási igényű, és különösen előnyös, amikor az f függvény kiértékelése költséges. Megjegyezzük, hogy ez a módszer is `[T1, Y113] = ode113(@diffegy, T1, 1)` szintaktikájú, és nem merev feladatokra alkalmazzuk. A módszer eredményeit a szokásos tesztfeladatunkon $h = 0.1$ és $h = 0.01$ lépéstávolság melletti 8.3 és 8.4 táblázatok tartalmazzák.

Az alábbi táblázatban foglaljuk össze néhány olyan módszer pontosságát, amelyet ebben a szakaszban tárgyaltunk. (Módszereinket a szokásos a (3.13) tesztfeladatra a $t^* = 1$ pontban hasonlítjuk össze.)

módszer	e_{n^*}	
	$h_1 = 0.1$	$h_2 = 0.01$
explicit Euler	1.9201e-002	1.8471e-003
javított Euler	6.6154e-004	6.1775e-006
implicit Euler (pontosabb)	1.7664e-002	1.8318e-003
implicit Euler	2.1537e-002	1.8687e-003
ODE45	1.2090e-009	1.0903e-009
ODE23	1.6607e-005	1.5087e-005
ODE113	1.1643e-005	1.6360e-008

t_i	a pontos megoldás	a numerikus megoldás	a hiba
0	1.0000	1.0000	0
0.0100	1.0000	1.0001	$1.6625e-007$
0.0200	1.0002	1.0002	$1.4800e-007$
⋮	⋮	⋮	⋮
0.1000	1.0048	1.0048	$1.4004e-007$
⋮	⋮	⋮	⋮
0.5000	1.1065	1.1065	$1.7178e-008$
⋮	⋮	⋮	⋮
0.8000	1.2493	1.2493	$2.0090e-008$
⋮	⋮	⋮	⋮
0.9000	1.3066	1.3066	$1.8052e-008$
⋮	⋮	⋮	⋮
0.9800	1.3553	1.3553	$1.6692e-008$
0.9900	1.3616	1.3616	$1.6525e-008$
1.0000	1.3679	1.3679	$1.6360e-008$

8.4. táblázat. Az ODE113 eredményei a $h = 0.01$ lépésközű rácshálón

Befejezésül megjegyezzük, hogy merev feladatok megoldására is vannak rutinok a Matlabban. Ilyenek a ODE15S (amelynek meghívása `[T1, Y15s] = ode15s(@diffegy, T1, 1)`), ODE23S (amelynek meghívása `[T1, Y23s] = ode23s(@diffegy, T1, 1)`). A ODE123T és az ODE123TB rutinok az enyhén merev rendszerek megoldására javasolhatók. Ezek a módszerek alacsony pontosságúak. Az ODE15S módszer a backward differentiation formulákon (BDF) (amelyet Gear-módszernek is szokásos nevezni) alapul és változó lépéshosszú módszer. Különösen akkor javasolt, amikor az ODE45 módszer nagyon lassú vagy egyáltalán nem működik. Az ODE23S egy másodrendű, módosított, egylépéses Rosenbrock-módszer. Mivel egylépéses, ezért általában hatékonyabb, mint az ODE15S módszer.
A módszerek részletei iránt érdeklődőknek javasoljuk a [1, 19] irodalmakat. A Matlab programok kapcsán javasoljuk a [6, 18] irodalmakat.

8.2. A peremérték-feladatok megoldása Matlab segítségével

A peremérték-feladatok numerikus megoldására a Matlab a BVP4C rutint javasolja. Ez a program a kétpontos peremértékfeladatokat elég általános esetben képes megoldani: egyrészt nem szeparált peremfeltételek is megadhatók (amikor nem csak a két végpontbeli függvényértékek illetve deriváltak adottak külön-külön a végpontokban, hanem azok kombinációi), másrészt a peremfeltételben paraméter is megadható. Módszerként a kollokációs módszert alkalmazza, amelynek eredményeként egy nemlineáris algebrai egyenletrendszert nyerünk. Ennek megoldására a Newton módszert alkalmazzuk. Mivel ehhez a bemenő függvények parciális deriváltjai is szükségesek, ezeket is közelítőleg, nevezetesen véges differenciákkal határozzuk meg. A módszer részletei megtalálhatók a
http://200.13.98.241/~martin/irq/tareas1/bvp_paper.pdf
linken a "Solving Boundary Value Problems for Ordinary Differential Equations in Matlab with bvp4c (Lawrence F. Shampine, Jacek Kierzenka, Mark W. Reichelt) leírásban. Nagyon hasznos továbbá a
http://www.mathworks.com/matlabcentral/fileexchange/3819
link, ahol a BVP4C rutin használatára található egy oktatóanyag és számos kidolgozott példa. Javasoljuk még a
http://www.mathworks.com/help/techdoc/ref/bvp4c.html
linket, ahol a BVP4C leírása mellett további területeken való alkalmazhatósága is szerepel.

Az anyagunkban szereplő két alapvető módszerünkre, a belövéses módszerre illetve a véges differenciák módszerére a megfelelő Matlab programok önállóan is elkészíthetők. Ennek leírásával és egy modellfeladaton való alkalmazásával a továbbiakban foglalkozunk.

8.2.1. A modellfeladat: stacionárius hőeloszlás homogén vezetékben

Tegyük fel, hogy egy hosszú és vékony vezeték két, állandó hőmérsékletű fal között helyezkedik el. A vezeték vastagsága a hosszúságához képest elhanyagolható, így a sugárirányú hőmérsékletváltozás (radiális sugárzás) elhanyagolható, és ezért a hőmérséklet csak az x egydimenziós térbeli koordinátától függ. A hőáramlás egyrészt a vezetékben történő hosszirányban kondukció, másrészt a vezeték és a vezetéket körbevevő állandó hőmérsékletű gáz közötti konvekció hatására történik. Feladatunk a stacionárius hőeloszlás meghatározása.
Írjuk fel először a mérlegegyenletet egy Δx hosszúságú elemre! Jelölje $q(x)$ az x ponthoz tartozó hőfluxust [mértékegysége: $J/(m^2s)$], A_c a keresztmetszet területét [m^2]), azaz $A_c = \pi r^2$, ahol r a vezeték sugara. Legyen D_{kon} a konvekciós hővezetési együttható $[J/(Km^2s)]$, ahol K a Kelvin hőfok, A_s az elem felülete [m^2], azaz $A_s = 2\pi r\Delta x$, és T_∞ a körbevevő gáz hőmérséklete [K]. Az x pontbeli ismeretlen hőmérsékletet $T(x)$ jelöli. Ekkor a mérlegegyenlet alapján

$$q(x)A_c = q(x+\Delta x)A_c - D_{kon}A_s(T_\infty - T(x)), \tag{8.1}$$

ahol a bal oldalon a bemenő hőmennyiség, a jobb oldalon pedig a kimenő hőmennyiség és a radiális áramlás okozta veszteség szerepel. Leosztva a (8.1) egyenlőséget az elemi rész térfogatával ($\pi r^2\Delta x$), átrendezés után a

$$-\frac{q(x+\Delta x)-q(x)}{\Delta x} + \frac{2D_{kon}}{r}(T_\infty - T(x)) = 0 \tag{8.2}$$

egyenlőséget kapjuk. Áttérve a $\Delta x \to 0$ határértékre, (8.2) a

$$-q'(x) + \frac{2D_{kon}}{r}(T_\infty - T(x)) = 0 \tag{8.3}$$

egyenletet eredményezi. Mivel Fourier törvénye alapján a fluxusra $q(x) = -D_{hov}T'(x)$, ahol D_{hov} $[J/(Kms)]$ a hővezetési együttható, ezért a (8.3) egyenlet felírható

$$T''(x) + D(T_\infty - T(x)) = 0 \tag{8.4}$$

alakban, ahol

$$D = \frac{2D_{kon}}{rD_{hov}}$$

a hőleadást jellemző állandó [m^{-2}]. Ha a vezeték hosszúsága L, akkor

$$T(0) = T_0, \quad T(L) = T_1, \tag{8.5}$$

ahol T_0 és T_1 a bal illetve jobb oldali fal hőmérséklete. Ekkor tehát matematikai modellünk a (8.4)–(8.5) kétpontos peremérték-feladat.

A (8.4)–(8.5) feladat pontos megoldása analitikusan előállítható. Ugyanis az egyenlet átírható

$$T''(x) - DT(x)) = -DT_\infty \tag{8.6}$$

alakra, amelynek megoldásához a homogén egyenlet általános megoldása és egy partikuláris megoldás ismerete szükséges. Mivel a $\lambda = \pm\sqrt{D}$ jelölésekkel a homogén feladat általános megoldása $T_{hom}(x) = C_1 e^{\lambda x} + C_2 e^{-\lambda x}$, ahol C_1 és C_2 tetszőleges állandók, a $T_{part}(x) = T_\infty$ pedig egy partikuláris megoldás, ezért a (8.6) egyenlet általános megoldása

$$T(x) = C_1 e^{\lambda x} + C_2 e^{-\lambda x} + T_\infty. \tag{8.7}$$

A képletben szereplő állandókat a (8.5) peremfeltételekből határozhatjuk meg:

$$C_1 = \frac{(T_0 - T_\infty)e^{-\lambda L} - (T_1 - T_\infty)}{e^{-\lambda L} - e^{\lambda L}}, \quad C_2 = \frac{-(T_0 - T_\infty)e^{\lambda L} + (T_1 - T_\infty)}{e^{-\lambda L} - e^{\lambda L}}. \tag{8.8}$$

A továbbiakban az alábbi adatokkal számolunk: $L = 10$, $D_{kon} = 1$, $D_{hov} = 200$, $r = 0.2$, $T_\infty = 200$, $T_0 = 300$, $T_1 = 400$. (Adataink a korábban megadott mértékegységekben értendők.) Ekkor $D = 0.05$, és a feladatunk pontos megoldása a

$$T(x) = 20.4671 e^{\sqrt{0.5}x} + 79.5329 e^{-\sqrt{0.5}x} + 200 \tag{8.9}$$

függvény.[5]

8.2.2. A tesztfeladat numerikus megoldása Matlab segítségével

Alkalmazzuk a belövéses módszert a (8.4)–(8.5) feladatra. Első lépésben átírjuk a másodrendű közönséges differenciálegyenletet egy kétismeretlenes, elsőrendű rendszerekre:

$$\begin{aligned} T'(x) &= z(x), \\ z'(x) &= -D(T_\infty - T(x)). \end{aligned} \tag{8.10}$$

Kezdeti feltételként a

$$T(0) = T_0, \quad z(0) = z_0 \tag{8.11}$$

feltételeket adjuk meg, ahol T_0 ismert, z_0 értékét a belövéses módszernek megfelelően pedig úgy választjuk meg, hogy a (8.10)–(8.11) kezdetiérték-feladat $T(x)$ megoldására a $T(L) = T_1$ feltétel teljesüljön.

A belövéses módszer algoritmusát követve Matlab programunkat két FUNCTION segítségével adjuk meg. A BVPS rutin adott z_0 érték mellett az explicit Euler-módszerrel kiszámolja a numerikus megoldást. Bemenő paraméterként a $[0, L]$ intervallum felosztásához szükséges osztásrészek száma (Nx) illetve az ismeretlen z komponens-függvény

[5] Amikor a megoldást beprogramozzuk, nem a (8.9) képletben szereplő állandókat adjuk meg közvetlenül, hanem a Matlab segítségével a (8.7) képletből számoljuk ki.

kezdeti értéke ($z0$) szerepel. Kimenő paraméterként a végponti ($x = L$) hőmérséklet (TL), a csomópontokban számolt hőmérsékletek ($Tvect$) és a csomópontok koordinátái ($xvect$) szerepelnek. A rutin a következő:

```
function [TL,Tvect,xvect] = bvpsee(Nx,z0)
%
% Belövéses módszer (shooting method) egy L hosszúságú rúd
% stacionárius höeloszlásának kiszámítására.
% Ez a következö kétpontos peremérték-feladat megoldását
%igényli:
%
% d^2T
% --- +D(Tinf - T) = 0,   T(0) = T0,   T(L) = T1
% dx^2
% A kezdetiérték-feladat megoldására az egyszerü explicit
% Euler módszert alkalmazzuk.
% Nx:  a térbeli osztásrészek száma
% z0:  a kezdeti meredekség
T0 = 300; % bal oldali végpontbeli peremfeltétel
T1 = 400; % jobb oldali végpontbeli peremfeltétel
Tinf = 200; % külsö hömérséklet
D = 0.05;  % állandó
L = 10; % a rúd hossza
xs = 0; % kezdöpont koordinátája
T = T0; % kezdeti feltétel
deltax=L/Nx; Tvect(1) = T; xvect(1) = xs; z=z0;
for i=2:Nx+1
dTdx = z;
dzdx = -c * (Tinf - T);
T = T + dTdx * deltax; % Euler módszer
z = z + dzdx * deltax;  xvect(i) = xs + (i-1)*deltax;
Tvect(i) = T; end;
TL=T;
```

Ezzel a rutinnal tehát egy adott kezdeti z értékkel tudjuk meghatározni a hőmérséklet-eloszlást. Nyilvánvalóan ez önmagában nem elegendő, hiszen a kiszámolt TL különbözik $T1$-től. A belövéses módszer algoritmusát követve a BVPS rutinnal különböző kezdeti $z(0)$ értékekre kiszámoljuk a végpontbeli TL hőmérsékletet, és az így kapott értékekkel kiszámoljuk a $z(0)$ értékhez tartozó $\varepsilon(z(0)) = TL(z(0)) - T1$ eltéréseket. Mivel olyan $z^\star(0)$ értéket keresünk, amelyre $\varepsilon(z^\star(0)) = 0$, ezért a $(z(0), \varepsilon(z(0)))$ pontokra interpolációs polinomot fektetünk, és a keresett $z^\star(0)$ ennek a zérushelye lesz. Ezt a két lépést a programban a (7.10) szerinti inverz interpolációval végezzük el.

A fenti lépéseket a SHOOTING rutin hajtja végre, amelynek bemenő paraméterei a következők:

- Nx: a $[0, L]$ intervallum felosztásához szükséges osztásrészek száma,
- $zstart$: a $z(0)$ kezdeti első értéke (ami az ismeretlen T függvény gradiensét jelenti az $x = 0$ pontban, azaz a hőmérsékletváltozás gradiense a kezdet időpontban),
- $deltaz$: a különböző $z(0)$ értékek meghatározására szolgáló megváltozás,
- Nz: a kezdeti $zstart$ értékektől jobbra és balra további Nz számú $z(0)$ értékkel számolunk, nevezetesen a $z(0) = zstart \pm k \cdot deltaz$ $(k = 1, 2, \ldots, Nz)$ kezdeti értékekkel is meghatározzuk TL értékét. (Ezzel állítjuk elő az interpolációs alappontokat.)

A rutin opcionálisan kiírja a numerikus megoldást, és a pontos megoldás ismeretében kiírja a maximumnormabeli hibát és kirajzolja a pontos és közelítő megoldásokat.

```
function shooting(Nx,zstart,deltaz,Nz)
%
% Belövéses módszer (shooting method) egy L hosszúságú rúd
% stacionárius höeloszlásának kiszámítására.
% Ez a következö kétpontos peremérték-feladat
% megoldását igényli:
%
% d^2T
% --- +D(Tinf - T) = 0,   T(0) = T0,   T(L) = T1
% dx^2
% A kezdetiérték-feladat megoldására a BVPS
%nevü rutint alkalmazzuk.
% Nx:  a térbeli osztásrészek száma
% z0:  a kezdeti meredekség
T0 = 300; % bal oldali végpontbeli peremfeltétel
T1 = 400; % jobb oldali végpontbeli peremfeltétel
Tinf = 200; % külsö hömérséklet
D = 0.05; % állandó
L = 10; % a rúd hossza
xs = 0; % kezdöpont koordinátája
T = T0; % kezdeti feltétel
deltax=L/Nx; zv(Nz+1)=zstart; z=zv(Nz+1);
[T,Tvect,xvect]=bvpsee(Nx,z);Tvegpont(Nz+1)=T;
for i=1:Nz
zv(i)=zstart-(Nz+1-i)*deltaz; z=zv(i);
```

```
[T,Tvect,xvect]=bvpsee(Nx,z);Tvegpont(i)=T;
zv(Nz+1+i)=zstart+i*deltaz; z=zv(Nz+1+i);
[T,Tvect,xvect]=bvpsee(Nx,z);Tvegpont(Nz+1+i)=T;
end
for i=1:2*Nz+1
Tvegpont(i); zv(i);
end
% A gyök megkeresésére az inverz interpolációt alkalmazzuk.
Tint=Tvegpont-Tb; z=interp1(Tint,zv,0);
fprintf('A meredekség:  %fn',z)
[Tfinal,Tvect,xvect]=bvpsee(Nx,z);
% A Meg nevü mátrixba összerakjuk a numerikus megoldást
% és utána szükség esetén kiiratjuk
Meg=ones(Nx+1,3);
for i=1:Nx+1
Meg(i,1)=i; Meg(i,2)=xvect(i); Meg(i,3)=Tvect(i);
end
% disp(' csomópont koordin.  hömérséklet')
% disp(Meg)
% plot(xvect,Tvekfinal);
% Ha ismerjük a pontos megoldást, akkor adjuk meg
% egy PONTOSSHOOTING nevü függvényben.
% Ekkor a hiba kiszámítható és opcionálisan kirajzolhatjuk
% a pontos és közelítö megoldásokat.
% [Tpontos,zpontos]=pontosshooting(Nx,xvect);
% hiba=norm(Tvect-Tpontos,inf);
% disp('lépésköz és hiba max.  normában:'),deltax, hiba
% i=1:Nx+1;
% plot(i,Tvect,'r', i, Tpontos,'b') xlabel('rúd'),
% ylabel('belövéses módszer (piros),
% pontos megoldás (kék)')
```

Mivel a (8.9) képlet segítségével ismerjük a pontos megoldást, ezért a belövéses módszert különböző paraméterek mellett tesztelni tudjuk. A Δx lépésköz függvényében a hibavektor maximumnormáját a 8.5 táblázat mutatja. A pontos és numerikus megoldások az egyes rácshálókon a 8.5-8.8 ábrákon láthatók.

8.1. Megjegyzés Futtatásainkat $Nz = 4$ és $zstart = -12$ értékekkel hajtottuk végre. Ha nagyobb Nz értéket választunk, akkor sem változik a futás eredménye. Ennek oka, hogy a tesztfeladatunk lineáris, azaz a 7.1.2 szakasz értelmében direkt módon is kiszámolható két adat ismeretében a kezdeti meredekség. Ezért lényegében nem szükséges

Δx	0.1	0.01	0.001	0.0001
e_h	$4.6133e-001$	$4.6786e-002$	$4.6852e-003$	$4.6859e-004$
rend	$1.1453e-001$	$1.0142e-001$	$1.0014e-001$	$1.0001e-001$

8.5. táblázat. A belövéses módszer explicit Euler-módszeres hibája a maximumnormában a tesztfeladatra, és a konvergencia rendje.

Δx	0.1	0.01	0.001	0.0001
e_h	$6.6929e-003$	$6.7386e-005$	$6.7433e-007$	$6.7446e-009$
rend	$1.0611e-002$	$1.0068e-002$	$1.0007e-002$	$1.0002e-002$

8.6. táblázat. A belövéses módszer javított explicit Euler-módszeres változatának hibája a maximumnormában a tesztfeladatra és a konvergencia rendje.

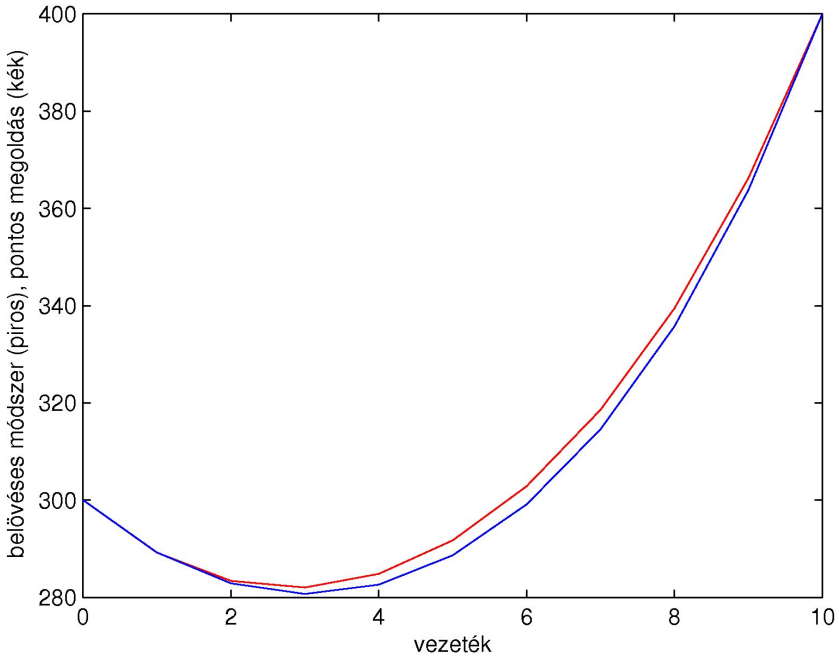
több pont megadása, sőt, valójában az $Nz = 2$ is elegendő. Ehhez elegendő összehasonlítani a 8.9 és a 8.5 ábrákat, amelyeket ugyanolyan térbeli felosztásra, de különböző Nz értékekre ($Nz = 2$ és $Nz = 10$) kaptunk.

8.2. Megjegyzés Ha nagyon távoli *zstart* értéket adunk meg, akkor kevés interpolációs alappont esetén előfordulhat, hogy az inverz interpoláció nem működik. (A nulla érték kívül esik az alappontokon.) Ilyenkor célszerű előzetesen néhány önálló futtatást végezni a BVPS rutinnal, és közelítőleg meghatározni a keresett z értéket.

8.3. Megjegyzés Ha az explicit Euler-módszer helyett a másodrendű javított Euler-módszerrel számolunk, akkor eredményeink hibái a 8.6 táblázatban láthatók. Mint az várható volt, a módszerünk másodrendben konvergens.

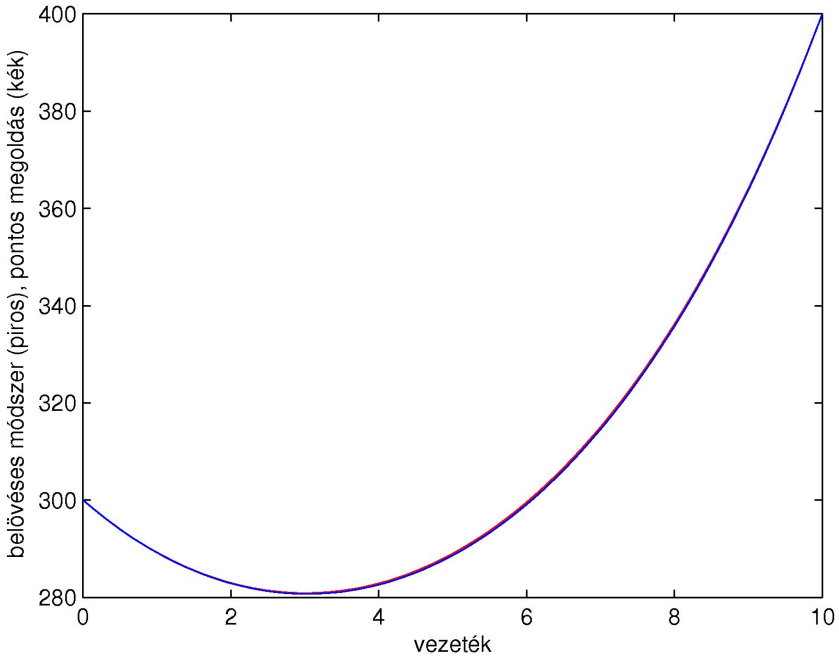
Térjünk át a (8.4)–(8.5) feladat véges differenciás megoldására. Nem részletezve az algoritmust, megadjuk a VDM1 rutint, amely a véges differenciák módszerével megoldja a fenti feladatot. Itt bemenő paraméterként az L hosszúságú vezeték osztásrészeinek számát (Nx) kell megadni. Kimenő adatként a csomópontok koordinátáit ($xvect$) illetve a csomópontokhoz tartozó közelítő megoldást tartalmazó vektort ($Tmego$) kapjuk.

```
function[xvect,Tmego]=vdm1(Nx)
% Véges differenciák módszere egy L hosszúságú rúd stacionárius
% höeloszlásának kiszámítására.  Ez a
% d^2 T
% --- + c (Tinf - T) = 0
% dx^2
```



8.5. ábra. A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben 10 osztáspontú ekvidisztáns rácshálón.

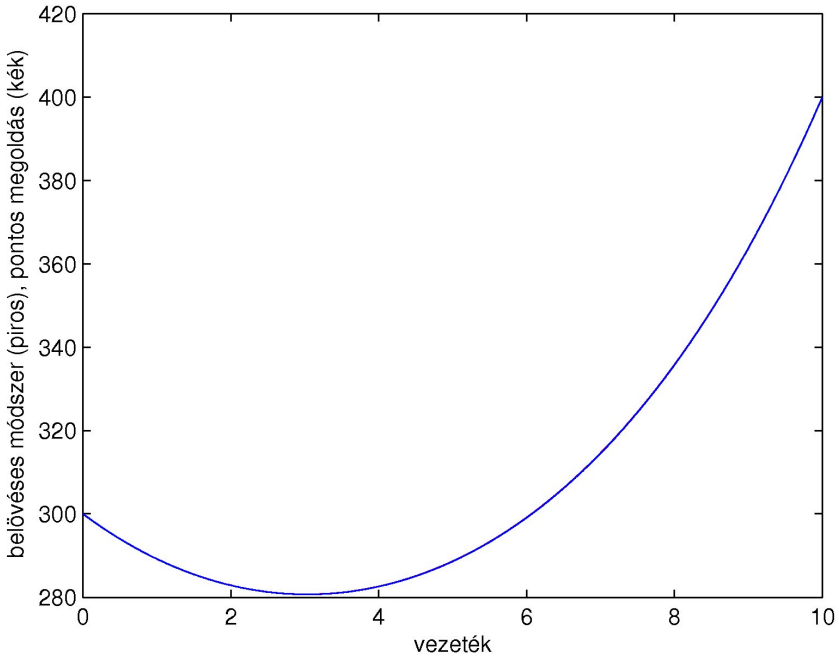
```
% típusú kétpontos peremérték feladatok megoldását igényli,
% amelyet a szokásos differenciasémával approximálunk.
% A peremfeltételeket közvetlenül beírhatjuk a sémába.
% Nx:  a térbeli osztásrészek száma
Ta = 300; % bal oldali végpontbeli peremfeltétel
Tb = 400; % jobb oldali végpontbeli peremfeltétel
Tinf = 200; % külsö hömérséklet
c = 0.05; % állandó
L = 10; % a vezeték hossza
ndivs = Nx; nunknowns = ndivs - 1; deltax = L/ndivs;
A = -(2 + deltax^2*c); B = -deltax^2*c*Tinf;
for i=1:Nx+1, xvect(i)=(i-1)*deltax; end
% a diszkretizációs alappontok elöállítása
matrix = zeros(nunknowns);
% a lineáris egyenletrendszer összeállítása kezdödik.
matrix(1,1) = A;
% az elsö egyenlet összeállítása
```



8.6. ábra. A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben 100 osztáspontú ekvidisztáns rácshálón.

```
matrix(1,2) = 1; rhs(1)= B - Ta;
for i = 2:nunknowns - 1
% a belsö pontokhoz tartozó egyenletek
matrix(i,i-1) = 1;
matrix(i,i) = A;
matrix(i,i+1) = 1;
rhs(i)= B;
end;
matrix(nunknowns, nunknowns-1) = 1;
% az utolsó egyenlet összeállítása
matrix(nunknowns, nunknowns) = A; rhs(nunknowns)= B - Tb;
T = matrix \ rhs'; % a lineáris egyenlet megoldása
Tmego(1)= Ta; % a teljes megoldásvektor elöállítása
Tmego(2:1 + nunknowns) = T(:); Tmego(nunknowns + 2) = Tb;
```

Eredményeinket a tesztfeladatra a 8.7 táblázatban adjuk meg. Az első három esethez (azaz amikor $Nx = 3, 6, 9$) tartozó véges differenciás megoldást a pontos megoldással



8.7. ábra. A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben 1000 osztáspontú ekvidisztáns rácshálón.

együtt a 8.10-8.12 ábrákon mutatjuk be.

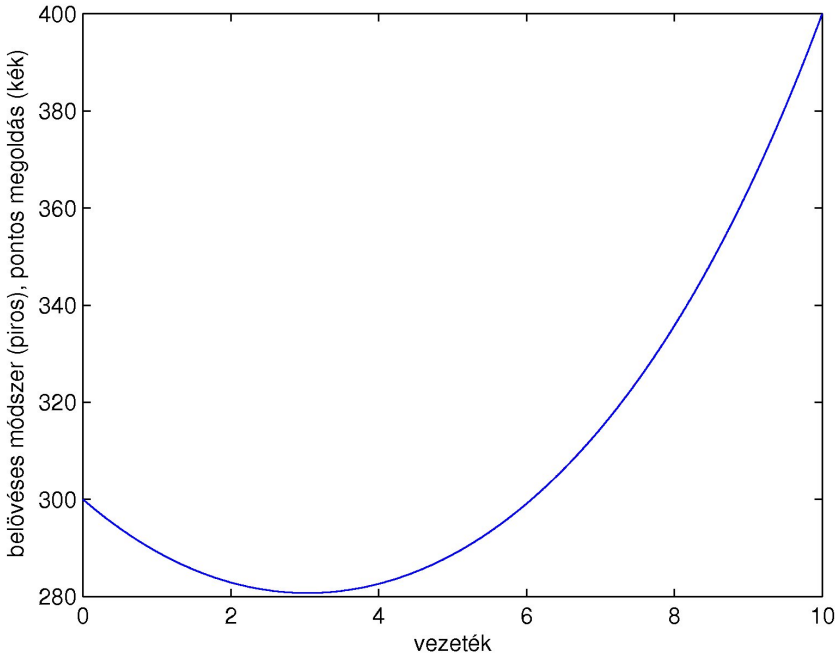
Befejezésül hasonlítsuk össze módszereinket ugyanazon diszkrét rácshálón! Legyen $Nx = 8$, azaz 9 osztáspontú, $\Delta x = 1.25$ lépésközű rácshálón hasonlítjuk össze az explicit Euler- és javított explicit Euler-módszeres belövéses eredményeinket, valamint a véges differenciás módszert a pontos megoldással. Eredményeinket a 8.8 táblázat tartalmazza. A megoldásokat a 8.13 ábrán láthatjuk. Mivel a megoldások közel haladnak egymáshoz, az ábra egy része kinagyítva látható a 8.14 ábrán.

osztásrészek száma	hiba max. normában	rend
3	$1.7002e+000$	
6	$4.5604e-001$	$2.6823e-001$
12	$1.1647e-001$	$2.5540e-001$
24	$2.9205e-002$	$2.5074e-001$
48	$7.3158e-003$	$2.5050e-001$
96	$1.8293e-003$	$2.5004e-001$
192	$4.5734e-004$	$2.5001e-001$
384	$1.1434e-004$	$2.5000e-001$
768	$2.8582e-005$	$2.4999e-001$
1536	$7.1573e-006$	$2.5041e-001$

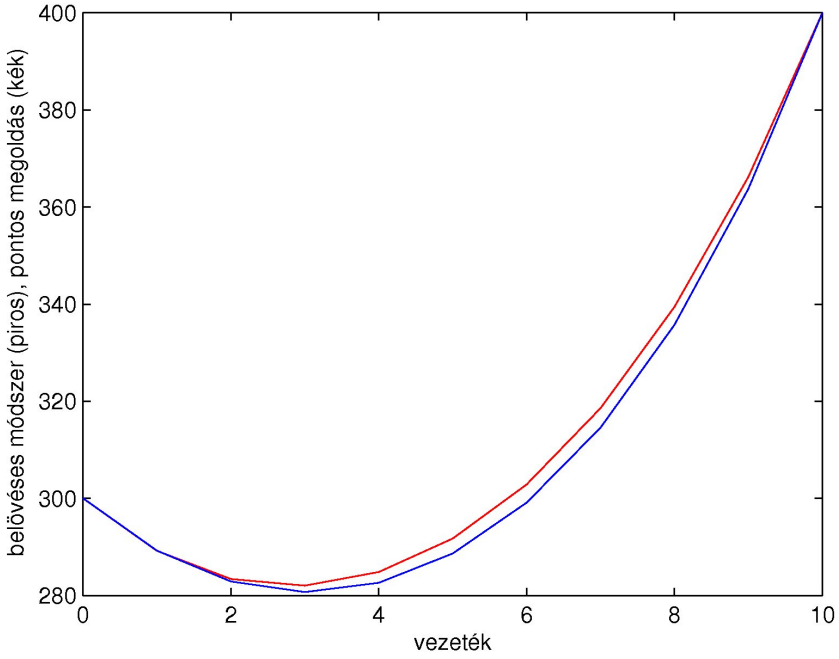
8.7. táblázat. A véges differenciák módszerének hibája feleződő lépéshosszal. A harmadik oszlop a rendet mutatja. (Az elméleti rend 0.25.)

rácspont	EEM	JEE	VDM	pontos megoldás
0	300.0000	300.0000	300.0000	300.0000
1.2500	287.2008	287.6551	287.3173	287.3173
2.5000	282.2141	282.0058	281.4562	281.4562
3.7500	284.0400	282.6293	281.9589	281.9589
5.0000	292.2889	289.5832	288.8646	288.8646
6.2500	307.1033	303.4096	302.7129	302.7129
7.5000	329.1279	325.1783	324.5856	324.5856
8.7500	359.5199	356.5687	356.1916	356.1916
10.0000	400.0000	400.0000	400.0000	400.0000

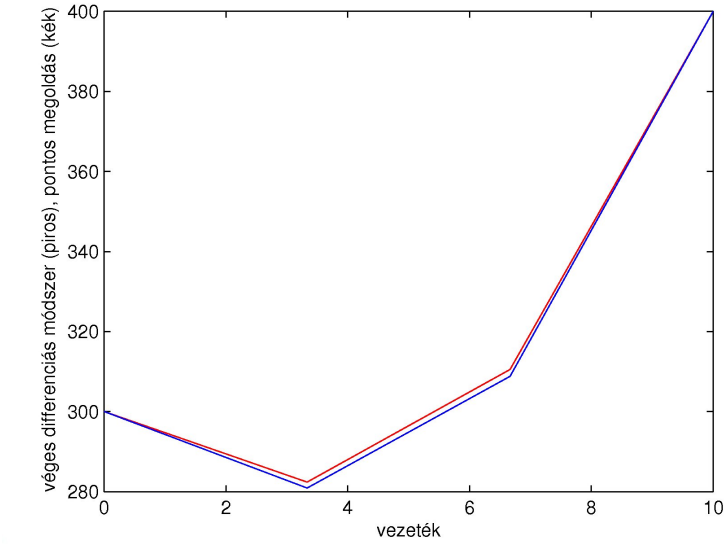
8.8. táblázat. A különböző módszerek eredményei a tesztfeladatra $Nx = 8$ felosztás esetén. A pontos megoldás piros, az explicit Euler-módszeres belövéses módszeré zöld, a javított explicit Euler-módszeres belövéses módszeré fekete, a véges differenciás megoldásé pedig kék.



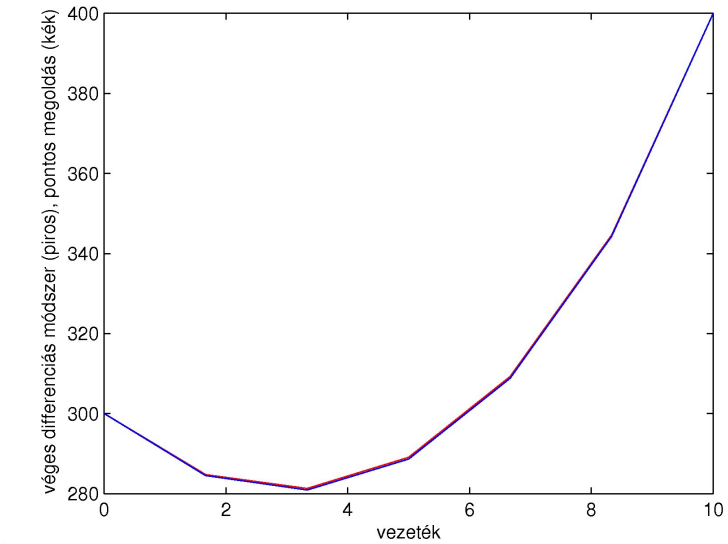
8.8. ábra. A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben 10000 osztáspontú ekvidisztáns rácshálón.



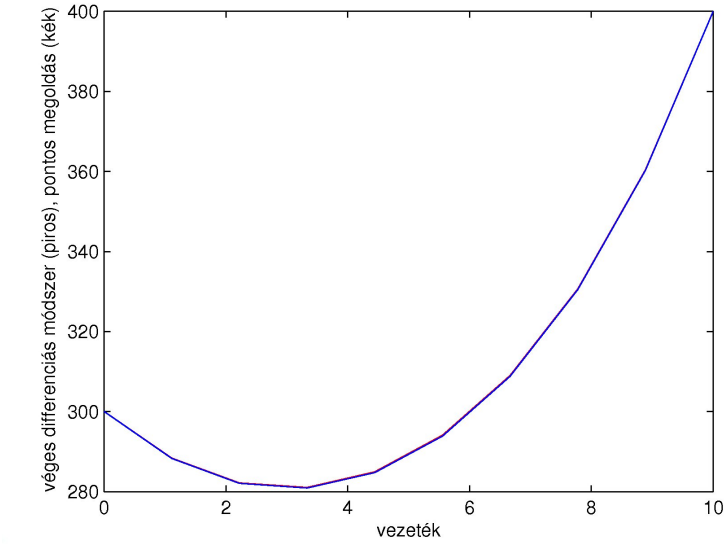
8.9. ábra. A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben 10 osztáspontú ekvidisztáns rácshálón két alappontra ($Nz = 2$) támaszkodó interpolációval



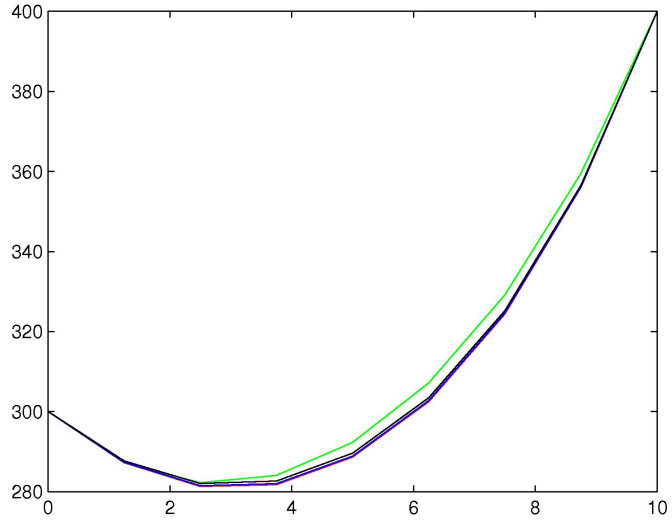
8.10. ábra. A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben véges differenciákkal a négy pontból álló ekvidisztáns rácshálón.



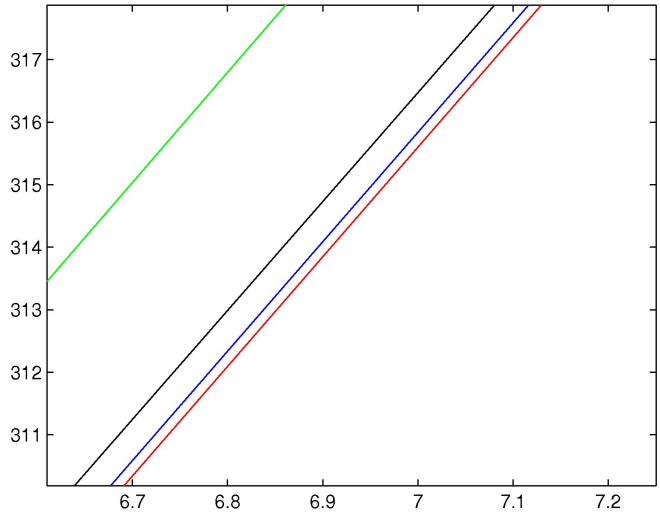
8.11. ábra. A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben véges differenciákkal a hét pontból álló ekvidisztáns rácshálón.



8.12. ábra. A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben véges differenciákkal a tíz pontból álló ekvidisztáns rácshálón.



8.13. ábra. A stacionárius hőmérsékleteloszlás az $L = 10m$ hosszúságú vezetékben a különböző módszerekkel a $\Delta x = 1.25$ lépésközű ekvidisztáns rácshálón.



8.14. ábra. Kinagyítva az előző ábra egy részletét, a pontos megoldás színe piros, az explicit Euler-módszeres belövéses módszeré zöld, a javított explicit Euler-módszeres belövéses módszeré fekete, a véges differenciás megoldásé pedig kék.

Irodalomjegyzék

[1] Asher, U., Petzold, L. Computer Methods for Ordinary Differential Equations. SIAM, Philadelphia, 1998.

[2] Berman, A., Plemmons, R. J. Nonnegative Matrices in the Mathematical Sciences. New York, SIAM, 1987.

[3] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis. Heidelberg, Springer, 1980.

[4] Coddington, E. A., Levinson, N. Theory of differential equations. New York, McGraw-Hill, 1955.

[5] Faragó I., Horváth R. Numerikus módszerek. Typotech, 2011. (internetjegyzet) http://www.tankonyvtar.hu/

[6] Galántai A., Jeney A. Numerikus módszerek. Miskolci Egyetemi Kiadó, 2002

[7] Hairer, E., Norsett, S P., Wanner, G. Solving Ordinary Differential Equations I: Nonstiff Problems. Springer Verlag, 1990.

[8] Kincaid, D., Cheney, W. Numerical Analysis. Mathematics of Scientific Computing, American Mathematical Society, 2009.

[9] Lambert, J. D. Numerical Methods for Ordinary Differential Systems, Wiley Publ., 2000.

[10] Larsson, S., Thomee, V. Partial Differential Equations with Numerical Methods. Springer, 2005.

[11] Petruska Gy. Komplex függvénytan. Nemzeti Tankönyvkiadó, 1998.

[12] Simon L., Baderko E. A. Másodrendű lineáris parciális differenciálegyenletek. Tankönyvkiadó, 1983.

[13] Simon L., Czách L. Parciális differenciálegyenletek. Tankönyvkiadó, 1974.

[14] Süli, E., Mayers, D. F. An Introduction to Numerical Analysis. Cambridge University Press, 2003.

[15] Rontó M. Dinamikus gazdasági modellek. Miskolci Egyetem, 2006. (internetjegyzet) http://www.uni-miskolc.hu/~matronto/

[16] Rózsa P. Lineáris algebra és alkalmazásai. Műszaki Könyvkiadó, Budapest, 1974.

[17] Simon P., Tóth J. Differenciálegyenletek. Bevezetés az elméletben és az alkalmazásokba. Typotech, Budapest, 2005.

[18] Stoyan G. Matlab. Typotex, Budapest, 2005.

[19] Stoyan G., Takó G. Numerikus módszerek 2. Typotex, Budapest, 1995.

[20] Vinogradov, R. E. A közönséges differenciálegyenlet-rendszer Ljapunov-értelembe instabil megoldásának egy kritériuma., DAN, 84, 1952, 201-204. (oroszul)