EÖTVÖS LORÁND TUDOMÁNYEGYETEM
INFORMATIKAI KAR

KÖZÖNSÉGES DIFFERENCIÁLEGYENLETEK NUMERIKUS MÓDSZEREI

JEGYZET

Krebsz Anna

Lektorálta: dr. Hegedűs Csaba

A jegyzet az ELTE Informatikai Karának 2014. évi Jegyzetpályázatának támogatásával készült.

Budapest, 2014

Tartalomjegyzék

Előszó **4**

1. A Csererendszer modell **5**

2. Alapfogalmak **8**
2.1. Monoton mátrixok 8
2.2. M-mátrixok 10
2.3. A mátrix exponenciális függvény 17
2.4. Logaritmikus norma 19
2.5. Állandó együtthatós lineáris differenciaegyenletek 26
2.6. Differenciálegyenletek alapvető tulajdonságai 32

3. Numerikus módszerek bevezetése **44**
3.1. Alapfogalmak 44
3.2. A legegyszerűbb numerikus módszerek 49
3.2.1. Euler-módszer 49
3.2.2. Implicit Euler-módszer 51
3.2.3. Javított (módosított) Euler-módszer 52
3.3. A stabilitás általános definíciója 54
3.4. Változó lépéstávolság 55

4. Explicit Runge–Kutta-módszerek (ERK) **57**
4.1. Az ERK-módszerek általános jellemzése 57
4.2. Harmadrendű ERK-módszerek 58
4.3. Az ERK-módszerek stabilitása és konvergenciája 62
4.4. Ismert ERK-módszerek 64
4.5. Összefüggések a lépcsőszám és a rend között 65
4.6. Beágyazott módszerek 65
4.7. Az ERK-módszerek hatékonysága 67
4.8. Hibabecslések 68
4.9. Lépésválasztás 72

5. Lineáris többlépéses módszerek (LTM) **75**
5.1. Általános lineáris többlépéses módszerek 75
5.2. Adams-módszerek 76
5.3. A középpont szabály 80
5.4. Többlépéses módszerek 0-stabilitása 83
5.5. Többlépéses módszerek konzisztenciája 85
5.6. 0-stabil többlépéses módszerek maximális rendje 92

6. Implicit Runge–Kutta-módszerek (IRK) 93
6.1. IRK-módszerek 93
6.2. IRK-módszerek konstrukciója 93
6.3. IRK-módszerek konvergenciája 96
6.4. Rosenbrock-módszerek 97

7. Stabilitás 99
7.1. Belső, lényegi instabilitás 99
7.2. Aszimptotikus stabilitás 99
7.3. Merev (stiff) differenciálegyenletek 106

8. Közönséges differenciálegyenletek peremérték feladatai 109
8.1. Peremérték feladatok 109
8.2. Fredholm alternatíva tétel 112
8.3. A másodrendű egyenlet és klasszikus peremfeltételei 116
8.4. A peremérték feladatok kondicionáltsága 117
8.5. Egy modellfeladat 121

9. Véges differencia eljárások 124
9.1. Bevezetés, alapvető fogalmak 124

ELŐSZÓ

A jegyzet az ELTE IK MSc Modellalkotó szakirányának Közönséges differenciálegyenletek numerikus megoldása tárgyához készült. A jegyzet alapjául szolgáló irodalom Stoyan Gisbert, Takó Galina: Numerikus módszerek 2. könyve, [9] volt.

Olyan jegyzetet szerettem volna készíteni, amely tartalmazza az elméletet a fontos definíciókkal, tételekkel és részletes bizonyításokkal. Emellett példákon keresztül segíti a megértést és a kitűzött feladatok megoldásával a tudás elmélyítését.

A logaritmikus normáról szóló fejezet feladataihoz dr. Hegedűs Csaba kézzel írt előadásjegyzetét és a [12], [5], [7] irodalmakat alapul véve készítettem. [6] felhasználásával készült a Runge–Kutta-módszerek és a Lineáris többlépéses módszerek fejezet. Mivel a BSc képzésben nem tananyag a diffenciaegyenletek elmélete, a lineáris többlépéses módszerek vizsgálatához viszont nélkülözhetetlen, ezért egy alfejezet került a jegyzetbe [1] és [4] felhasználásával.

Ezúton szeretnék köszönetet mondani dr. Hegedűs Csabának, akivel a jegyzet születésekor hétről hétre megbeszélhettem a soron következő anyagrészt. Tanácsot adott, hogyan lehetne rövidebben, egyszerűbben bizonyítani. Köszönöm ezúton is, hogy bármikor fordulhattam hozzá kérdéseimmel.

Köszönöm továbbá az ELTE IK MSc Modellalkotó Informatikus szakirány 2013/14-es tanév I. féléves hallgatóinak a jegyzettel kapcsolatos megjegyzéseit és a hibák felderítését. Fiamnak a tanácsokat és az anyaghoz elkészített programokat. Utoljára és legfőképpen páromnak köszönöm a segítségét, mellyel levette a vállamról az otthoni munkák terhét.

Budapest, 2014. november 7.

Krebsz Anna

1. fejezet

A Csererendszer modell

[SG 10.1.1] Tekintsünk egy gazdasági vagy ökológiai rendszert, ami n alrendszerből áll. Az alrendszerek termelnek árut, energiát, koncentrációkat és azokat cserélik ki egymással és a külvilággal. A következő adatokkal dolgozunk:

- c_j: „általánosított koncentráció" a j. alrendszer állapotát mutatja a közös egységben (pl. pénzben).
- $a_{ij}c_j\tau$: mutatja, hogy a j. alrendszer τ időegység alatt ennyi egységet ad át az i. alrendszernek $(i \neq j) - re$.
- $d_jc_j\tau$: mutatja, hogy a j. alrendszer τ időegység alatt ennyi egységet ad át a külvilágnak.
- $b_j\tau$: mutatja, hogy a j. alrendszer τ időegység alatt ennyi egységet kap a külvilágtól.

A j. alrendszer által elvesztett és kapott mennyiség a $[t; t+\tau]$ időintervallumban

$$Veszt_j = \tau \left(\sum_{i=1, i\neq j}^{n} a_{ij} + d_j \right) c_j, \quad Kap_j = \tau \left(\sum_{i=1, i\neq j}^{n} a_{ji}c_i + b_j \right).$$

Vezessük be a következő jelöléseket a mátrixos alak felírásához.

$$\mathbf{A} = (\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}) = (a_{ij}), \quad a_{ii} = 0, \quad \mathbf{e} = (1, 1, \ldots, 1)^T$$

A j. alrendszer által elvesztett (átadott) és kapott mennyiség a $[t; t+\tau]$ időintervallumban

$$Veszt_j = \tau(\mathbf{e^T a_j} + d_j)\, c_j, \quad Kap_j = \tau(\mathbf{Ac})_j + \tau b_j.$$

A $t+\tau$ időpontban a $c_j(t+\tau)$ koncentráció a $c_j(t)$ koncentrációból számítható a „mérlegegyenlet" alapján.

$$c_j(t+\tau) = c_j(t) + Kap_j - Veszt_j$$

Részletesen felírva

$$c_j(t+\tau) = c_j(t) + \tau(\mathbf{Ac}(t))_j + \tau b_j - \tau(\mathbf{e^T a_j} + d_j)\, c_j(t) = c_j(t) + \tau(\mathbf{b} - \mathbf{Bc}(t))_j,$$

ahol

$$\mathbf{B} = (b_{ij}) = -\mathbf{A} + diag(\mathbf{e^T a_j} + d_j), \quad b_{ij} = \begin{cases} \mathbf{e^T a_j} + d_j & \text{ha } i = j \\ -a_{ij} & \text{ha } i \neq j. \end{cases}$$

Az egyenletet átrendezve és τ-val leosztva.

$$\frac{c_j(t+\tau) - c_j(t)}{\tau} = (\mathbf{Ac}(t))_j + b_j - (\mathbf{e^T a_j} + d_j)\, c_j(t) = (\mathbf{b} - \mathbf{Bc}(t))_j.$$

$\tau \to 0$ esetén a

$$\mathbf{c}' = \mathbf{b} - \mathbf{B}\mathbf{c}, \quad \mathbf{c}(0) = \mathbf{c_0}$$

közönséges differenciálegyenletet kapjuk, melyet még ki kell egészítenünk a kezdetiértékekkel, hogy a megoldás egyértelmű legyen. A $\mathbf{c}(t)$ vektorfüggvény deriválását elemenként értjük. Ezzel elkészült a *csererendszer* matematikai modellje.

A kezdetiértékproblémát megoldva arra kaphatunk választ, hogyan reagál a rendszer a külső (b_i) vagy a belső (a_{ij}) változásokra, $t \to \infty$ esetén beáll-e a stacionárius állapot. Ha az összes $d_i > 0$, akkor a $\mathbf{B}$ mátrixnak nem csak az előjel elrendeződése megfelelő, hanem főátló domináns is az oszlopaira nézve és M-mátrix is. Ebből következnek a fenti modell előnyös tulajdonságai.

1-1. Példa. Konkrét példaként tekintsük a Balaton szennyeződésének egy egyszerű *kompartment* (csererendszer) modelljét. A tavat 3 részre osztjuk:

- 1. medence: Keszthely/Szigliget
- 2. medence: Szemes
- 3. medence: Siófok.

Minden medencében (azaz alrendszerben, *kompartment*ben) feltételezzük, hogy egységes a koncentráció. Ennek a *jó keveredés*nek a feltételezése éppen azt eredményezi, hogy nem parciális, hanem közönséges differenciálegyenletekre jutunk.

Az egyes medencék között mindkét irányban történik áramlás. A modell a_{ij} számai azt a víztérfogatot jelölik, mely időegység alatt a j-edik medencéből az i-edikbe áramlik. A könnyebb számolás miatt tényleges mérési eredmények helyett

$$a_{21} = 2, \;\; a_{12} = 1, \;\; a_{31} = a_{13} = 0, \;\; a_{32} = 1.2, \;\; a_{23} = 0.2.$$

Így az $\mathbf{A}$ mátrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 2 & 0 & 0.2 \\ 0 & 1.2 & 0 \end{bmatrix}.$$

Az i. medencében a víztérfogat változását α_i-val jelöljük.

$$\alpha_i = \sum_{j=1, j\neq i}^{n} (a_{ji} - a_{ij}).$$

$\alpha_1 = 1$, az 1. medence a Zalából
$\alpha_2 = 0$, a 2. medence kívülről nem kap utánpótlást
$\alpha_3 = -1$, a 3. medence vízvesztesége a Sióba folyik.
A kiáramló víz mennyisége az egyes medencékben:

$$d_1 = d_2 = 0, \quad d_3 = 1 = -\alpha_3.$$

A modellben tükröződik a szállítóanyag (víz) megmaradása. A definiált $\mathbf{B}$ mátrixunk és inverze

$$\mathbf{B} = \underbrace{\begin{bmatrix} 0 & -1 & 0 \\ -2 & 0 & -0.2 \\ 0 & -1.2 & 0 \end{bmatrix}}_{-\mathbf{A}} + \underbrace{\begin{bmatrix} 2+0 & 0 & 0 \\ 0 & 2.2+0 & 0 \\ 0 & 0 & 0.2+1 \end{bmatrix}}_{diag(\mathbf{e^T a_j} + d_j)} = \begin{bmatrix} 2 & -1 & 0 \\ -2 & 2.2 & -0.2 \\ 0 & -1.2 & 1.2 \end{bmatrix}$$

$$\Rightarrow \quad \mathbf{B^{-1}} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{12} \\ 1 & 1 & \frac{1}{6} \\ 1 & 1 & 1 \end{bmatrix} > \mathbf{0},$$

ahol a pozitivitást elemenként értjük. A későbbiekben látni fogjuk, hogy $\mathbf{B}$ M-mátrix, ugyanis az átlón kívüli elemei nem pozitívak és megadható olyan pozitív elemű $\mathbf{g}$ vektor, hogy $\mathbf{Bg} > \mathbf{0}$.

$$\mathbf{g} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} > \mathbf{0} \quad \Rightarrow \quad \mathbf{Bg} = \begin{bmatrix} 1 \\ 1.8 \\ 1.2 \end{bmatrix} > \mathbf{0}$$

Ezen kívül $\mathbf{B}$ minden sajátértéke valós és pozitív:

$$\lambda_1 \approx 0.51, \quad \lambda_2 \approx 1.32, \quad \lambda_3 \approx 3.57.$$

A bevitt szennyeződés tömegét időegységenként jelöljük b_j-vel. Pl.

$$b_1 = 30, \quad b_2 = 20, \quad b_3 = 40.$$

Feltételezzük, hogy kezdetben nincs szennyeződés. Ezután a (passzív) környezetvédelem néhány kérdésére kereshetünk választ, pl. a koncentráció küszöbértékeivel kapcsolatban.

2. fejezet

Alapfogalmak

2.1. Monoton mátrixok

A valós számok körében, ha $0 < a \leq b$, akkor $\frac{1}{a} \geq \frac{1}{b}$. Mátrixok körében ez általában nem teljesül.

2-1. Példa. Tekintsük a következő invertálható mátrixokat.

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}$$

Mutassuk meg, hogy bár $\mathbf{A} \leq \mathbf{B}$ (elemenként értve a relációt), mégsem igaz a fordított reláció az inverzeikre.

Megoldás.

$$\mathbf{A^{-1}} = -\frac{1}{2}\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}, \quad \mathbf{B^{-1}} = -\frac{1}{2}\begin{bmatrix} 5 & -3 \\ -4 & 2 \end{bmatrix} \quad \Rightarrow \quad \mathbf{B^{-1} - A^{-1}} = -\frac{\mathbf{1}}{\mathbf{2}}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

Látjuk, hogy a kapott eredmény nem pozitív elemű mátrix. ■

2-1. Definíció. Az $\mathbf{A} \in \mathbb{R}^{n\times n}$ invertálható mátrixot monoton mátrixnak nevezzük, ha $\mathbf{A^{-1}} \geq \mathbf{0}$ (minden eleme nemnegatív).

Jelölés: Az $\mathbf{x} \leq \mathbf{y}$ és $\mathbf{A} \leq \mathbf{B}$ relációkat elemenként értjük. Csak részben rendezettséget ad.

2-2. Példa. Melyek a 2×2-es monoton mátrixok?

Megoldás. Legyen $a, b, c, d \geq 0$ és $\det(\mathbf{A}) = ad - bc > 0$.

$$\mathbf{A} = \begin{bmatrix} a & -b \\ -c & d \end{bmatrix}, \quad \mathbf{A^{-1}} = \frac{1}{ad-bc}\begin{bmatrix} d & b \\ c & a \end{bmatrix} \geq \mathbf{0}$$

A $\det(\mathbf{A}) = ad - bc < 0$ esethez $a, b, c, d \leq 0$ feltétel kell. ■

2-1. T *Legyen* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *monoton mátrix,* $\mathbf{x}, \mathbf{y}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^n$*, melyekre* $\mathbf{Ax} = \mathbf{b}$ *és* $\mathbf{Ay} = \mathbf{c}$*. Ekkor*

$$\mathbf{b} \geq \mathbf{c} \quad \Rightarrow \quad \mathbf{x} \geq \mathbf{y}.$$

Bizonyítás.

$$\mathbf{x} - \mathbf{y} = \mathbf{A^{-1}}(\mathbf{b} - \mathbf{c}) \geq 0$$

■

2-2. T *Legyenek* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ *monoton mátrixok. Ha*

$$\mathbf{A} \geq \mathbf{B} \quad \Rightarrow \quad \mathbf{B^{-1}} \geq \mathbf{A^{-1}}.$$

Bizonyítás.

$$\mathbf{A} \geq \mathbf{B} \quad \Rightarrow \quad \mathbf{B^{-1}A} \geq \mathbf{B^{-1}B} = \mathbf{I}$$

Az egyenlőtlenség mindkét oldalán nemnegatív elemű mátrixok állnak.
Szorozzuk jobbról mindkét oldalt $\mathbf{A^{-1}}$-zel.

$$\mathbf{B^{-1}} \geq \mathbf{A^{-1}}$$

■

2-3. Példa. Igazoljuk, hogy az $\mathbf{A} = \text{tridiag}(-1, 2, -1)$ mátrix monoton.

Megoldás. 1. mo: Az $\mathbf{A}$ LU-felbontása $\mathbf{A} = \mathbf{LU}$, ahol

$$\mathbf{L} = \text{tridiag}(-\frac{i-1}{i}, 1, 0), \quad \mathbf{U} = \text{tridiag}(0, \frac{i+1}{i}, -1).$$

Oldjuk meg a felbontás segítségével az $\mathbf{Ax_i} = \mathbf{LUx_i} = \mathbf{e_i}$ lineáris egyenletrendszereket $i = 1, \ldots, n$-re. Először oldjuk meg az $\mathbf{Lh_i} = \mathbf{e_i}$,

$$\begin{aligned}
h_1 &= 0 \\
-\frac{1}{2}\,h_1 + h_2 = 0 \to &\quad h_2 = 0 \\
-\frac{i-1}{i}\,h_{i-1} + h_i = 1 \to &\quad h_i = 1 > 0 \\
-\frac{j-1}{j}\,h_{j-1} + h_j = 0 \to &\quad h_j = \frac{j-1}{j}\,h_{j-1} > 0 \quad (j = i+1, \ldots, n)
\end{aligned}$$

majd az $\mathbf{Ux_i} = \mathbf{h_i}$ háromszög alakú egyenletrendszert.

$$\begin{aligned}
\frac{n+1}{n}\,x_n = h_n \to &\qquad x_n = \frac{n}{n+1}\,h_n > 0 \\
\frac{n}{n-1}\,x_{n-1} - x_n = h_{n-1} \to &\quad x_{n-1} = \frac{n-1}{n}\,(h_{n-1} + x_n) > 0 \\
\frac{j+1}{j}\,x_j - x_{j+1} = h_j \to &\qquad x_j = \frac{j}{j+1}\,(h_j + x_{j+1}) > 0 \quad (j = n-2, \ldots, 1)
\end{aligned}$$

A kapott $\mathbf{x_i}$ megoldások lesznek az inverz oszlopai, innen $\mathbf{A^{-1}}$ pozitivitása nyilvánvaló.

2. mo: Belátható, hogy $\mathbf{A^{-1}} = (\alpha_{ij})_{i,j=1}^{n}$ elemei

$$\alpha_{ij} = (n+1) \cdot \begin{cases} i(n+1-j) & \text{ha } i \leq j \\ j(n+1-i) & \text{ha } i \geq j. \end{cases}$$

Innen a pozitivitás nyilvánvaló. ■

Feladatok

2-1. Tekintsük az $\mathbf{A}$=tridiag(a_i, d_i, c_i) 3-átlós mátrixot, ahol

$$\begin{gathered} a_1 = c_n = 0, \\ a_i < 0 \ \ (i = 2, \ldots, n) \quad c_i < 0 \ \ (i = 1, \ldots, n-1) \\ a_i + d_i + c_i \geq 0 \ \ (i = 1, \ldots, n) \ \text{és} \\ \exists\, j: \ \ a_j + d_j + c_j > 0 \end{gathered}$$

Mutassuk meg, hogy monoton mátrix.

2.2. M-mátrixok

2-2. Definíció. Az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix főátló domináns a soraira, ha

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \quad (i = 1, 2, \ldots, n).$$

2-3. Definíció. Az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix *Z*-mátrix, ha $a_{ij} \leq 0$, minden $i \neq j$-re.

2-4. Definíció. Az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix *M*-mátrix, ha *Z*-mátrix és

$$\exists\, \mathbf{g} > \mathbf{0}: \quad \mathbf{Ag} > \mathbf{0}.$$

Az *M*-mátrixok tulajdonságai

2-3. T *Ha* $\mathbf{A}$ *M-mátrix, akkor* $a_{ii} > 0$ *minden* i*-re.*

Bizonyítás. Vegyük a definícióban szereplő $\mathbf{g} > \mathbf{0}$ vektort, melyre $\mathbf{Ag} > \mathbf{0}$.
Írjuk fel az $\mathbf{Ag}$ vektor i. komponensét $(i = 1, 2, \ldots, n)$

$$0 < (\mathbf{Ag})_i = a_{ii}g_i + \underbrace{\sum_{j=1, j\neq i}^{n} a_{ij}g_j}_{\leq 0} \quad \Rightarrow \quad a_{ii}g_i > 0$$

Mivel $j \neq i$-re $a_{ij} \leq 0$ és $g_j > 0$, a szumma értéke nem pozitív,
amiből $g_i > 0$ miatt $a_{ii} > 0$ következik minden i-re. ■

2-4. T *Ha* $\mathbf{A}$ *M-mátrix, akkor a definícióban szereplő* $\mathbf{g}$ *vektorral elkészített* $\mathbf{G} = \mathrm{diag}(\mathbf{g})$ *mátrixra* $\mathbf{AG}$ *főátló domináns a soraira.*

Bizonyítás. Legyen $\mathbf{e} = [1, 1, \ldots, 1]^T$ és $\widetilde{\mathbf{A}} = \mathbf{AG}$.

$$0 < \mathbf{Ag} = \mathbf{AGe} = \widetilde{\mathbf{A}}\mathbf{e} \quad \Leftrightarrow \quad 0 < (\widetilde{\mathbf{A}}\mathbf{e})_i = \sum_{j=1}^{n} \widetilde{a}_{ij} \quad (i = 1, 2, \ldots, n)$$

Mivel $\widetilde{a}_{ij} = a_{ij}g_j \leq 0$ $i \neq j$-re és $\widetilde{a}_{ii} = a_{ii}g_i > 0$, az előző összeget az elemek abszolút értékével is kifejezhetjük.

$$0 < \widetilde{a}_{ii} + \sum_{j=1, j\neq i}^{n} \widetilde{a}_{ij} = |\widetilde{a}_{ii}| - \sum_{j=1, j\neq i}^{n} |\widetilde{a}_{ij}| \quad (i = 1, 2, \ldots, n)$$

Átrendezve éppen a sorokra vonatkozó főátló dominanciát kapjuk. ■

2-5. T *Ha* $\mathbf{A}$ *M-mátrix, akkor a definícióban szereplő* $\mathbf{g}$ *vektorral elkészített* $\mathbf{G} = \mathrm{diag}(\mathbf{g})$ *mátrixra* $\mathbf{G^{-1}AG}$ *főátló domináns a soraira és* $\mathrm{Re}\,(\lambda_i(\mathbf{A})) > 0$.

Bizonyítás. Legyen $\mathbf{e} = [1, 1, \ldots, 1]^T$ és $\mathbf{B} = \mathbf{G^{-1}AG}$, ekkor

$$b_{ij} = \frac{a_{ij}g_j}{g_i} \leq 0 \;\; (i \neq j) \quad \text{és} \quad b_{ii} = a_{ii} > 0.$$

Be kell látnunk, hogy $\mathbf{B}$ főátló domináns a soraira, azaz

$$a_{ii} = |b_{ii}| > \sum_{j=1, j\neq i}^{n} |b_{ij}| = \sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|g_j}{g_i} \quad (i = 1, 2, \ldots, n).$$

Ha az egyenlőtlenséget beszorozzuk a $g_i > 0$ értékkel, akkor

$$a_{ii}g_i > \sum_{j=1, j\neq i}^{n} |a_{ij}|g_j \quad (i = 1, 2, \ldots, n).$$

Ez éppen az $\mathbf{AG}$ mátrix főátló dominanciáját adja, amit a 4. tételben bizonyítottunk.
A sajátértékekre vonatkozó állítást az első egyenlőtlenségből kapjuk, ha a Gersgorin tételt alkalmazzuk a $\mathbf{B}$ mátrixra. Az a_{ii} középpontú Gersgorin körök a jobboldali félsíkon vannak. Másrészt a tételből $\mathbf{B}$ és a hasonlósági transzformáció miatt $\mathbf{A}$ invertálhatósága is következik. ■

2-5. Definíció. Az $\mathbf{A}$ mátrixot stabilnak nevezzük, ha bármely λ_i sajátértékre $\mathrm{Re}\,(\lambda_i) < 0$.

2-6. T *Ha* $\mathbf{A}$ *M-mátrix, akkor* $-\mathbf{A}$ *stabil.*

Bizonyítás. Láttuk, hogy M-mátrixok esetén $\mathrm{Re}\,(\lambda_i) > 0$. A $-\mathbf{A}$ mátrix sajátértékei $-\lambda_i$-k, így $\mathrm{Re}\,(-\lambda_i) < 0$ miatt $-\mathbf{A}$ stabil. ■

2-7. T *A Schur-komplemens megőrzi az M-mátrix tulajdonságot.*

Bizonyítás. A Gauss-eliminációnak csak az első lépését vizsgáljuk. A tanult képletben

$$a_{ij}^{(1)} = \underbrace{a_{ij}}_{\leq 0} - \underbrace{\frac{a_{i1}}{a_{11}} \cdot a_{1j}}_{\geq 0}, \quad i, j = 2, \ldots, n$$

az előjel viszonyokat vizsgálva $i \neq j$-re $a_{11} > 0, a_{i1} \leq 0, a_{1j} \leq 0$, így $a_{ij}^{(1)} \leq 0$.
Tehát az $[\mathbf{A}|a_{11}]$ Schur-komplementerben az előjelviszonyok megfelelőek.
Tekintsük a elimináció 1. lépését leíró $\mathbf{L_1}$ mátrixot és a $\mathbf{g} > 0$ vektort, melyre $\mathbf{Ag} > \mathbf{0}$.
Mivel $\mathbf{L_1} = \mathbf{I} - \mathbf{l_1 e_1^T}$ és $i \neq 1$-re $(\mathbf{l_1})_i = \frac{a_{i1}}{a_{11}} \leq 0$, így $\mathbf{L_1} \geq \mathbf{0}$.

$$0 < \mathbf{L_1 Ag} = \begin{bmatrix} a_{11} & \mathbf{u_1^T} \\ 0 & [\mathbf{A}|a_{11}] \end{bmatrix} \cdot \mathbf{g} = \begin{bmatrix} (\mathbf{Ag})_1 \\ [\mathbf{A}|a_{11}]\mathbf{g'} \end{bmatrix}$$

ahol $\mathbf{g'} = [g_2, \ldots, g_n]^T$. Ezzel megkaptuk az $[\mathbf{A}|a_{11}]$ Schur-komplementerhez az M-mátrix definíciójában szereplő $\mathbf{g'} > 0$ vektort, melyre $[\mathbf{A}|a_{11}]g' > 0$. ■

2-8. T *Ha az* $\mathbf{L}$ *alsóháromszög- és* $\mathbf{U}$ *felsőháromszög mátrix Z-mátrix és a diagonális elemek pozitívak, akkor*

a) $\mathbf{L}$ *és* $\mathbf{U}$ *M-mátrix, továbbá*

b) $\mathbf{L^{-1}} \geq \mathbf{0}$ *és* $\mathbf{U^{-1}} \geq \mathbf{0}$.

Bizonyítás. a) A diagonálison kívüli elemekre a feltételek teljesülnek. Keressünk olyan $\mathbf{g} > \mathbf{0}$ vektort, melyre $\mathbf{Lg} > \mathbf{0}$. Írjuk fel a feltételeket minden komponensre.

$$\begin{aligned} l_{11}g_1 > 0 \qquad & g_1 > 0 \\ l_{21}g_1 + l_{22}g_2 > 0 \rightarrow \quad & g_2 > -\frac{1}{l_{22}} l_{21}g_1 \geq 0 \\ l_{i1}g_1 + \ldots + l_{ii}g_i > 0 \rightarrow \quad & g_i > -\frac{1}{l_{ii}} \left(l_{i1}g_1 + \ldots + l_{ii-1}g_{i-1}\right) \geq 0 \quad (i = 1, \ldots, n) \end{aligned}$$

A fenti egyenlőtlenségek a keresett vektor konstrukcióját mutatják.
Felsőháromszög mátrixra ugyanígy elkészíthető $\mathbf{g} > \mathbf{0} : \mathbf{Ug} > \mathbf{0}$.

b) $\mathbf{L^{-1}}$ meghatározásához az $\mathbf{Lx_i} = \mathbf{e_i}$ egyenleteket kell megoldani $i = 1, \ldots, n$-re, a kapott $\mathbf{x_i}$ vektorok lesznek $\mathbf{L^{-1}}$ oszlopai.

$$\begin{aligned} l_{11}x_{11} = 0 \qquad & x_{11} = 0 \\ l_{21}x_{11} + l_{22}x_{21} = 0 \rightarrow \quad & x_{21} = -\frac{1}{l_{22}} l_{21}x_{21} = 0 \\ l_{i1}x_{11} + \ldots + l_{ii}x_{i1} = 1 \rightarrow \quad & x_{i1} = \frac{1}{l_{ii}} \left(1 - \underbrace{\left[l_{i1}x_{11} + \ldots + l_{i,i-1}x_{i-1,1}\right]}_{=0}\right) > 0 \end{aligned}$$

$j = i + 1, \ldots, n$-re

$$l_{j1}x_{11} + \ldots + l_{jj}x_{j1} = 0 \rightarrow \quad x_{j1} = -\frac{1}{l_{jj}} \left(\underbrace{l_{j1}x_{11} + \ldots}_{=0} + \underbrace{\ldots + l_{j,j-1}x_{j-1,1}}_{\leq 0}\right) \geq 0$$

Felsőháromszög mátrixra ugyanígy bizonyítható az inverz elemeinek előjele.
■

2-9. T *Ha* $\mathbf{A}$ *M-mátrix, akkor* $\mathbf{A^{-1}} \geq \mathbf{0}$. *(*$A \geq$ *reláció elemenként értendő.)*

Bizonyítás. Könnyen meggondolható (lásd feladatok), hogy ha $\mathbf{A}$ M-mátrix, akkor létezik LU-felbontása és $\mathbf{L}, \mathbf{U}$ is M-mátrix. Ekkor $\mathbf{L^{-1}}, \mathbf{U^{-1}} \geq \mathbf{0}$.

$$\mathbf{A^{-1}} = (\mathbf{LU})^{-1} = \mathbf{U^{-1}L^{-1}} \geq 0$$

■

2-10. T *Ha* $\mathbf{A}$ *M-mátrix, akkor* $\mathbf{g} = \mathbf{A^{-1}e}$ *jó a definícióban szereplő* $\mathbf{g}$ *vektornak, ahol* $\mathbf{e} = [1, 1, \ldots, 1]^T$.

Bizonyítás. Mivel $\mathbf{A^{-1}} \geq \mathbf{0}$, így $\mathbf{g} = \mathbf{A^{-1}e} \geq \mathbf{0}$ (, ami kevés) és $\mathbf{Ag} = \mathbf{AA^{-1}e} = \mathbf{e} > \mathbf{0}$.
$\mathbf{g}$ pozitivitását indirekt bizonyítjuk. Tegyük fel, hogy $\exists\, i : \;\; g_i = 0$, ekkor

$$g_i = \sum_{j=1}^{n} (\mathbf{A^{-1}})_{ij} = 0,$$

vagyis $\mathbf{A^{-1}}$ tartalmaz csupa nullákból álló sort. Ez ellentmond az invertálhatóságnak. ■

A fentiek alapján az *M*-mátrix definícióját kicserélhetnénk a következővel.

2-6. Definíció. (Ekvivalens a korábbi definícióval.)
Az $\mathbf{A} \in \mathbb{R}^{n\times n}$ mátrix *M*-mátrix, ha *Z*-mátrix és $\mathbf{A^{-1}} \geq \mathbf{0}$.

Megjegyzések.

1. Ha $\mathbf{A^{-1}} \geq \mathbf{0}$ (pl. monoton mátrix vagy *M*-mátrix), akkor az $\mathbf{Ag} = \mathbf{e}$ megoldása pozitív. Tegyük fel, hogy $\exists\, i: \ g_i = 0$, ekkor

$$0 = g_i = \sum_{j=1}^{n} (\mathbf{A^{-1}})_{ij},$$

vagyis $\mathbf{A^{-1}}$ tartalmaz csupa nullákból álló sort. Ez ellentmond az invertálhatóságnak.

2. Ha az $\mathbf{Ax} = \mathbf{b}$, $\mathbf{b} \geq \mathbf{0}$ és $\mathbf{A}$ *M*-mátrix, akkor $\mathbf{A^{-1}} \geq \mathbf{0}$-ból következik, hogy $\mathbf{x} \geq \mathbf{0}$.

3. Ha $\mathbf{A}$ *M*-mátrix, akkor monoton is, ugyanis $\mathbf{A^{-1}} \geq \mathbf{0}$.

2-11. T *Ha* $\mathbf{A}$ *M-mátrix, akkor* $\mathbf{A^T}$ *is M-mátrix.*

Bizonyítás. $(\mathbf{A^T})_{ij} = a_{ji} \leq 0$ minden $i \neq j$-re, tehát $\mathbf{A^T}$ is *Z*-mátrix.
Válasszuk a $\mathbf{g} = (\mathbf{A^T})^{-1}\mathbf{e} = (\mathbf{A^{-1}})^{\mathbf{T}}\mathbf{e} \geq \mathbf{0}$ vektort. Ez $\mathbf{A^{-1}} \geq \mathbf{0}$ miatt teljesül.
Már csak azt kell belátnunk, hogy nem lehet $\mathbf{g}$ egyik komponense sem 0.
Ha az i. komponense 0 lenne, akkor

$$0 = g_i = \sum_{j=1}^{n} (\mathbf{A^{-1}})_{ji} \quad \Rightarrow \quad (\mathbf{A^{-1}})_{ji} = 0 \quad \forall\, j = 1, \ldots, n$$

vagyis az inverz j. oszlopának összege nulla, ami csak úgy teljesülhet a nemnegativitás miatt, hogy az oszlop minden eleme nulla. Ez ellentmond az invertálhatóságnak. ■

2-12. T *Ha* $\mathbf{A}$ *M-mátrix, akkor* $\exists\, \mathbf{g}, \mathbf{h} > \mathbf{0}$ *és* $\mathbf{G} = \mathrm{diag}(\mathbf{g})$, $\mathbf{H} = \mathrm{diag}(\mathbf{h})$ *diagonális mátrixokra* $\mathbf{HAG}$ *sorok és oszlopok szerint is főátló domináns.*

Bizonyítás. Vezessük be a következő jelöléseket $\mathbf{B} = \mathbf{HAG}$ és
$\mathbf{G} = \mathrm{diag}(\mathbf{g}) = \mathrm{diag}(g_i)$, $\mathbf{H} = \mathrm{diag}(\mathbf{h}) = \mathrm{diag}(h_i)$. Ekkor

$$b_{ij} = h_i a_{ij} g_j \quad (i, j = 1, \ldots, n).$$

Írjuk fel $\mathbf{B}$-re, mit jelent, hogy a sorokra főátló domináns. Használjuk fel az előjelekre tett feltételeket.

$$h_i a_{ii} g_i > \sum_{j=1, j\neq i}^{n} h_i |a_{ij}| g_j \quad \Leftrightarrow \quad a_{ii} g_i > \sum_{j=1, j\neq i}^{n} |a_{ij}| g_j = -\sum_{j=1, j\neq i}^{n} a_{ij} g_j \quad (i = 1, \ldots, n)$$

A kapott feltétel azt jelenti, hogy az $\mathbf{A}$ *M*-mátrix definíciójában szereplő $\mathbf{g} > \mathbf{0} : \mathbf{Ag} > \mathbf{0}$.
Nézzük meg, mit jelent, hogy $\mathbf{B}$ az oszlopokra főátló domináns.

$$h_i a_{ii} g_i > \sum_{j=1, j\neq i}^{n} h_j |a_{ji}| g_i \quad \Leftrightarrow \quad h_i a_{ii} > \sum_{j=1, j\neq i}^{n} h_j |a_{ji}| = -\sum_{j=1}^{n} a_{ji} h_j \quad (i = 1, \ldots, n)$$

Átrendezve

$$(\mathbf{A^T h})_i = h_i a_{ii} + \sum_{j=1}^{n} a_{ji} h_j > 0 \quad (i = 1, \ldots, n).$$

A kapott feltétel épp azt jelenti, hogy $\mathbf{h} > \mathbf{0}$ vektorra $\mathbf{A^T h} > \mathbf{0}$. ■

2-13. L *(Szétbontásból származó mátrix normájának becslése)*
Legyen $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n)$ *diagonális mátrix és* $\mathbf{A_1}, \mathbf{A_2}$ *mátrixok, hogy* $\mathbf{A_1}$ *soraira*

$$\|\mathbf{e_i^T A_1}\|_1 < |d_i| \quad i = 1, \ldots, n.$$

Ekkor

$$\|(\mathbf{A_1} + \mathbf{D})^{-1}\mathbf{A_2}\|_\infty \leq \max_{i=1}^{n} \frac{\|\mathbf{e_i^T A_2}\|_1}{|d_i| - \|\mathbf{e_i^T A_1}\|_1}.$$

Bizonyítás. Az $\mathbf{A_1}$-re tett feltételből következik, hogy $(\mathbf{A_1} + \mathbf{D})$ főátló domináns a soraira, így létezik inverze. A mátrixnorma definícióját felhasználva

$$\|(\mathbf{A_1} + \mathbf{D})^{-1}\mathbf{A_2}\|_\infty \leq \max_{\|\mathbf{x}\|_\infty = 1} \| \underbrace{(\mathbf{A_1} + \mathbf{D})^{-1}\mathbf{A_2} \cdot \mathbf{x}}_{=\mathbf{y}} \|_\infty.$$

Tegyük fel, hogy $|y_i| = \|\mathbf{y}\|_\infty$.

$$\mathbf{A_2} \cdot \mathbf{x} = (\mathbf{A_1} + \mathbf{D}) \cdot \mathbf{y} \quad \Rightarrow \quad \mathbf{D} \cdot \mathbf{y} = \mathbf{A_2} \cdot \mathbf{x} - \mathbf{A_1} \cdot \mathbf{y}$$

Az i. egyenletet felírva és becsülve

$$|d_i y_i| = \left| \sum_{j=1}^{n} a_{ij}^{(2)} x_j - \sum_{j=1}^{n} a_{ij}^{(1)} y_j \right| \leq \sum_{j=1}^{n} |a_{ij}^{(2)}| \cdot |x_j| + \sum_{j=1}^{n} |a_{ij}^{(1)}| \cdot |y_j|.$$

$\|\mathbf{x}\|_\infty = 1$ és $|y_i| = \|\mathbf{y}\|_\infty$ miatt

$$|d_i| \cdot |y_i| \leq \sum_{j=1}^{n} |a_{ij}^{(2)}| + |y_i| \cdot \sum_{j=1}^{n} |a_{ij}^{(1)}| \leq \|\mathbf{e_i^T A_2}\|_1 + |y_i| \cdot \|\mathbf{e_i^T A_1}\|_1.$$

Átrendezve

$$|y_i|(|d_i| - \|\mathbf{e_i^T A_1}\|_1) \leq \|\mathbf{e_i^T A_2}\|_1 \quad \Rightarrow \quad |y_i| \leq \frac{\|\mathbf{e_i^T A_2}\|_1}{|d_i| - \|\mathbf{e_i^T A_1}\|_1.}$$

Mivel nem tudjuk, milyen i-re veszi fel $\mathbf{y}$ a végtelen normáját, ezért a jobboldali tört értékét minden i-re kiértékeljük és maximumot veszünk. Ezzel a lemmát beláttuk. ■

2-14. T *Ha* $\mathbf{A}$ *M-mátrix és* $\mathbf{g} > \mathbf{0}$ *a definícióban szereplő vektor, akkor* $\mathbf{A^{-1}}$*-re a következő becslés adható*

$$\|\mathbf{A^{-1}}\|_\infty \leq \frac{\|\mathbf{g}\|_\infty}{\min_{i=1}^{n} (\mathbf{Ag})_i}.$$

Bizonyítás. Lemmával:

Legyen $\mathbf{G} = \mathrm{diag}(\mathbf{g})$, ekkor $\mathbf{AG} = \mathbf{Ag}$ és $\|\mathbf{G}\|_\infty = \|\mathbf{g}\|_\infty$.

$$\|\mathbf{A^{-1}}\|_\infty = \|\mathbf{GG^{-1}A^{-1}}\|_\infty = \|\mathbf{G(AG)^{-1}}\|_\infty \leq \|\mathbf{(AG)^{-1}}\|_\infty \cdot \|\mathbf{G}\|_\infty$$

Mivel $\mathbf{AG}$ főátló domináns a soraira, ezért alkalmazhatjuk a Lemmát $\|\mathbf{(AG)^{-1}}\|_\infty$-re a következő mátrixokkal.

$$\mathbf{A_2} = \mathbf{I}, \quad \mathbf{D} + \mathbf{A_1} = \mathbf{AG}, \quad \mathbf{D} = \mathrm{diag}(\mathbf{AG}), \quad \mathbf{A_1} = \mathbf{AG} - \mathbf{D}$$

$$\|\mathbf{(AG)^{-1}}\|_\infty \leq \max_{i=1}^{n} \frac{\|\mathbf{e_i^T I}\|_1}{|\mathbf{(AG)}_{ii}| - \|\mathbf{e_i^T(AG - D)_{ii}}\|_1} = \max_{i=1}^{n} \frac{1}{|a_{ii}g_i| - \sum_{j=1,i\neq j}^{n} |a_{ij}g_j|}$$

A mátrixokra ismert előjelfeltételeket felhasználva

$$\|\mathbf{(AG)^{-1}}\|_\infty \leq \max_{i=1}^{n} \frac{1}{a_{ii}g_i + \sum_{j=1,i\neq j}^{n} a_{ij}g_j} = \frac{1}{\min_{i=1}^{n}(\mathbf{Ag})_i}$$

A kapott egyenletlenséget beírva a korábbi becslésbe

$$\|\mathbf{A^{-1}}\|_\infty \leq \|\mathbf{(AG)^{-1}}\|_\infty \cdot \|\mathbf{G}\|_\infty \leq \frac{\|\mathbf{g}\|_\infty}{\min_{i=1}^{n}(\mathbf{Ag})_i}$$

Lemma nélkül:

Tekintsük az $\mathbf{Ax} = \mathbf{b}$ LER-t tetszőleges $\mathbf{b}$-re.

$$\|\mathbf{A^{-1}}\|_\infty = \max_{\mathbf{b}\neq\mathbf{0}} \frac{\|\mathbf{A^{-1}b}\|_\infty}{\|\mathbf{b}\|_\infty} = \max_{\mathbf{b}\neq\mathbf{0}} \frac{\|\mathbf{x}\|_\infty}{\|\mathbf{b}\|_\infty}$$

A továbbiakban $\|\mathbf{x}\|_\infty$-ra fogunk felső becslést adni.
Legyen $\mathbf{g} > \mathbf{0}$ vektor az M-mátrix definícójában szereplő, melyre $\mathbf{Ag} > \mathbf{0}$.
Rögzített $\mathbf{b}$ esetén keressünk olyan $m > 0$ számot, melyre

$$\mathbf{A}(m \cdot \mathbf{g} \pm \mathbf{x}) = m \cdot \mathbf{Ag} \pm \mathbf{b} \geq 0 \quad \Leftrightarrow \quad m \cdot (\mathbf{Ag})_i \pm b_i \geq 0 \quad (i = 1, \ldots, n)$$

Mivel $\mathbf{Ag} > \mathbf{0}$, így osztáskor a reláció iránya nem változik,

$$m \geq \mp \frac{b_i}{(\mathbf{Ag})_i} \quad (i = 1, \ldots, n) \quad \Rightarrow \quad m = \frac{\|\mathbf{b}\|_\infty}{\min_{i=1}^{n}(\mathbf{Ag})_i}$$

jó választás. A LER-t a nemnegatív $\mathbf{A^{-1}}$-zel megszorozva

$$m \cdot \mathbf{g} \pm \mathbf{x} \geq 0 \quad \Leftrightarrow \quad mg_i \geq \mp x_i \quad (i = 1, \ldots, n) \quad \Rightarrow \quad |x_i| \leq m|g_i| \quad (i = 1, \ldots, n).$$

Innen

$$\|x\|_\infty \leq m \cdot \|g\|_\infty = \frac{\|\mathbf{b}\|_\infty \cdot \|\mathbf{g}\|_\infty}{\min_{i=1}^{n}(\mathbf{Ag})_i}.$$

A becslést beírva az inverz normabecslésébe, kapjuk az állítást. ■

2-4. Példa. Készítsünk az $\mathbf{A}$=tridiag(-1,2,-1) mátrixhoz a definícióban szereplő $\mathbf{g} > \mathbf{0}$ vektort.

Megoldás. 1. mo:
Az $\mathbf{A}$ LU-felbontása $\mathbf{A} = \mathbf{LU}$, ahol

$$\mathbf{L} = tridiag(-\frac{i-1}{i}, 1, 0), \quad \mathbf{U} = tridiag(0, \frac{i+1}{i}, -1).$$

Oldjuk meg a felbontás segítségével az $\mathbf{Ag} = \mathbf{LUg} = \mathbf{e}$ LER-t. Először oldjuk meg az $\mathbf{Lh} = \mathbf{e}$,

$$\begin{aligned} h_1 &= 1 \\ -\frac{1}{2}\,h_1 + h_2 = 1 \to &\quad h_2 = 1 + \frac{1}{2}\,h_1 = \frac{3}{2} > 0 \\ -\frac{i-1}{i}\,h_{i-1} + h_i = 1 \to &\quad h_i = 1 + \frac{i-1}{i}\,h_{i-1} > 0 \quad (i = 2, \ldots, n) \end{aligned}$$

majd az $\mathbf{Ug} = \mathbf{h}$ háromszög alakú egyenletrendszert.

$$\begin{aligned} \frac{n+1}{n}\,g_n = h_n \to &\quad g_n = \frac{n}{n+1}\,h_n > 0 \\ \frac{n}{n-1}\,g_{n-1} - g_n = h_{n-1} \to &\quad g_{n-1} = \frac{n-1}{n}\,(h_{n-1} + g_n) > 0 \\ \frac{i+1}{i}\,g_i - g_{i+1} = h_i \to &\quad g_i = \frac{i}{i+1}\,(h_i + g_{i+1}) > 0 \quad (i = n-2, \ldots, 1) \end{aligned}$$

2. mo:
Írjuk fel $\mathbf{g}$-re az $\mathbf{Ag} > \mathbf{0}$ feltételeket.

$$\begin{aligned} 2g_1 - g_2 > 0 \to &\quad 0 < g_2 < 2g_1 \\ -g_1 + 2g_2 - g_3 > 0 \to &\quad g_2 > \frac{1}{2}(g_1 + g_3) \\ -g_{i-1} + 2g_i - g_{i+1} > 0 \to &\quad g_i > \frac{1}{2}(g_{i-1} + g_{i+1}) \quad (i = 3, \ldots, n-1) \\ -g_{n-1} + 2g_n > 0 \to &\quad 0 < g_n < \frac{1}{2}\,g_{n-1} \end{aligned}$$

Ha a g_i értékek egy szigorúan konkáv függvény egyenletes felosztáshoz tartozó értékei, akkor a fenti feltételek teljesülnek. ■

2-5. Példa. Igazoljuk, hogy ha $\mathbf{A}$ M-mátrix,

$$a_{ij} \le b_{ij} \le 0 \;\; i \ne j, \quad \text{és} \quad 0 < a_{ii} \le b_{ii},$$

akkor $\mathbf{B}$ is M-mátrix.

Bizonyítás. A feltétel miatt $\mathbf{B}$ előjeleloszlása megfelel és $\mathbf{A} \le \mathbf{B}$.
Ha $\mathbf{g} > \mathbf{0}$-ra $\mathbf{Ag} > \mathbf{0}$, akkor $\mathbf{Bg} \ge \mathbf{Ag} > \mathbf{0}$. ■

2-6. Példa. Igazoljuk, hogy ha $\mathbf{A}$ M-mátrix és $\mathbf{P}$ permutációmátrix, akkor $\mathbf{P^T AP}$ is M-mátrix.

Bizonyítás. A permutációmátrixszal végzett hasonlósági transzformáció az i, j-edik sorokat és i, j-edik oszlopokat cseréli meg. Az általbeli két elem helyet cserél az átlón kívüli elemek ott maradnak. Ezzel az előjel viszonyok a mátrixon belül nem változnak.
Mivel $\mathbf{A}$ M-mátrix

$$\exists\, \mathbf{g} > \mathbf{0} : \mathbf{Ag} > \mathbf{0} \quad \Rightarrow \quad \mathbf{APP^T g} > \mathbf{0} \quad \Rightarrow \quad \mathbf{P^T AP}(\mathbf{P^T g}) > \mathbf{0},$$

így a $\widetilde{\mathbf{g}} = \mathbf{P^T g}$ választással megkaptuk a keresett pozitív elemű vektort, melyre $\mathbf{P^T AP}\widetilde{\mathbf{g}} > 0$.
■

Feladatok

2-2. Igazoljuk, hogy ha $\mathbf{A} = s \cdot \mathbf{I} - \mathbf{B}$ alakba írható, ahol $b_{ij} \geq 0$ minden i, j-re és $s > \rho(\mathbf{B})$, akkor $\mathbf{A}$ M-mátrix.

2-3. Igazoljuk, hogy ha $\mathbf{A}$ M-mátrix, akkor $\mathbf{A} = s \cdot \mathbf{I} - \mathbf{B}$ alakba írható, ahol $b_{ij} \geq 0$ minden i, j-re és $s > \rho(\mathbf{B})$.

2-4. Igazoljuk, hogy ha $\mathbf{A}$ M-mátrix, akkor az LU-felbontásból kapott $\mathbf{L}$ és $\mathbf{U}$ mátrix is M-mátrix.

2-5. Igazoljuk, hogy ha $\mathbf{A}$ tridiagonális M-mátrix, akkor a Gauss-elimináció végrehajtható.

2-6. Igazoljuk, hogy ha $\mathbf{A}$ M-mátrix, akkor a főminorok pozitívak.

2-7. Igazoljuk, hogy ha $\mathbf{A}$ M-mátrix, akkor az ILU-felbontás végrehajtható és a kapott $\mathbf{L}, \mathbf{U}$ is M-mátrix.

2-8. Igazoljuk, hogy ha $\mathbf{A}$ M-mátrix, akkor a Jacobi-iteráció konvergál.

2-9. Igazoljuk, hogy ha $\mathbf{A}$ M-mátrix, akkor a Gauss–Seidel-iteráció konvergál.

2-10. Igazoljuk, hogy ha $\mathbf{A}$ M-mátrix pontosan akkor, ha van olyan $\mathbf{D} > \mathbf{0}$ diagonális mátrix, hogy $\mathbf{B} = \mathbf{DAD^{-1}}$-re $\frac{1}{2}(\mathbf{B} + \mathbf{B^T})$ pozitív definit (lásd Miroslav Fiedler: Special Matrices and Their Applications in Numerical Mathematics).

2-11. Igazoljuk, hogy ha $\mathbf{A}$ M-mátrix pontosan akkor, ha van van olyan $\mathbf{D} > \mathbf{0}$ diagonális mátrix, melyre $\mathbf{DA} + \mathbf{A^T D}$ pozitív definit (lásd Miroslav Fiedler: Special Matrices and Their Applications in Numerical Mathematics).

2.3. A mátrix exponenciális függvény

Valós és komplex számok esetén az exponenciális függvényt a

$$\sum_{k=0}^{\infty} \frac{x^k}{k!}, \quad x \in \mathbb{K}$$

hatványsor összegfüggvényeként definiáljuk. Mátrixok esetén is ez lesz a definíció.
Legyen $\mathbf{A}$ egy $n \times n$-es mátrix, ekkor

$$e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 + \ldots = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!}.$$

Ez a sor abszolút konvergens, mert a mátrixnorma szorzatra vonatkozó tulajdonságát felhasználva

$$\left\| \frac{\mathbf{A}^k}{k!} \right\| \leq \frac{\|\mathbf{A}\|^k}{k!}, \quad \forall \ k \in \mathbb{N},$$

innen a majoránskritérium feltételét vizsgálva kapjuk, hogy

$$\sum_{k=0}^{\infty} \left\| \frac{\mathbf{A}^k}{k!} \right\| \leq \sum_{k=0}^{\infty} \frac{\|\mathbf{A}\|^k}{k!} = e^{\|\mathbf{A}\|}.$$

A $\phi(t)$ mátrixfüggvényt ugyanígy definiáljuk:

$$\phi(t) = e^{\mathbf{A}t} = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k t^k}{k!}.$$

Az előzőekből következik, hogy minden t-re a végtelen sor abszolút konvergens. Sőt minden véges intervallumon a hatványsor egyenletesen is konvergens, így akárhányszor differenciálható és

$$\phi'(t) = \sum_{k=0}^{\infty} k\frac{\mathbf{A}^k t^{k-1}}{k!} = \sum_{k=1}^{\infty} \mathbf{A}\frac{\mathbf{A}^{k-1} t^{k-1}}{(k-1)!} = \sum_{k=0}^{\infty} \mathbf{A}\frac{\mathbf{A}^k t^k}{k!} = \mathbf{A}e^{\mathbf{A}t}.$$

2-15. T *(A mátrix exponenciális függvény tulajdonságai)*

1. $e^{\mathbf{A}0} = \mathbf{I}$,
2. $e^{\mathbf{A}t}$ *invertálható és inverze* $e^{-\mathbf{A}t}$,
3. $e^{\mathbf{A}t} \cdot e^{\mathbf{A}s} = e^{\mathbf{A}(s+t)}$, $s, t \in \mathbb{R}$,
4. $e^{\mathbf{A}t}$ *deriválható minden* t*-re és* $\left(e^{\mathbf{A}t}\right)' = \mathbf{A}e^{\mathbf{A}t}$.

2-16. T *Ha* $\mathbf{A}$ *egyszerű struktúrájú mátrix (diagonalizálható), azaz létezik* $\mathbf{C}$ *invertálható mátrix, melyre* $\mathbf{A} = \mathbf{CDC^{-1}}$*, ahol* $\mathbf{D}$ *diagonális, (*$\mathbf{C}$ *oszlopai a sajátvektorok). Továbbá* $f(z) = \sum_{k=0}^{\infty} c_k z^k$ *egy konvergens hatványsor összegfüggvénye az origó körüli* $|z| < R$ *körlemez minden pontjában és az összes sajátérték benne van a körben, akkor*

$$f(\mathbf{A}) = \mathbf{C} \cdot f(\mathbf{D}) \cdot \mathbf{C^{-1}}.$$

2-7. Példa. A következő $\mathbf{A}$ mátrix esetén határozzuk meg az $e^{\mathbf{A}}$ mátrixot.

$$\mathbf{A} = \begin{bmatrix} 3 & -3 & 2 \\ -1 & 5 & -2 \\ -1 & 3 & 0 \end{bmatrix}$$

Megoldás. A mátrix sajátértékei: $\lambda_{1,2} = 2$, $\lambda_3 = 4$.
A $\lambda_{1,2} = 2$-höz tartozó sajátvektorok: $\mathbf{v_1} = [1,1,1]^T$, $\mathbf{v_2} = [0,2,3]^T$.
A $\lambda_3 = 4$-hez tartozó sajátvektor: $\mathbf{v_3} = [1,-1,-1]^T$. A spektrálfelbontás:

$$\mathbf{A} = \begin{bmatrix} 3 & -3 & 2 \\ -1 & 5 & -2 \\ -1 & 3 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & -1 \\ 1 & 3 & -1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{2} & \frac{3}{2} & -1 \\ 0 & -1 & 1 \\ \frac{1}{2} & -\frac{3}{2} & 1 \end{bmatrix} = \mathbf{CDC^{-1}}.$$

Innen

$$e^{\mathbf{A}} = \mathbf{C}(e^{\mathbf{D}})\mathbf{C^{-1}} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & -1 \\ 1 & 3 & -1 \end{bmatrix} \cdot \begin{bmatrix} e^2 & 0 & 0 \\ 0 & e^2 & 0 \\ 0 & 0 & e^4 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{2} & \frac{3}{2} & -1 \\ 0 & -1 & 1 \\ \frac{1}{2} & -\frac{3}{2} & 1 \end{bmatrix}.$$

■

2-17. L *(A mátrix exponenciális függvény nemnegativitása)*
Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ *tetszőleges mátrix. Ekkor*

$$e^{\mathbf{A}t} \geq \mathbf{0}, \ \forall\ t \geq 0 \quad \Leftrightarrow \quad a_{ij} \geq 0 \ \ i \neq j.$$

Bizonyítás. $\Rightarrow$: Legyen $t \geq 0$ ekkor az exponenciális mátrixfüggvény Taylor-sorfejtéséből

$$e^{\mathbf{A}t} \geq \mathbf{0} \quad \Rightarrow \quad (e^{\mathbf{A}t})_{ij} = a_{ij}t + \frac{1}{2}(\mathbf{A^2})_{ij}\, t^2 + \frac{1}{3!}(\mathbf{A^3})_{ij}\, t^3 + \ldots \geq 0, \;\; i \neq j.$$

$t \geq 0$-val leosztva az egyenlőtlenséget kapjuk, hogy

$$a_{ij} + \frac{1}{2}(\mathbf{A^2})_{ij}\, t + \frac{1}{3!}(\mathbf{A^3})_{ij}\, t^2 + \ldots \geq 0, \;\; i \neq j.$$

Elégendően kicsi t-t választva ebből következik, hogy $i \neq j$-re $a_{ij} \geq 0$ lehet csak.

$\Leftarrow$: Rögzítsük a $t \geq 0$-t és bontsuk fel $\mathbf{A}\frac{t}{m}$-et két mátrixra: $\mathbf{A_1} = -\mathbf{I}$, $\mathbf{A_2} = \mathbf{I} + \mathbf{A}\frac{t}{m}$, ahol $m > 0$ egész (értékét később fogjuk megadni).

$$\left(e^{\mathbf{A}\frac{t}{m}}\right)^m = \left(e^{\mathbf{A_1}+\mathbf{A_2}}\right)^m = \left(e^{\mathbf{A_1}} \cdot e^{\mathbf{A_2}}\right)^m$$

Mivel $e^{\mathbf{A_1}} = e^{-\mathbf{I}} \geq \mathbf{0}$, elegendő $\mathbf{A_2} \geq \mathbf{0}$-t megmutatni elég nagy m-re.

$$(\mathbf{A_2})_{ij} = a_{ij}\frac{t}{m} \geq 0, \;\; i \neq j.$$

A diagonális elemre

$$(\mathbf{A_2})_{ii} = 1 + a_{ii}\frac{t}{m} \geq 0 \quad \Leftrightarrow \quad a_{ii}t \geq -m \quad \Leftrightarrow \quad -a_{ii}t \leq m.$$

Ha $a_{ii} \geq 0$, akkor $(\mathbf{A_2})_{ii} \geq 0$ triviális.
Az $a_{ii} < 0$ esetben válasszuk a feltételnek eleget tevő m-et (minden i-re), majd ezek közül a legnagyobbat. Az exponenciális mátrixfüggvény Taylor-sorfejtéséből

$$e^{\mathbf{A_2}} = \mathbf{I} + \mathbf{A_2} + \frac{1}{2}\mathbf{A_2^2} + \ldots \geq \mathbf{0},$$

ami csupa nemnegatív mátrix összegéből áll. ■

Megjegyzés.
Ha $\mathbf{A}$ nem konstans mátrix, hanem $\mathbf{A} = \mathbf{A}(t)$ folytonos t-ben, akkor $a_{ij} \geq 0, \;\; i \neq j$ még mindig elégséges ahhoz, hogy $\mathbf{c}(t) \geq 0$ legyen, ahol $\mathbf{c}'(t) = \mathbf{A}\mathbf{c}(t) + \mathbf{b}(t)$ és $\mathbf{c}(0) \geq 0$, $\mathbf{b}(t) \geq 0$. A bizonyítást lásd [10] 22. oldalán.

2.4. Logaritmikus norma

2-7. Definíció. Legyen $\mathbf{A} \in \mathbb{R}^{n\times n}$ és $\|.\|$ egy vektornorma által indukált mátrixnorma. Ekkor az $\mathbf{A}$ mátrix logaritmikus normája a következő mennyiség:

$$\mu(\mathbf{A}) = \lim_{h\to 0+} \frac{\|\mathbf{I} + h\mathbf{A}\| - 1}{h}.$$

Megjegyezzük, hogy a logaritmikus norma nem mátrixnorma, mert negatív értéke is lehet, de sok tulajdonságában hasonlít a normára. A későbbiekben látni fogjuk, hogy a differenciálegyenletek stabilitásában lesz fontos szerepe. Alsó indexben jelöljük, hogy mely normában számoljuk a logaritmikus normát, például $\mu_2(\mathbf{A})$ az euklideszi normára vonatkozik. Nézzük a speciális $n = 1$ esetet. Ha $\mathbf{A} = a \in \mathbb{R}$, akkor $\mu(\mathbf{A}) = a$, ha pedig $\mathbf{A} = a \in \mathbb{C}$, akkor $\mu(\mathbf{A}) = \mathrm{Re}\,(a)$, vagyis a logaritmikus norma a „valós rész” kiterjesztése. A [5] és a [12] irodalom részletesen tárgyalja a logaritmikus norma történetét és alkalmazásait. A kitűzött feladatok egy része is innen származik.

2-18. T *(A logaritmikus norma tulajdonságai)*

- *a) A logaritmikus norma jól definált és* $|\mu(\mathbf{A})| \leq \|\mathbf{A}\|$.
- *b)* $\mu_2(\mathbf{A}) = \frac{1}{2}\lambda_{\max}(\mathbf{A} + \mathbf{A}^{\mathbf{T}})$
- *c) Ha a* $\|.\|$ *skaláris szorzat segítségével definiált, akkor* $\mu(\mathbf{A}) = \sup_{\|\mathbf{y}\|=1}(\mathbf{A}\mathbf{y}, \mathbf{y})$.
- *d) A c) pont normájával* $\|e^{\mathbf{A}t}\| \leq e^{\mu(\mathbf{A})t}$ *minden* $t \geq 0$*-ra.*

Bizonyítás. a) Először belátjuk, hogy az

$$f(h) = \frac{\|\mathbf{I} + h\mathbf{A}\| - 1}{h} : \mathbb{R} \to \mathbb{R}$$

függvény $h > 0$-ra monoton növő és alulról korlátos. A vektornorma összegre és különbségre vonatkozó háromszög egyenlőtlenségéből és az indukált normákra ismert $\|\mathbf{I}\| = 1$-ből

$$1 - h\|\mathbf{A}\| \leq \|\mathbf{I} + h\mathbf{A}\| \leq 1 + h\|\mathbf{A}\|$$

$$-\|\mathbf{A}\| \leq \frac{\|\mathbf{I} + h\mathbf{A}\| - 1}{h} \leq \|\mathbf{A}\|.$$

$0 < h' < h$ esetén újra a háromszög egyenlőtlenséggel

$$\|\mathbf{I} + h'\mathbf{A}\| = \left\| \left(1 - \frac{h'}{h}\right)\mathbf{I} + \frac{h'}{h}(\mathbf{I} + h\mathbf{A}) \right\| \leq 1 - \frac{h'}{h} + \frac{h'}{h}\|\mathbf{I} + h\mathbf{A}\| = 1 + h'\frac{\|\mathbf{I} + h\mathbf{A}\| - 1}{h},$$

$$\frac{\|\mathbf{I} + h'\mathbf{A}\| - 1}{h'} \leq \frac{\|\mathbf{I} + h\mathbf{A}\| - 1}{h},$$

tehát az f monoton, korlátos függvénynek létezik határértéke 0-ban jobbról.

$$\lim_{h \to 0+} f(h) = \mu(\mathbf{A})$$

b) A valós skaláris szorzatban

$$(\mathbf{A}\mathbf{y}, \mathbf{y}) = (\mathbf{y}, \mathbf{A}^{\mathbf{T}}\mathbf{y}) = (\mathbf{A}^{\mathbf{T}}\mathbf{y}, \mathbf{y}) \quad \Rightarrow \quad (\mathbf{A}\mathbf{y}, \mathbf{y}) = \left(\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{A} + \mathbf{A}^{\mathbf{T}})\mathbf{y}, \mathbf{y}\right).$$

Az $\mathbf{A_{sym}} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^{\mathbf{T}})$ szimmetrikus mátrix, az $\mathbf{A}$ szimmetrikus részének is nevezik. A Rayleigh-hányadosról tanultak szerint

$$\frac{(\mathbf{A}\mathbf{y}, \mathbf{y})}{(\mathbf{y}, \mathbf{y})} = \frac{\left(\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{A} + \mathbf{A}^{\mathbf{T}})\mathbf{y}, \mathbf{y}\right)}{\|\mathbf{y}\|^2} \leq \lambda_{\max}\left(\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{A} + \mathbf{A}^{\mathbf{T}})\right).$$

Itt a maximális sajátérték negatív is lehet. A $\mathbf{B} := \mathbf{I} + h\mathbf{A}$ jelöléssel

$$\|\mathbf{B}\|_2^2 = \lambda_{\max}(\mathbf{B}^{\mathbf{T}}\mathbf{B})$$

$$\mathbf{B}^{\mathbf{T}}\mathbf{B} = (\mathbf{I} + h\mathbf{A})^T(\mathbf{I} + h\mathbf{A}) = \mathbf{I} + h(\mathbf{A} + \mathbf{A}^{\mathbf{T}}) + h^2\mathbf{A}^{\mathbf{T}}\mathbf{A}$$

A kapott kifejezés $\mathbf{I} + h\mathbf{C}$ alakú, melynek sajátértékei $\lambda_i = 1 + h\lambda_i(\mathbf{C})$ alakúak.

$$\lambda_{\max}(\mathbf{B}^{\mathbf{T}}\mathbf{B}) = 1 + h\lambda_{\max}(\mathbf{A} + \mathbf{A}^{\mathbf{T}} + h\mathbf{A}^{\mathbf{T}}\mathbf{A}).$$

A Bauer-Fike tételt illetve a sajátértékek folytonos függését a mátrix elemeitől felhasználva

$$\lambda_{\max}(\mathbf{A} + \mathbf{A}^{\mathbf{T}} + h\mathbf{A}^{\mathbf{T}}\mathbf{A}) = \lambda_{\max}(\mathbf{A} + \mathbf{A}^{\mathbf{T}}) + O(h).$$

A kapott eredményeket egyberakva

$$\|\mathbf{I}+h\mathbf{A}\|_2^2 = \|\mathbf{B}\|_2^2 = 1 + h\lambda_{\max}(\mathbf{A}+\mathbf{A}^{\mathbf{T}}) + O(h^2).$$

A gyökvonáshoz a Taylor-formulát használjuk. Vezessük be a következő valós függvényt:

$$g(h) := \sqrt{1+h\lambda+Kh^2}, \quad \lambda = \lambda_{\max}(\mathbf{A}+\mathbf{A}^{\mathbf{T}}), \quad O(h^2) \le Kh^2.$$

$$g(0) = 1, \quad g'(h) := \frac{\lambda + 2Kh}{2\sqrt{1+h\lambda+Kh^2}} \quad \to \quad g'(0) = \frac{\lambda}{2}$$

$$g(h) = 1 + \frac{\lambda}{2}\,h + O(h^2).$$

Ezt felhasználva

$$\|\mathbf{I}+h\mathbf{A}\|_2 = 1 + \frac{1}{2}\,h\lambda_{\max}(\mathbf{A}+\mathbf{A}^{\mathbf{T}}) + O(h^2)$$

$$\frac{\|\mathbf{I}+h\mathbf{A}\|_2 - 1}{h} = \frac{1}{2}\,\lambda_{\max}(\mathbf{A}+\mathbf{A}^{\mathbf{T}}) + O(h),$$

ahonnan $h \to 0+$ esetén következik az állítás.

c) Legyen $\|\mathbf{y}\| = 1$ tetszőleges vektor.

$$\frac{\|(\mathbf{I}+h\mathbf{A})\mathbf{y}\| - 1}{h} = \frac{(\|(\mathbf{I}+h\mathbf{A})\mathbf{y}\| - 1)(\|(\mathbf{I}+h\mathbf{A})\mathbf{y}\| + 1)}{h(\|(\mathbf{I}+h\mathbf{A})\mathbf{y}\| + 1)} = \frac{\|(\mathbf{I}+h\mathbf{A})\mathbf{y}\|^2 - 1}{h(\|(\mathbf{I}+h\mathbf{A})\mathbf{y}\| + 1)}$$

Írjuk át a valós skaláris szorzat segítségével a számlálót más alakba

$$\|(\mathbf{I}+h\mathbf{A})\mathbf{y}\|^2 - 1 = (\mathbf{y}+h\mathbf{A}\mathbf{y}, \mathbf{y}+h\mathbf{A}\mathbf{y}) - 1 = \underbrace{\|\mathbf{y}\|^2}_{=1} + 2h(\mathbf{A}\mathbf{y},\mathbf{y}) + h^2\|\mathbf{A}\mathbf{y}\|^2 - 1 =$$

$$= 2h(\mathbf{A}\mathbf{y},\mathbf{y}) + h^2\|\mathbf{A}\mathbf{y}\|^2.$$

Az előző törtbe visszaírva

$$\frac{\|(\mathbf{I}+h\mathbf{A})\mathbf{y}\|^2 - 1}{h\|(\mathbf{I}+h\mathbf{A})\mathbf{y}\| + 1)} = \frac{2(\mathbf{A}\mathbf{y},\mathbf{y}) + h\|\mathbf{A}\mathbf{y}\|^2}{(\|(\mathbf{I}+h\mathbf{A})\mathbf{y}\| + 1)} \to (\mathbf{A}\mathbf{y},\mathbf{y}) \;\; (h \to 0+).$$

Ugyanis a baloldalon álló tört h-ban folytonos. Az $\|\mathbf{y}\| = 1$ által meghatározott kompakt halmazon pedig felveszi a maximumát. Innen

$$\mu(\mathbf{A}) = \max_{\|\mathbf{y}\|=1} (\mathbf{A}\mathbf{y},\mathbf{y}).$$

d) Tegyük fel, hogy μ-re

$$\mu\|\mathbf{v}\|^2 \ge (\mathbf{A}\mathbf{v},\mathbf{v}), \;\; \forall\, \mathbf{v} \in \mathbb{R}^n.$$

Ez a c) pont alapján azt jelenti, hogy $\mu \ge \mu(\mathbf{A})$.
Tekintsük az $\mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t)$ differenciálegyenletet $t \ge 0$-ra és szorozzuk skalárisan $\mathbf{y}(t)$-vel

$$\mu\|\mathbf{y}(t)\|^2 \ge (\mathbf{A}\mathbf{y}(t),\mathbf{y}(t)) = (\mathbf{y}'(t),\mathbf{y}(t)) = \sum_{i=1}^n y_i'(t)y_i(t) = \frac{1}{2}\left(\sum_{i=1}^n y_i'(t)2y_i(t)\right) =$$

$$= \frac{1}{2}\left(\sum_{i=1}^n y_i^2(t)\right)' = \frac{1}{2}\left(\|\mathbf{y}(t)\|^2\right)' = \frac{1}{2}\frac{d}{dt}\|\mathbf{y}(t)\|^2 =$$

$$= \left(\frac{1}{2}\left(\sqrt{\sum_{i=1}^n y_i^2(t)}\right)^2\right)' = \frac{1}{2}\cdot 2\cdot\left(\sqrt{\sum_{i=1}^n y_i^2(t)}\right)\cdot\frac{d}{dt}\|\mathbf{y}(t)\| =$$

$$= \|\mathbf{y}(t)\|\frac{d}{dt}\|\mathbf{y}(t)\|$$

A kapott egyenlőtlenséget osztva $\|\mathbf{y}(t)\|$-vel és szorozva $e^{-\mu t}$-vel

$$e^{-\mu t}\mu\|\mathbf{y}(t)\| \geq e^{-\mu t}\frac{d}{dt}\|\mathbf{y}(t)\|.$$

Átrendezve

$$0 \geq -e^{-\mu t}\mu\|\mathbf{y}(t)\| + e^{-\mu t}\frac{d}{dt}\|\mathbf{y}(t)\| = \left(e^{-\mu t}\|\mathbf{y}(t)\|\right)'.$$

Tehát az $(e^{-\mu t}\|\mathbf{y}(t)\|)$ függvény t-ben fogyó, így

$$e^{-\mu t}\|\mathbf{y}(t)\| \leq \|\mathbf{y}(0)\| \quad \Rightarrow \quad \|\mathbf{y}(t)\| \leq e^{\mu t}\|\mathbf{y}(0)\|.$$

Ide beírva a differenciálegyenlet $\mathbf{y}(t) = e^{\mathbf{A}t}\mathbf{y}(0)$ megoldását

$$\|e^{\mathbf{A}t}\mathbf{y}(0)\| \leq e^{\mu t}\cdot\|\mathbf{y}(0)\| \quad \Rightarrow \quad \frac{\|e^{\mathbf{A}t}\mathbf{y}(0)\|}{\|\mathbf{y}(0)\|} \leq e^{\mu t}.$$

Mivel $\mathbf{y}(0)$ tetszőleges, a mátrixnormára is ez a korlát.

$$\|e^{\mathbf{A}t}\| = \sup_{\mathbf{y}(0)\neq\mathbf{0}} \frac{\|e^{\mathbf{A}t}\mathbf{y}(0)\|}{\|\mathbf{y}(0)\|} \leq e^{\mu t}$$

■

Megjegyzés.
A skaláris szorzattal definiált normában felírt logaritmikus normát a

$$\mu(\mathbf{A}) = \max_{\|\mathbf{y}\|=1}(\mathbf{A}\mathbf{y},\mathbf{y}) = \max_{\mathbf{y}\neq 0}\frac{(\mathbf{A}\mathbf{y},\mathbf{y})}{(\mathbf{y},\mathbf{y})}$$

alakban is felírhatjuk. A jobboldalon szereplő hányados a Rayleigh-hányados, melynek tulajdonságairól szimmetrikus mátrix esetén tanultak. Összehasonlításul a norma négyzetére

$$\|\mathbf{A}\|^2 = \max_{\mathbf{y}\neq 0}\frac{\|\mathbf{A}\mathbf{y}\|^2}{\|\mathbf{y}\|^2} = \max_{\mathbf{y}\neq 0}\frac{(\mathbf{A}\mathbf{y},\mathbf{A}\mathbf{y})}{(\mathbf{y},\mathbf{y})} = \max_{\mathbf{y}\neq 0}\frac{(\mathbf{A}^\mathbf{T}\mathbf{A}\mathbf{y},\mathbf{y})}{(\mathbf{y},\mathbf{y})}.$$

Komplex elemű mátrixok és komplex skaláris szorzat esetén az állítás változik:

$$\mu(\mathbf{A}) = \max_{\mathbf{y}\neq 0}\frac{\mathrm{Re}\,(\mathbf{A}\mathbf{y},\mathbf{y})}{(\mathbf{y},\mathbf{y})}$$

2-8. Példa. Számítsuk ki a következő mátrix logaritmikus normáját a spektrálnormában.

$$\mathbf{A} = \begin{bmatrix} -4 & -2 \\ 2 & -6 \end{bmatrix}$$

Megoldás.

$$\frac{1}{2}\left(\mathbf{A}+\mathbf{A}^\mathbf{T}\right) = \frac{1}{2}\left(\begin{bmatrix} -4 & -2 \\ 2 & -6 \end{bmatrix} + \begin{bmatrix} -4 & 2 \\ -2 & -6 \end{bmatrix}\right) = \begin{bmatrix} -4 & 0 \\ 0 & -6 \end{bmatrix}$$

A sajátértékei $\lambda_1 = -4$, $\lambda_2 = -6$, így

$$\mu_2(\mathbf{A}) = \max_i \lambda_i\left(\frac{1}{2}(\mathbf{A}+\mathbf{A}^\mathbf{T})\right) = -4.$$

Az eredményünkből az is látható, hogy

$$\mu_2(-\mathbf{A}) = \max_i \lambda_i\left(\frac{1}{2}(-\mathbf{A}-\mathbf{A}^\mathbf{T})\right) = 6.$$

■

2-9. Példa. Számítsuk ki a következő mátrix logaritmikus normáját a spektrálnormában.

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

Megoldás.

$$\frac{1}{2}\left(\mathbf{A} + \mathbf{A}^{\mathbf{T}}\right) = \frac{1}{2}\left(\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} + \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\right) = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

A sajátértékei $\lambda_1 = 1$, $\lambda_2 = 3$, így

$$\mu_2(\mathbf{A}) = \max_i \lambda_i \left(\frac{1}{2}(\mathbf{A} + \mathbf{A}^{\mathbf{T}})\right) = 3.$$

Az eredményünkből az is látható, hogy

$$\mu_2(-\mathbf{A}) = \max_i \lambda_i \left(\frac{1}{2}(-\mathbf{A} - \mathbf{A}^{\mathbf{T}})\right) = -1.$$

■

2-10. Példa. Számítsuk ki a következő mátrix logaritmikus normáját a ∞ normában.

$$\mathbf{A} = \begin{bmatrix} -4 & -2 \\ 2 & -6 \end{bmatrix}$$

Megoldás.

$$\mu_\infty(\mathbf{A}) = \lim_{h \to 0+} \frac{\|\mathbf{I} + h\mathbf{A}\|_\infty - 1}{h}$$

Mivel $h \to 0+$ határértéket kell számolnunk, ezért feltehető, hogy $h < \frac{1}{6}$, így a diagonális elemek pozitívak.

$$\|\mathbf{I} + h\mathbf{A}\|_\infty = \left\| \begin{bmatrix} 1-4h & -2h \\ 2h & 1-6h \end{bmatrix} \right\|_\infty = \max(1-2h, 1-4h) = 1-2h$$

$$\mu_\infty(\mathbf{A}) = \lim_{h \to 0+} \frac{1-2h-1}{h} = -2$$

■

2-11. Példa. Számítsuk ki a következő mátrix logaritmikus normáját az 1-es normában.

$$\mathbf{A} = \begin{bmatrix} -4 & -2 \\ 2 & -6 \end{bmatrix}$$

Megoldás.

$$\mu_1(\mathbf{A}) = \lim_{h \to 0+} \frac{\|\mathbf{I} + h\mathbf{A}\|_1 - 1}{h}$$

Mivel $h \to 0+$ határértéket kell számolnunk, ezért feltehető, hogy $h < \frac{1}{6}$, így a diagonális elemek pozitívak.

$$\|\mathbf{I} + h\mathbf{A}\|_1 = \left\| \begin{bmatrix} 1-4h & -2h \\ 2h & 1-6h \end{bmatrix} \right\|_1 = \max(1-2h, 1-4h) = 1-2h$$

$$\mu_1(\mathbf{A}) = \lim_{h \to 0+} \frac{1-2h-1}{h} = -2$$

■

2-19. T *(További tulajdonságok) Az* 1*-es és* ∞ *mátrixnormában*

$$\mu_1(\mathbf{A}) = \max_i \left(a_{ii} + \sum_{j \neq i} |a_{ji}| \right)$$
$$\mu_\infty(\mathbf{A}) = \max_i \left(a_{ii} + \sum_{j \neq i} |a_{ij}| \right).$$

Megjegyezzük, hogy komplex elemű mátrixok esetén a_{ii} helyett $\operatorname{Re}(a_{ii})$-t kell írnunk.

Feladatok

2-12. Számítsuk ki a következő mátrixok logaritmikus normáját a spektrálnormában.

$$\mathbf{A_1} = \begin{bmatrix} -4 & -6 \\ 2 & -7 \end{bmatrix}, \quad \mathbf{A_2} = \begin{bmatrix} 2 & -1 \\ 0 & -1 \end{bmatrix}$$

2-13. Igazoljuk, hogy $\mu(\mathbf{0}) = 0$, $\mu(\mathbf{I}) = 1$ és $\mu(-\mathbf{I}) = -1$.

2-14. Igazoljuk a definícióval, hogy $\mu_1(\mathbf{A}) = \mu_\infty(\mathbf{A}^*)$.

2-15. Igazoljuk, hogy $c \geq 0$ esetén $\mu(c \cdot \mathbf{A}) = c \cdot \mu(\mathbf{A})$.

2-16. Igazoljuk, hogy $c < 0$ esetén $\mu(c \cdot \mathbf{A}) = |c| \cdot \mu(-\mathbf{A})$.

2-17. Igazoljuk, hogy $\mu(\mathbf{A} + \mathbf{B}) \leq \mu(\mathbf{A}) + \mu(\mathbf{B})$.

Bizonyítás. A számlálót és nevezőt 2-vel szorozva és a normákra vonatkozó háromszögegyenlőtlenséget felhasználva bármely $h > 0$-ra teljesül a következő egyenlőtlenség:

$$\frac{\|\mathbf{I} + h(\mathbf{A} + \mathbf{B})\| - 1}{h} = \frac{\|\mathbf{I} + h\mathbf{A} + h\mathbf{B}\| - 1}{h} = \frac{\|\mathbf{I} + 2h\mathbf{A} + \mathbf{I} + 2h\mathbf{B}\| - 2}{2h} \leq$$
$$\leq \frac{\|\mathbf{I} + 2h\mathbf{A}\| - 1}{2h} + \frac{\|\mathbf{I} + 2h\mathbf{B}\| - 1}{2h}.$$

Mindkét oldalon $h \to 0+$ határértéket véve

$$\mu(\mathbf{A} + \mathbf{B}) \leq \lim_{h \to 0+} \frac{\|\mathbf{I} + 2h\mathbf{A}\| - 1}{2h} + \lim_{h \to 0+} \frac{\|\mathbf{I} + 2h\mathbf{B}\| - 1}{2h} = \mu(\mathbf{A}) + \mu(\mathbf{B}).$$

■

2-18. Igazoljuk, hogy $|\mu(\mathbf{A}) - \mu(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\|$.

2-19. Igazoljuk, hogy $z \in \mathbb{C}$ esetén $\mu(\mathbf{A} + z \cdot \mathbf{I}) = \mu(\mathbf{A}) + \operatorname{Re} z$.

2-20. Igazoljuk, hogy ha $\mathbf{D} = \operatorname{diag}(d_{ii})$ és $p \in \mathbb{N}$ vagy $p = \infty$, akkor a p normában felírt logaritmikus normában $\mu_p(\mathbf{D}) = \max d_{ii}$ és $-\mu_p(-\mathbf{D}) = \min d_{ii}$.

2-21. Igazoljuk, hogy ha $\mathbf{A}$ szimmetrikus mátrix, akkor $\mu_2(\mathbf{A}) = \lambda_{\max}(\mathbf{A})$.

2-22. Igazoljuk, hogy $z \in \mathbb{C}$ esetén $\mu(z \cdot \mathbf{I}) = \operatorname{Re}(z)$.

Bizonyítás. Legyen $z = z_1 + z_2 i \in \mathbb{C}$, ekkor

$$\|\mathbf{I} + hz\mathbf{I}\| = \left\| \begin{bmatrix} 1+hz & 0 & 0 \\ & \dots & \\ 0 & 0 & 1+hz \end{bmatrix} \right\| = |1+hz| = \sqrt{(1+hz_1)^2 + (hz_2)^2}.$$

A határérték kiszámításához a következő függvény deriváltját használjuk.

$$f(h) := \sqrt{(1+hz_1)^2 + (hz_2)^2}, \quad f(0) = 1, \quad f'(h) = \frac{2(1+hz_1)z_1 + 2hz_2^2}{2\sqrt{(1+hz_1)^2 + (hz_2)^2}}$$

$$\lim_{h\to 0+} \frac{\sqrt{(1+hz_1)^2 + (hz_2)^2} - 1}{h} = \lim_{h\to 0+} \frac{f(h) - f(0)}{h} = f'(0) = \frac{2z_1}{2} = z_1 = \mathrm{Re}\,(z)$$

■

2-23. Igazoljuk, hogy skaláris szorzattal definiált normában felírt logaritmikus norma esetén bármely $\mathbf{y} \neq \mathbf{0}$ vektorra

$$-\mu(-\mathbf{A}) \leq \frac{\mathbf{y^T A y}}{\mathbf{y^T y}} \leq \mu(\mathbf{A}).$$

2-24. Igazoljuk, hogy az $\mathbf{A}$ mátrixra felírt Gersgorin-körökre

$$-\mu_\infty(-\mathbf{A}) = \min\left\{(\cup_{i=1}^n G_i) \cap \mathbb{R}\right\}, \qquad \max\left\{(\cup_{i=1}^n G_i) \cap \mathbb{R}\right\} = \mu_\infty(\mathbf{A}),$$

és az $\mathbf{A^T}$ mátrixra felírt Gersgorin-körökre

$$-\mu_1(-\mathbf{A}) = \min\left\{\left(\cup_{i=1}^n \widetilde{G}_i\right) \cap \mathbb{R}\right\}, \qquad \max\left\{\left(\cup_{i=1}^n \widetilde{G}_i\right) \cap \mathbb{R}\right\} = \mu_1(\mathbf{A}).$$

A kétféle típusú Gersgorin-körök metszetében vannak az $\mathbf{A}$ sajátértékei.

Bizonyítás. Az i-edik Gersgorin-kör sugara $r_i = \sum_{j\neq i} |a_{ij}|$, ezért a logaritmikus norma alakja

$$\mu_\infty(\mathbf{A}) = \max_i (a_{ii} + \sum_{j\neq i} |a_{ij}|) = \max_i (a_{ii} + r_i).$$

Ez azt jelenti, hogy a ∞-normában a logaritmikus norma a Gersgorin-körök úniójának valós tengellyel vett metszetének legfelső pontja. Nézzük az alsó becslést:

$$-\mu_\infty(-\mathbf{A}) = -\max_i (-a_{ii} + \sum_{j\neq i} |-a_{ij}|) = -\max_i (-a_{ii} + r_i) = \min_i (a_{ii} - r_i).$$

Ez az eredmény azt jelenti, hogy az ∞-normában a logaritmikus norma a Gersgorin-körök úniójának valós tengellyel vett metszetének legalsó pontja.

Az 1-es normában felírt állítást ugyanígy bizonyítjuk, de ott $\mathbf{A^T}$-ra írjuk fel a Gersgorin-köröket.
Mivel $\mathbf{A}$ és $\mathbf{A^T}$ sajátértékei azonosak és $(\cup_{i=1}^n G_i) \cap \left(\cup_{i=1}^n \widetilde{G}_i\right)$ tartalmazza az összes sajátértéket, ezért ezzel a sajátértékek és a logaritmikus norma (az 1-es illetve ∞-normában) kapcsolatára is kaptunk egy állítást. ■

2-25. Igazoljuk, hogy az $\mathbf{A}$ bármely λ_i sajátértékére

$$-\mu(-\mathbf{A}) \leq \mathrm{Re}\,(\lambda_i) \leq \mu(\mathbf{A}).$$

(Az előző feladatban a Gersgorin-körökre vonatkozó állítás éppen az 1-es és ∞-normában bizonyítja az állítást. Általánosan nehéz bizonyítani.)

2-26. Igazoljuk, hogy ha $\mu(\mathbf{A}) < 0$, akkor

$$\|\mathbf{A^{-1}}\| \leq -\frac{1}{\mu(\mathbf{A})}.$$

2-27. Jelölje $\mathbf{A_{sym}} = \frac{1}{2}(\mathbf{A} + \mathbf{A^T})$ az $\mathbf{A}$ szimmetrikus részét. Igazoljuk, hogy

$$\mu_2(\mathbf{A}) \leq c \quad \Leftrightarrow \quad c\mathbf{I} - \mathbf{A_{sym}} \text{ pozitív szemidefinit.}$$

2-28. Jelölje $\alpha(\mathbf{A}) = \max_i \operatorname{Re} \lambda_i$ az $\mathbf{A}$ sajátértékeinek maximális valós részét (spectral abscissa). Igazoljuk, hogy

$$\alpha(\mathbf{A}) \leq \mu(\mathbf{A}).$$

2.5. Állandó együtthatós lineáris differenciaegyenletek

A legismertebb differenciaegyenlet a Fibonacci-sorozatra felírt

$$y_{n+1} = y_n + y_{n-1}, \quad y_0 = y_1 = 1$$

egyenlet a megadott kezdeti feltételekkel. Ehhez hasonló differenciaegyenleteket kapunk többlépéses módszerek eredményeképpen, ezért fontos külön foglalkoznunk az y_n n. tag felírásával és a differenciaegyenlet stabilitásával. Az egyenlet megoldásának vizsgálata nagyon hasonló az állandó együtthatós lineáris differenciálegyenletek megoldásához (lásd [1] 46. oldalán), így annak ismerete segít a témakör megértésében.

2-8. Definíció. Legyen $l \in \mathbb{N}$ és $\alpha_0, \ldots \alpha_l \in \mathbb{R}$, $\alpha_0, \alpha_l \neq 0$, $f_n \in \mathbb{R}, (n \in \mathbb{N}_0)$, ekkor a

$$\sum_{k=0}^{l} \alpha_k y_{n+k} = f_n \quad n \in \mathbb{N}_0$$

egyenletet l-edrendű lineáris differenciaegyenletnek nevezzük. Ha α_k független n-től, akkor állandó együtthatós (csak ezekkel foglalkozunk), ha $f_n \equiv 0$ minden n-re, akkor homogén egyenletnek nevezzük.

A differenciaegyenlet megoldásán az $\mathbf{y} = (y_n | n = 0, 1, \ldots)$ végtelen sorozatot értjük, melyre

$$\sum_{k=0}^{l} \alpha_k y_{n+k} = f_n \quad n = l, l+1, \ldots \ .$$

Ha az $y_0, \ldots, y_{l-1}$ kezdeti feltételeket megadjuk, akkor a differenciaegyenlet megoldása egyértelmű lesz. A

$$\varrho(z) = \sum_{k=0}^{l} \alpha_k z^k$$

l-edfokú polinomot a differenciaegyenlet karakterisztikus polinomjának (vagy karakterisztikus egyenletének) nevezzük.

2-20. T *A homogén differenciaegyenlet megoldásai lineáris alteret képeznek, vagyis ha* $\mathbf{y^{(1)}} = (y_n^{(1)})$ *és* $\mathbf{y^{(2)}} = (y_n^{(2)})$ *ugyanazon homogén differenciaegyenlet két megoldása, akkor tetszőleges* $c_1, c_2 \in \mathbb{R}$ *konstansokra* $c_1\mathbf{y^{(1)}} + c_2\mathbf{y^{(2)}}$ *is megoldása a homogén egyenletnek.*

Bizonyítás. Lásd Feladatok. ■

2-21. T *Az inhomogén differenciaegyenlet teteszőleges megoldása előállítható ugyanazon inhomogén egyenlet egy partikuláris megoldásának és a homogén egyenlet általános megoldásának összegeként.*

a) *Ha* $\mathbf{v} = (v_n)$ *az inhomogén egyenlet partikuláris megoldása és* $\mathbf{y} = (y_n)$ *a homogén egyenlet általános megoldása, akkor* $\mathbf{v} + \mathbf{y}$ *az inhomogén egyenlet megoldása.*

b) *Ha* $\mathbf{v^{(1)}} = (v_n^{(1)})$ *és* $\mathbf{v^{(2)}} = (v_n^{(2)})$ *ugyanazon inhomogén differenciaegyenlet két megoldása, akkor* $\mathbf{v^{(1)}} - \mathbf{v^{(2)}}$ *a homogén egyenlet megoldása.*

Bizonyítás. Lásd Feladatok. ■

2-22. T *A homogén differenciaegyenlet* $\mathbf{y^{(1)}}, \mathbf{y^{(2)}}, \ldots, \mathbf{y^{(m)}}$ *megoldásai akkor és csak akkor lineárisan függetlenek, ha az* $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_m} \in \mathbb{R}^l$ *vektorok lineárisan függetlenek, ahol* $\mathbf{y_j} = (y_0^{(j)}, y_1^{(j)}, \ldots, y_{l-1}^{(j)})^T$.

Bizonyítás. Lásd a [4] jegyzet 43. oldalán. ■

2-23. T *Ha* $\mathbf{y^{(1)}}, \ldots, \mathbf{y^{(m)}}$, $(m \leq l)$ *lineárisan független megoldásai a homogén differenciaegyenletnek, akkor a homogén egyenlet tetszőleges* $\mathbf{y}$ *megoldása felírható a következő alakban:*

$$\mathbf{y} = \sum_{k=1}^{m} c_k \mathbf{y^{(k)}}, \quad c_1, \ldots, c_m.$$

Bizonyítás. Lásd Feladatok. ■

A továbbiakban a homogén differeciaegyenlet lineárisan független megoldásainak egy maximális elemszámú rendszerét szeretnénk előállítani. Ebben fontos szerepet játszanak a karakterisztikus polinom gyökei.

2-24. T *Ha* μ *a* ϱ *karakterisztikus polinom gyöke, akkor* $\mathbf{y} = (y_n) = (\mu^n)$ *a homogén egyenlet megoldása.*

Bizonyítás. Helyettesítsük be $\mathbf{y} = (y_n) = (\mu^n)$-t a homogén egyenletbe.

$$\sum_{k=0}^{l} \alpha_k y_{n+k} = \sum_{k=0}^{l} \alpha_k \mu^{n+k} = \mu^n \sum_{k=0}^{l} \alpha_k \mu^k = \mu^n \varrho(\mu) = 0.$$

■

2-25. T *Ha* $\mu_1, \ldots, \mu_l$ *a* ϱ *karakterisztikus polinom gyökei, akkor*

$$\mathbf{y} = (y_n), \quad y_n = \sum_{j=1}^{l} c_j \mu_j^n$$

a homogén egyenlet megoldása. Ha a gyökök különbözők, akkor a (μ_j^n) *megoldások lineárisan függetlenek.*

Bizonyítás. Helyettesítsük be $\mathbf{y} = (y_n) = (\sum_{j=1}^{l} c_j \mu_j^n)$-t a homogén egyenletbe:

$$\sum_{k=0}^{l} \alpha_k \sum_{j=1}^{l} c_j \mu_j^{n+k} = \sum_{j=1}^{l} c_j \mu_j^n \sum_{k=0}^{l} \alpha_k \mu_j^k = \sum_{j=1}^{l} c_j \mu_j^n \varrho(\mu_j) = 0.$$

A függetlenség bizonyításához vegyük a megoldások lineáris kombinációját, mely 0-t ad minden n-re. Belátjuk, hogy ez csak úgy teljesülhet, ha minden $c_j = 0$.

$$\sum_{j=1}^{l} c_j \mu_j^n = 0 \quad n = 0, 1, \ldots$$

Azonos kezdeti feltételek esetén $c_1, \ldots, c_l$-re egy lineáris egyenletrendszert kapunk $n = 0, 1, \ldots, l-1$-re:

$$\begin{bmatrix} 1 & 1 & \ldots & 1 \\ \mu_1 & \mu_2 & \ldots & \mu_l \\ \vdots & \vdots & & \vdots \\ \mu_1^{l-1} & \mu_2^{l-1} & \ldots & \mu_l^{l-1} \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_l \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

A kapott mátrix a Vandermonde-mátrix transzponáltja, így különböző gyökök esetén a LER-nek egyetlen megoldása $c_1 = c_2 = \ldots = c_l = 0$. Ezzel a függetlenséget beláttuk. ■

2-26. T *Ha μ a ϱ karakterisztikus polinom gyöke m multiplicitással, akkor*

$$(\mu^n),\ (n\mu^{n-1}),\ (n(n-1)\mu^{n-2}),\ \ldots,\ \left(\frac{n!}{(n-m+1)!}\mu^{n-m+1}\right)$$

a homogén egyenlet megoldásai és lineárisan függetlenek.

Bizonyítás. Korábban már beláttuk, hogy (μ^n) megoldása a homogén egyenletnek.
Nézzük $(n\mu^{n-1})$-t és használjuk fel, hogy $\varrho(\mu) = 0$ és $\varrho'(\mu) = 0$:

$$\sum_{k=0}^{l} \alpha_k (n+k)\mu^{n+k-1} = n\mu^{n-1} \sum_{k=0}^{l} \alpha_k \mu^k + \mu^n \sum_{k=0}^{l} \alpha_k k \mu^{k-1} = n\mu^{n-1}\varrho(\mu) + \mu^n \varrho'(\mu) = 0.$$

Megjegyezzük, hogy

$$\left(\sum_{k=0}^{l} \alpha_k \mu^{n+k}\right)' = (\mu^n \varrho(\mu))' = n\mu^{n-1}\varrho(\mu) + \mu^n \varrho'(\mu),$$

vagyis éppen a differenciaegyenlet μ szerinti deriváltját kaptuk.
Nézzük a következő függvényt, $(n(n-1)\mu^{n-2})$-t:

$$\begin{aligned} &\sum_{k=0}^{l} \alpha_k (n+k)(n+k-1)\mu^{n+k-2} = \sum_{k=0}^{l} \alpha_k [n(n-1) + 2nk + k(k-1)]\mu^{n+k-2} = \\ &= n(n-1)\mu^{n-2} \sum_{k=0}^{l} \alpha_k \mu^k + 2n\mu^{n-1} \sum_{k=0}^{l} \alpha_k k \mu^{k-1} + \mu^n \sum_{k=0}^{l} \alpha_k k(k-1)\mu^{k-2} = \\ &= n(n-1)\mu^{n-2}\varrho(\mu) + 2n\mu^{n-1}\varrho'(\mu) + \mu^n \varrho''(\mu) = 0 = (\mu^n \varrho(\mu))''. \end{aligned}$$

Az általános eset a differenciaegyenlet μ szerinti j. deriváltja alapján $(j = 1, \ldots, m-1)$-re a szorzat deriválási szabályának felhasználásával kapható meg.

$$(\mu^n \varrho(\mu))^{(j)} = \mu^n \varrho(\mu)^{(j)} + \binom{j}{n} n\mu^{n-1} \varrho(\mu)^{(j-1)} + \ldots + \binom{j}{j} (\mu^n)^{(j)} \varrho(\mu) = 0,$$

mivel μ gyöke a ϱ deriváltjainak is. Ebből végigszámolható az általános képlet esete. Az m. megoldásfüggvényre

$$\sum_{k=0}^{l} \alpha_k \underbrace{(n+k)(n+k-1)\dots(n+k-m+2)\mu^{n+k-m+1}}_{=y_{n+k}} = (\mu^n \varrho(\mu))^{(m-1)} = 0.$$

A függetlenség bizonyításához vegyük a megoldások lineáris kombinációját, mely 0-t ad minden n-re. Belátjuk, hogy ez csak úgy teljesülhet, ha minden $c_j = 0$.

$$\sum_{j=1}^{m} c_j \frac{n!}{(n-m+j)!}\mu^{n-m+j} = 0 \quad n = 0, 1, \dots$$

Azonos kezdeti feltételek esetén $c_1, \dots, c_m$-re egy lineáris egyenletrendszert kapunk $n = 0, 1, \dots, m-1$-re:

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \mu & 1 & 0 & \dots & 0 \\ \mu^2 & 2\mu & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \mu^{m-1} & (m-1)\mu^{m-2} & (m-1)(m-2)\mu^{m-3} & \dots & \frac{m!}{(m-n)!} \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

A kapott alsóháromszögű LER-nek egyetlen megoldása $c_1 = c_2 = \dots = c_m = 0$. Ezzel a függetlenséget beláttuk. (A fenti mátrix általánosított Vandermonde-mátrix transzponáltja, az Hermite-interpolációnál találkozunk még vele.) ■

Megjegyzések.

1. m-szeres gyök esetén a tételben megadott megoldások helyett szokás a

$$(\mu^n),\ (n\mu^n),\ (n^2\mu^n), \dots,\ \left(n^{m-1}\mu^n\right)$$

rendszert is használni. A rendszer bármely eleme előállítható a tételbeli elemekkel és ez fordítva is igaz.

2. A tételt a következő alakban is meg szokták fogalmazni:
Ha μ a karakterisztikus polinom m-szeres gyöke és $P(z)$ egy legfeljebb $m-1$-edfokú polinom, akkor a P polinomból előállított $\mathbf{p} = (p_n) = (P(n)\mu^n)$ megoldása a homogén egyenletnek.

2-27. T *Tegyük fel, hogy $\mu_1, \dots, \mu_m$ egymástól és 0-tól különböző komplex számok. A $P_1(z), \dots, P_m(z)$ valós együtthatós polinomokkal felírt*

$$\sum_{k=1}^{m} P_k(n)\mu_k^n = 0, \quad n = 0, 1, \dots$$

azonosság pontosan akkor állhat fenn, ha mindegyik $P_k(z)$ polinom azonosan nulla.

Bizonyítás. Lásd a [4] jegyzet 46. oldalán. ■

2-28. T *A homogén differenciaegyenlet tetszőleges* $\mathbf{y} = (y_n)$ *megoldása előállítható a*

$$\mathbf{y} = \sum_{k=1}^{m} \sum_{j=0}^{m_k-1} c_{kj} n^j \mu_k^n, \quad n = 0, 1, \ldots$$

alakban, ahol $\mu_1, \ldots, \mu_m$ *a karakterisztikus polinomnak az* $m_1, \ldots, m_k$ *multiplicitású különböző gyökei, így* $\sum_{k=1}^{m} m_k = l$.

Bizonyítás. Az előző két tételből következik.

■

2-12. Példa. Írjuk fel az $y_{n+1} = y_n + y_{n-1}$, $y_0 = y_1 = 1$ Fibonacci-sorozat n. tagját és vizsgáljuk a sorozat stabilitását.

Megoldás. A differenciaegyenlet $y_{n+1} - y_n - y_{n-1} = 0$ átrendezéséből a karakterisztikus polinom: $\varrho(z) = z^2 - z - 1$. Ennek gyökei: $\mu_1 = \frac{1+\sqrt{5}}{2}$ és $\mu_2 = \frac{1-\sqrt{5}}{2}$. Az általános megoldás alakja:

$$y_n = c_1 \left(\frac{1+\sqrt{5}}{2}\right)^n + c_2 \left(\frac{1-\sqrt{5}}{2}\right)^n,$$

ahol $c_1, c_2 \in \mathbb{R}$. Ezeket a konstansokat a kezdeti feltételből határozzuk meg:

$$y_0 = c_1 + c_2 = 1$$
$$y_1 = c_1 \left(\frac{1+\sqrt{5}}{2}\right) + c_2 \left(\frac{1-\sqrt{5}}{2}\right) = 1$$

A 2. egyenletet átírva

$$1 = \frac{1}{2}(c_1 + c_2) + \frac{\sqrt{5}}{2}(c_1 - c_2) = \frac{1}{2} + \frac{\sqrt{5}}{2}(c_1 - c_2) \quad \rightarrow \quad \frac{1}{\sqrt{5}} = c_1 - c_2$$
$$1 = c_1 + c_2$$

Innen

$$c_1 = \frac{1}{2} + \frac{1}{2\sqrt{5}} = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right), \quad c_2 = \frac{1}{2} - \frac{1}{2\sqrt{5}} = -\frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right).$$

Tehát a Fibonacci-sorozat n. tagjának képlete:

$$y_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}.$$

Vizsgáljuk most a sorozat érzékenységét a kezdeti feltételek változására.
Ha y_0 helyett $y_0 - \varepsilon_0$-ból illetve y_1 helyett $y_1 - \varepsilon_1$-ből indulnánk,
akkor a δ_n hibasorozatra

$$\delta_{n+1} = y_{n+1} - \widetilde{y}_{n+1} = (y_n + y_{n-1}) - (\widetilde{y}_n + \widetilde{y}_{n-1}) =$$
$$= (y_n - \widetilde{y}_n) + (y_{n-1} - \widetilde{y}_{n-1}) = \delta_n + \delta_{n-1}.$$

Hozzávéve a kezdeti feltételeket

$$\delta_{n+1} = \delta_n + \delta_{n-1}, \quad \delta_0 = \varepsilon_0, \ \ \delta_1 = \varepsilon_1,$$

vagyis ugyanolyan differenciaegyenletet kapunk, melynek megoldása

$$\delta_n = d_1 \left(\frac{1+\sqrt{5}}{2}\right)^n + d_2 \left(\frac{1-\sqrt{5}}{2}\right)^n$$

alakú, ahol d_1 és d_2 értékét a kezdeti feltételekből egy 2×2-es LER megoldásával kell meghatározni. Nézzük meg, hogy a kezdeti feltételekben bekövetkező változás hogyan befolyásolja a megoldása változását. Mivel

$$\left|\frac{1-\sqrt{5}}{2}\right| < 1 \quad \Rightarrow \quad \left(\frac{1-\sqrt{5}}{2}\right)^n \to 0 \ \ (n \to \infty)$$
$$\frac{1+\sqrt{5}}{2} > 1 \quad \Rightarrow \quad \left(\frac{1+\sqrt{5}}{2}\right)^n \to \infty \ \ (n \to \infty),$$

így ha $d_1 \neq 0$, akkor

$$\lim_{n\to\infty} \delta_n = +\infty,$$

tehát az eltérés bármilyen nagy lehet. (A stabilitáshoz bármely kezdeti feltétel esetén teljesülnie kell, hogy az eltérés korlátos marad.) Látjuk, hogy a problémát az 1-nél nagyobb abszolútértékű gyök jelenti. ■

Feladatok

2-29. Igazoljuk, hogy ha $\mathbf{y^{(1)}} = (y_i^{(1)})$ és $\mathbf{y^{(2)}} = (y_i^{(2)})$ ugyanazon homogén differenciaegyenlet két megoldása, akkor tetszőleges $c_1, c_2 \in \mathbb{R}$ esetén $c_1\mathbf{y^{(1)}} + c_2\mathbf{y^{(2)}}$ is megoldása a homogén egyenletnek.

2-30. Ha $\mathbf{v^{(1)}} = (v_n^{(1)})$ és $\mathbf{v^{(2)}} = (v_n^{(2)})$ ugyanazon inhomogén differenciaegyenlet két megoldása, akkor $\mathbf{v^{(1)}} - \mathbf{v^{(2)}}$ a homogén egyenlet megoldása.

2-31. Igazoljuk, hogy ha $\mathbf{y^{(1)}}, \ldots, \mathbf{y^{(m)}}$ a homogén differenciaegyenlet megoldásai, akkor tetszőleges $c_1, \ldots, c_m \in \mathbb{R}$ esetén

$$\mathbf{y} = \sum_{k=1}^{m} c_k \mathbf{y^{(k)}}$$

is megoldása a homogén egyenletnek.

2-32. Igazoljuk, hogy ha $\mathbf{y^{(1)}}, \ldots, \mathbf{y^{(m)}}$ a homogén differenciaegyenlet lineárisan független megoldásai, akkor a homogén egyenlet $\mathbf{y}$ általános megoldása felírható a következő alakban:

$$\mathbf{y} = \sum_{k=1}^{m} c_k \mathbf{y^{(k)}}.$$

Az $y_0, \ldots, y_{l-1}$ kezdeti feltételek megadása esetén igazoljuk az egyértelműséget.

2-33. Írjuk fel a $y_{n+1} = y_n - \frac{1}{4}y_{n-1}$ differenciaegyenlet általános megoldását és vizsgáljuk a stabilitását a $\delta_0 = \varepsilon$ és $\delta_1 = 0$ esetben!

2-34. Írjuk fel a $y_{n+1} = 4y_n - 4y_{n-1}$ differenciaegyenlet általános megoldását és vizsgáljuk a stabilitását a $\delta_0 = \varepsilon$ és $\delta_1 = 0$ esetben!

2-35. Írjuk fel a $y_{n+1} = \frac{9}{2}y_n - 7y_{n-1} - \frac{9}{2}y_{n-2} - y_{n-3}$ differenciaegyenlet általános megoldását és vizsgáljuk a stabilitását! (A karakterisztikus polinom gyökei: $2, \frac{1}{2}, 1, 1$.)

2.6. Differenciálegyenletek alapvető tulajdonságai

2-9. Definíció. Legyen $f(x, \mathbf{y})$ az $x \in \mathbb{R}$ és $\mathbf{y} \in \mathbb{R}^n$ argumentumának folytonos vektorfüggvénye, $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$. A közönséges differenciálegyenletekből álló rendszer általános alakja

$$\mathbf{y}'(x) = f(x, \mathbf{y}(x)), \quad \mathbf{y}(x_0) = \mathbf{y_0}.$$

Az $\mathbf{y}(x_0) = \mathbf{y_0}$ feltételt kezdeti feltételnek nevezzük. A differenciálegyenlet rendszer megoldását egy véges intervallumon keressük. A differenciálegyenlet rendszer megoldása az $\mathbf{y} : [a; b] \to \mathbb{R}^n$ függvény, folytonosan differenciálható és kielégíti a fenti egyenletet és kezdeti feltételt az $x_0 \in [a; b]$ pontban. A derivált jelölésére a $\frac{d\mathbf{y}(x)}{dx}$ vagy $\frac{d}{dx}\mathbf{y}(x)$, az egyenletre az $\mathbf{y}' = f(x, \mathbf{y})$ jelölést is használjuk.

2-10. Definíció. A magasabbrendű differenciálegyenlet kezdetiérték feladatának általános alakja:

$$y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}),$$
$$y(x_0) = y_0^{(0)}, \;\; y'(x_0) = y_0^{(1)}, \ldots, y^{(n-1)}(x_0) = y_0^{(n-1)},$$

ahol $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$ adott függvény folytonos az $(x_0, y_0^{(0)}, \ldots, y_0^{(n-1)})$ pont környezetében. A differenciálegyenlet megoldása az $y \in \mathbb{R} \to \mathbb{R}$ függvény, mely az x_0 pont környezetében értelmezett, n-szer folytonosan differenciálható és kielégíti a fenti egyenletet és kezdeti feltételeket.

2-29. T *A fenti magasabbrendű differenciálegyenlet kezdetiérték feladatának megoldása azonos a következő elsőrendű diff. egy. rendszer kezdetiérték feladatának megoldásával:*

$$y_1' = y_2,$$
$$y_2' = y_3,$$
$$\ldots$$
$$y_{n-1}' = y_n,$$
$$y_n' = f(x, y_1, y_2, \ldots, y_n),$$
$$y_1(x_0) = y_0^{(0)}, \;\; y_2(x_0) = y_0^{(1)}, \ldots, y_n(x_0) = y_0^{(n-1)}.$$

Bizonyítás. Nyilvánvaló. ■

Mivel bármely véges zárt intervallum egy lineáris transzformációval a $[0; 1]$-re transzformálható, a továbbiakban olyan diff. egy. rendszerekkel foglalkozunk, melynek megoldása az $\mathbf{y} : [0; 1] \to \mathbb{R}^n$ függvény. $\mathbf{y} \in (C^1[0; 1])^n$ ($[0; 1]$-en folytonosan differenciálható) és kielégíti a következő egyenletet és kezdeti feltételt ($x_0 = 0$).

$$\mathbf{y}'(x) = f(x, \mathbf{y}(x)), \quad x \in [0; 1],$$
$$\mathbf{y}(0) = \mathbf{y_0}$$

2-11. Definíció. Az $f(x, \mathbf{y})$ függvény a második változójában eleget tesz a Lipschitz-feltételnek $\mathbb{R}^n$-en, ha létezik olyan $L_f \geq 0$ állandó, melyre

$$\|f(x, \mathbf{u}) - f(x, \mathbf{v})\| \leq L_f \|\mathbf{u} - \mathbf{v}\|, \quad x \in [0; 1], \quad \mathbf{u}, \mathbf{v} \in \mathbb{R}^n.$$

Ekkor f-et Lipschitz-folytonosnak nevezzük és $f \in Lip(\mathbf{y})$-nal jelöljük.

Megjegyzések.

1. Az analízisben $D := [x_0 - a; x_0 + a] \times \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{y_0}\| \leq b\} \subset \mathbb{R} \times \mathbb{R}^n$ hengerre szokás definiálni a Lipschitz-feltételt, ahol $(x, \mathbf{u}), (x, \mathbf{v}) \in D$ és $y(x_0) = \mathbf{y_0}$ a kezdeti feltétel.

2. Ha $f(x, \mathbf{y})$ folytonos, akkor a Cauchy–Peano-féle egzisztencia tétel szerint mindig van a kezdetiérték problémának megoldása.

3. Ha $f(x, \mathbf{y})$ folytonos és második változójában eleget tesz a Lipschitz-feltételnek, akkor a Picard–Lindelöf tétel szerint a kezdetiérték-feladat lokálisan egyértelműen megoldható.

A pontosítást lásd Differenciálegyenletek tárgyból.

2-13. Példa. Az $f(x, y) = cy(1 - y)$, ($c > 0$ konstans) folytonos függvény a második változójában Lipschitz-folytonos $[0; 1]$-en, így $0 < y_0 \leq 1$ esetén az

$$\begin{aligned} y' &= cy(1 - y), \\ y(0) &= y_0 \end{aligned}$$

kezdetiérték feladatnak van egyértelmű megoldása. A konkrét példa a hírek terjedésére a [11] jegyzetben található.

Megoldás. Tetszőleges $x \in [0; 1]$ és $y_1 \in (0; 1)$, $y_2 \in [0; 1]$ esetén

$$\begin{aligned} |f(x, y_1) - f(x, y_2)| = c|y_1(1 - y_1) - y_2(1 - y_2)| = c|y_1 - y_2 - (y_1^2 - y_2^2)| = \\ = c|y_1 - y_2| \cdot |1 - (y_1 + y_2)| \leq c|y_1 - y_2|. \end{aligned}$$

Tehát f Lipschitz-folytonos ($L = c$), a tanult tételek szerint a kezdetiérték feladatnak egyértelmű a megoldása.

- Ha $y_0 = 0$, akkor $y(x) \equiv 0$.
- Ha $y_0 = 1$, akkor $y(x) \equiv 1$.
- Ha $0 < y_0 < 1$, akkor $y(x) = \frac{1}{1+de^{-cx}}$, ahol $d = \frac{1-y_0}{y_0}$.

Ellenőrizzük az utóbbit!

$$\begin{aligned} y'(x) &= \frac{dce^{-cx}}{(1 + de^{-cx})^2} \\ y(x)(1 - y(x)) &= \frac{1}{1 + de^{-cx}} \cdot \left(1 - \frac{1}{1 + de^{-cx}}\right) = \frac{dce^{-cx}}{(1 + de^{-cx})^2} \\ y(0) &= \frac{1}{1 + d} = y_0 \quad \Leftrightarrow \quad d = \frac{1 - y_0}{y_0}. \end{aligned}$$

■

2-14. Példa. Az $f(x, y) = \sqrt{y}$ függvény a második változójában nem Lipschitz-folytonos $[0; 1]$-en, az

$$\begin{aligned} y' &= \sqrt{y}, \\ y(0) &= 0 \end{aligned}$$

kezdetiérték feladatnak nincs egyértelmű megoldása.

Megoldás. Tetszőleges $x \in [0;1]$ és $y_1 > 0$, $y_2 \geq 0$ esetén

$$|f(x,y_1) - f(x,y_2)| = |\sqrt{y_1} - \sqrt{y_2}| = \frac{|y_1 - y_2|}{\sqrt{y_1} + \sqrt{y_2}} \leq \frac{|y_1 - y_2|}{\sqrt{y_1}}.$$

Mivel $\frac{1}{\sqrt{y_1}}$ nem korlátos, ezért f nem Lipschitz-folytonos. A példánk mutatja, hogy ha nem teljesül a tulajdonság, akkor az egyértelműség nem biztos, hogy teljesül.

A kezdetiérték feladatnak $y(0) = 0$ kezdeti feltétel esetén két megoldása van:

1) $y \equiv 0$, ugyanis $y' \equiv 0 \equiv 0$ és $y(0) = 0$.
2) $y(x) = \frac{1}{4}x^2$, mivel $y'(x) = \frac{1}{2}x = \sqrt{y(x)}$ és y(0)=0.

Ha $y_1 \geq y_0 > 0$ lenne, akkor $\frac{1}{\sqrt{y_1}} \leq \frac{1}{\sqrt{y_0}} =: L$ lenne a Lipschitz konstans. Tehát egy pozitív kezdeti feltétel megoldaná a problémát, de feladatunkban $y_0 = 0$ szerepel.

■

Feladatok

2-36. Az $f(x,y) = \sqrt{y^2 + 5}$ függvény a második változójában az egész $\mathbb{R}$-en Lipschitz-folytonos. ($L_f = 1$)

2-37. Lipschitz-folytonosak-e a következő függvények a második változójukban a $[0;1]$-en?

a) $f(x,y) = e^y$,
b) $f(x,y) = y^3$,
c) $f(x,y) = \sqrt[3]{y}$,
d) $f(x,y) = \frac{1}{1-y}$,

2-38. Igazoljuk, hogy ha $f(t,\mathbf{y}), g(t,\mathbf{y}) \in Lip(\mathbf{y})$ az L_f, L_g Lipschitz-konstanssal (ugyanazon a halmazon) és $a, b \in \mathbb{R}$, akkor

$$\varphi(t,\mathbf{y}) = a \cdot f(t,\mathbf{y}) + b \cdot g(t,\mathbf{y}) \in Lip(\mathbf{y})$$

az $L_\varphi = |a|L_f + |b|L_g$ Lipschitz-konstanssal.

2-39. Igazoljuk, hogy ha $f(t,\mathbf{y}), g(t,\mathbf{y}) \in Lip(\mathbf{y})$ az L_f, L_g Lipschitz-konstanssal (ugyanazon a halmazon), akkor

$$\varphi(t,\mathbf{y}) = f(t, g(t,\mathbf{y})) \in Lip(\mathbf{y})$$

az $L_\varphi = L_f \cdot L_g$ Lipschitz-konstanssal.

2-40. Igazoljuk, hogy ha $f(t,\mathbf{y}), g(t,\mathbf{y}) \in Lip(\mathbf{y})$ az L_f, L_g Lipschitz-konstanssal (ugyanazon a halmazon) és $a, b \in \mathbb{R}$, akkor

$$\varphi(t,\mathbf{y}) = a \cdot f(t, b \cdot g(t,\mathbf{y})) \in Lip(\mathbf{y})$$

az $L_\varphi = |a||b| \cdot L_f \cdot L_g$ Lipschitz-konstanssal.

A numerikus megoldás előtt vizsgáljuk a hibák szemszögéből az egyenleteket. Milyenek azok az egyenletek, melyeknél az egyszer elkövetett hibáktól később nem kell tartani?

2-12. Definíció. Legyen $f : [0;1] \times \mathbb{R}^n \to \mathbb{R}^n$ folytonos és $\|.\|$ vektornorma $\mathbb{R}^n$-en. A differenciálegyenletet disszipatívnak (lecsengőnek) nevezzük az adott vektornormában, ha

$$\|\mathbf{u}(x) - \mathbf{v}(x)\| \leq \|\mathbf{u}(0) - \mathbf{v}(0)\|$$

minden tetszőleges (a kezdeti értékben különböző) $\mathbf{u}(x)$, $\mathbf{v}(x)$ megoldásra és minden $x \in [0;1]$-re. A megoldásokat ilyenkor kontraktívnak nevezzük.

2-30. T *(Skaláris egyenlet disszipativitása, $n = 1$ eset)*
Legyen $f(x, y)$ folytonos mindkét argumentumában és folytonosan differenciálható y szerint. Az

$$\begin{aligned} y'(x) &= f(x, y(x)), \quad x \in [0;1], \\ y(0) &= y_0 \end{aligned}$$

differenciálegyenlet disszipatív, ha

$$\frac{\partial f(x,y)}{\partial y} \leq 0, \quad \forall\, y, \forall\, x \in [0;1].$$

Bizonyítás. Tekintsünk két megoldást, u, v-t és vezessük be a $z(x) = u(x) - v(x)$ eltérést és a

$$\phi(x) = \begin{cases} \frac{f(x,u(x)) - f(x,v(x))}{u(x) - v(x)} & \text{ha } u(x) \neq v(x) \\ \frac{\partial f}{\partial u}\,(x, u(x)) & \text{ha } u(x) = v(x) \end{cases}$$

segédfüggvényt. Ekkor

$$z'(x) = u'(x) - v'(x) = f(x, u(x)) - f(x, v(x)) = \phi(x)(u(x) - v(x)) = (\phi z)(x).$$

A z-re kapott differenciálegyenlet megoldása

$$z(x) = z(0) \cdot \exp\left(\int_0^x \phi(s)\, ds\right).$$

A deriváltra vonatkozó feltétel miatt

$$\phi(s) \leq 0 \quad \Rightarrow \quad \exp\left(\int_0^x \phi(s)\, ds\right) \leq 1 \quad \Rightarrow \quad |z(x)| \leq |z(0)|.$$

Ezzel igazoltuk, hogy $|u(x) - v(x)| \leq |u(0) - v(0)|$. ■

Megjegyzés.

Ha éppen ellenkezően

$$0 \leq \frac{\partial f(x,y)}{\partial y} \leq M,$$

akkor az előző bizonyításból $|\int_0^x \phi(s)\, ds| \leq Mx$, így

$$|u(x) - v(x)| \leq e^{Mx}|u(0) - v(0)|$$

azaz a két megoldás eltérése exponenciálisan nőhet.

2-31. T *(A perturbált egyenlet megoldása)*
Tekintsük a következő u, v megoldásokat:

$$v' = f(x,v), \quad u' = f(x,u) + g(x,u),$$

ahol g folytonos függvény, $|g(x,y)| \le \varepsilon$ minden (x,y)-ra és

$$\frac{\partial f(x,y)}{\partial y} \le 0, \quad \forall\ y, \forall\ x \in [0;1].$$

Ekkor

$$|u(x) - v(x)| \le |u(0) - v(0)| + x\varepsilon.$$

Bizonyítás. Az előző bizonyításban leírtak szerint vezessük be a $z(x) = u(x) - v(x)$ eltérést és a

$$\phi(x) = \begin{cases} \frac{f(x,u(x)) - f(x,v(x))}{u(x)-v(x)} & \text{ha } u(x) \neq v(x) \\ \frac{\partial f}{\partial u}(x,u(x)) & \text{ha } u(x) = v(x) \end{cases}$$

segédfüggvényt. Ekkor

$$\begin{aligned} z'(x) &= u'(x) - v'(x) = f(x,u(x)) + g(x,u(x)) - f(x,v(x)) \\ &= \phi(x)(u(x) - v(x)) + g(x,u(x)) = (\phi z + g)(x). \end{aligned}$$

A z-re kapott differenciálegyenlet:

$$z' = \phi z + g.$$

Vezessük be a $\Phi(x) = \int_0^x \phi(s)\,ds$ jelölést és ellenőrizzük, hogy a megoldása

$$z(x) = z(0) \cdot \exp\left(\int_0^x \phi(s)\,ds\right) + \int_0^x \exp\left(\int_t^x \phi(s)\,ds\right) g(t,u(t))\,dt$$

Amit más alakban is felírhatunk.

$$\begin{aligned} z(x) &= z(0) \cdot \exp(\Phi(x) - \Phi(0)) + \int_0^x \exp(\Phi(x) - \Phi(t)) g(t,u(t))\,dt = \\ &= z(0) \cdot \exp(\Phi(x) - \Phi(0)) + \exp(\Phi(x)) \cdot \int_0^x \exp(-\Phi(t)) g(t,u(t))\,dt \end{aligned}$$

A kezdeti feltétel $z(0) = u(0) - v(0)$ rendben, deriváljuk a $z(x)$-et.

$$\begin{aligned} z'(x) = z(0) \cdot \underbrace{\Phi'(x)}_{\phi(x)} \cdot \exp(\Phi(x) - \Phi(0)) + \underbrace{\Phi'(x)}_{\phi(x)} \cdot \exp(\Phi(x)) \int_0^x \exp(-\Phi(t)) g(t,u(t))\,dt + \\ + \underbrace{\exp(\Phi(x)) \cdot \exp(-\Phi(x))}_{=1} \cdot g(x,u(x)) = \\ = \phi(x) \underbrace{\left(z(0) \cdot \exp(\Phi(x) - \Phi(0)) + \exp(\Phi(x)) \cdot \int_0^x \exp(-\Phi(t)) g(t,u(t))\,dt \right)}_{=z(x)} + g(x,u(x)) \end{aligned}$$

Így a $z(x)$ megoldást becsülve

$$\phi(s) \le 0 \quad \Rightarrow \quad \exp\left(\int_0^x \phi(s)\,ds\right) \le 1, \quad \exp\left(\int_t^x \phi(s)\,ds\right) \le 1, \quad \Rightarrow \quad |z(x)| \le |z(0)| + x\varepsilon.$$

$$|z(x)| = |u(x) - v(x)| \le |u(0) - v(0)| + x\varepsilon.$$

■

2-15. Példa. Mielőtt a rendszerekre vonatkozó disszipativitási feltételt megnézzük, egy fizikai példát vizsgálunk. Látni fogjuk, hogy az egyváltozós disszipativitási feltétel hogyan változik. A rugón felfüggesztett test differenciálegyenletét elemezzük (lásd [9] és [11]), nézzük a legegyszerűbb esetet.

Legyen m a tömeg, t az idő és k a rugó merevségére jellemző állandó, külső erő nem hat a testre, ekkor a normálhelyzettől való $x(t)$ eltérést az

$$mx''(t) + kx(t) = 0$$

differenciálegyenlet írja le. Ha külső erő hat a testre, akkor a jobboldalon megjelenik egy $f(t)$ függvény, ha a súrlódást is figyelembe akarjuk venni a rezgés során, akkor egy $rx'(t)$ tagot is hozzá kell vennünk. Így a differenciálegyenlet általános alakja a következő lesz:

$$mx''(t) + rx'(t) + kx'(t) = f(t).$$

Itt $x'(t)$ a test sebessége, $x''(t)$ a gyorsulása. Az m (tömeg) és k (rugóállandó) konstansokról feltesszük, hogy pozitívak. Szorozzuk az egyenletet x'-vel és rendezzük át:

$$-r(x'(t))^2 + x'(t)f(t) = mx''(t)x'(t) + kx(t) = \frac{d}{dt}\left[\frac{m}{2}(x'(t))^2 + \frac{k}{2}x^2(t)\right] = \frac{d}{dt}(K+P),$$

ahol K illetve P a kinetikus illetve helyzeti energia. Elemezzük, hogy mit kaptunk.

- Ha $f \equiv 0$ (nincs külső erő), $r > 0$ (van súrlódás) és $x' \not\equiv 0$, akkor az egyenlet baloldala negatív, így a $K + P$ összenergia idővel csökken, tehát a rendszer disszipatív.
- Ha $f \equiv 0$ (nincs külső erő) és $r = 0$ (nincs súrlódás), akkor az egyenlet baloldala 0, így az összenergia megmarad.

A továbbiakban a megoldások kontraktivitását vizsgáljuk. Legyen $x(t)$ és $u(t)$ az egyenlet két megoldása, azaz

$$\begin{aligned} -rx'(t) + f(t) &= mx''(t) + kx(t) \\ -ru'(t) + f(t) &= mu''(t) + ku(t). \end{aligned}$$

Vonjuk ki egymásból a két egyenletet

$$-r(x'(t) - u'(t)) = m(x''(t) - u''(t)) + k(x(t) - u(t)),$$

majd szorozzuk mindkét oldalt $x'(t) - u'(t)$-vel.

$$\begin{aligned} -r(x'(t) - u'(t))^2 &= m(x''(t) - u''(t))(x'(t) - u'(t)) + k(x(t) - u(t))(x'(t) - u'(t)) = \\ &= \frac{1}{2}\frac{d}{dt}\left[m(x'(t) - u'(t))^2 + k(x(t) - u(t))^2\right]. \end{aligned}$$

Ha $r \geq 0$, akkor a baloldal negatív, vagyis a derivált negatív, ami a jobboldali zárójeles kifejezés monoton csökkenését jelenti. $t_1 \leq t_2$ esetén

$$\left[m(x' - u')^2 + k(x - u)^2\right](t_1) \geq \left[m(x' - u')^2 + k(x - u)^2\right](t_2),$$

ebben az értelemben tehát a megoldások kontraktívak.

Nézzük most ugyanezt az eredményt az eredeti definíciónk alapján. Alakítsuk először a másodrendű egyenletet elsőrendű rendszerré a következő jelölésekkel:

$$y_1(t) := x(t), \quad y_2(t) := x'(t).$$

A másodrendű egyenletbe helyettesítve

$$y_1' = x' = y_2,$$
$$y_2' = x'' = \frac{1}{m}(-rx' - kx + f) = \frac{1}{m}(-ry_2 - ky_1 + f).$$

Ezt vektoralakban felírva

$$\mathbf{y} := \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x \\ x' \end{bmatrix}, \quad \mathbf{A} := -\frac{1}{m}\begin{bmatrix} 0 & -m \\ k & r \end{bmatrix}, \quad \mathbf{g} := \begin{bmatrix} 0 \\ \frac{f}{m} \end{bmatrix}$$

a következő differenciálegyenlet-rendszert kapjuk:

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g}.$$

Egy kis kitérőt teszünk a következőkben, megadjuk azt a vektornormát, amiben a disszipativitás definícióját felírjuk.

2-32. L *Legyen* $\mathbf{H}$ *szimmetrikus, pozitív definit mátrix, ekkor*

$$\|\mathbf{y}\|_H := (\mathbf{H}\mathbf{y}, \mathbf{y})^{1/2} \quad \textit{vektornorma} \quad \mathbb{R}^2\textit{-en.}$$

Feladatok

2-41. Igazoljuk az előző lemmát. Vegyük észre, hogy a $\mathbf{H}$-ra tett feltételek fontosak.

2-33. L *Legyen* $\|\mathbf{y}\|_H := (\mathbf{H}\mathbf{y}, \mathbf{y})^{1/2}$ *vektornorma* $\mathbb{R}^2$*-en, ahol* $\mathbf{H}$ *szimmetrikus, pozitív definit mátrix. Ekkor*

$$\frac{d}{dt}\|\mathbf{y}\|_H^2 = (\mathbf{H}\mathbf{y}', \mathbf{y}) + (\mathbf{H}\mathbf{y}, \mathbf{y}'), \quad \textit{ahol} \quad \mathbf{y} = \begin{bmatrix} x \\ x' \end{bmatrix}.$$

Bizonyítás. A norma definíciója alapján

$$\|\mathbf{y}\|_H^2 = (\mathbf{H}\mathbf{y}, \mathbf{y}) = \left(\begin{bmatrix} h_{11}x + h_{12}x' \\ h_{21}x + h_{22}x' \end{bmatrix}, \begin{bmatrix} x \\ x' \end{bmatrix} \right) = h_{11}x^2 + h_{12}x'x + h_{21}xx' + h_{22}(x')^2.$$

Deriválva t szerint

$$\frac{d}{dt}\|\mathbf{y}\|_H^2 = 2h_{11}x'x + h_{12}(x''x + (x')^2) + h_{21}(xx'' + (x')^2) + 2h_{22}x'x''.$$

Ugyanígy

$$(\mathbf{H}\mathbf{y}', \mathbf{y}) = \left(\begin{bmatrix} h_{11}x' + h_{12}x'' \\ h_{21}x' + h_{22}x'' \end{bmatrix}, \begin{bmatrix} x \\ x' \end{bmatrix} \right) = h_{11}x'x + h_{12}x''x + h_{21}(x')^2 + h_{22}x''x'$$

és

$$(\mathbf{H}\mathbf{y}, \mathbf{y}') = \left(\begin{bmatrix} h_{11}x + h_{12}x' \\ h_{21}x + h_{22}x' \end{bmatrix}, \begin{bmatrix} x' \\ x'' \end{bmatrix} \right) = h_{11}xx' + h_{12}(x')^2 + h_{21}xx' + h_{22}x'x''.$$

Összegezve a jobboldalra kapott formulákat a derivált képleteit kapjuk. Megjegyezzük, hogy a $\mathbf{H}$-ra tett feltételre a vektornormához van szükség. ■

Folytatva a példa gondolatmenetét a disszipativitást feltétele a fenti H-normában, hogy az $\mathbf{y} = [x, x']^T$ és $\mathbf{z} = [u, u']^T$ megoldások esetén

$$\frac{d}{dt}\|\mathbf{y} - \mathbf{z}\|_H^2 \leq 0$$

legyen. A lemmát felhasználva

$$\begin{aligned}\frac{d}{dt}\|\mathbf{y} - \mathbf{z}\|_H^2 &= (\mathbf{H}(\mathbf{y}' - \mathbf{z}'), (\mathbf{y} - \mathbf{z})) + (\mathbf{H}(\mathbf{y} - \mathbf{z}), (\mathbf{y}' - \mathbf{z}')) = \\ &= (\mathbf{HA}(\mathbf{y} - \mathbf{z}), (\mathbf{y} - \mathbf{z})) + (\mathbf{H}(\mathbf{y} - \mathbf{z}), \mathbf{A}(\mathbf{y} - \mathbf{z})) = \\ &= (\mathbf{HA}(\mathbf{y} - \mathbf{z}), (\mathbf{y} - \mathbf{z})) + (\mathbf{A^T H}(\mathbf{y} - \mathbf{z}), (\mathbf{y} - \mathbf{z})) = \\ &= ((\mathbf{HA} + \mathbf{A^T H})(\mathbf{y} - \mathbf{z}), (\mathbf{y} - \mathbf{z}))\end{aligned}$$

Olyan $\mathbf{H}$ mátrixot keresünk, melyre

$$((\mathbf{HA} + \mathbf{A^T H})(\mathbf{y} - \mathbf{z}), (\mathbf{y} - \mathbf{z})) \leq 0.$$

Legyen $r > 0$ és

$$\mathbf{H} := \begin{bmatrix} 2k + \frac{r^2}{m} & r \\ r & 2m \end{bmatrix}.$$

Ellenőrizzük Maple-lel, hogy ekkor

$$-(\mathbf{HA} + \mathbf{A^T H}) = 2\frac{r}{m}\begin{bmatrix} k & 0 \\ 0 & m \end{bmatrix}$$

és így

$$\frac{d}{dt}\|\mathbf{y} - \mathbf{z}\|_H^2 = ((\mathbf{HA} + \mathbf{A^T H})(\mathbf{y} - \mathbf{z}), (\mathbf{y} - \mathbf{z})) = -2\frac{r}{m}\left[k(x - u)^2 + m(x' - u')^2\right] \leq 0.$$

Az $r = 0$ esetben $\mathbf{H} = \mathbf{0}$, vagyis nem pozitív definit, de ekkor $\mathbf{HA} + \mathbf{A^T H} = \mathbf{0}$ miatt igaz az a deriváltra vonatkozó állítás. Ezért $r \geq 0$ esetén a definíció szerint disszipatív a rendszer és a H-normában nem nő a megoldások távolsága:

$$\|\mathbf{y}(t_2) - \mathbf{z}(t_2)\|_H \leq \|\mathbf{y}(t_1) - \mathbf{z}(t_1)\|_H, \quad t_2 \geq t_1,$$

$$\|\mathbf{y}(t) - \mathbf{z}(t)\|_H := \left[\left(2k + r + \frac{r^2}{m}\right)(x(t) - u(t))^2 + (2m + r)(x'(t) - u'(t))^2\right]^{1/2}.$$

Ez az eredmény előre vetíti, hogy differenciálegyenlet rendszerek esetén a disszipativitás feltétele a második változó szerinti Jacobi-mátrix valamely tulajdonsága lesz.

2-34. T *(Diff.egy. rendszerek disszipativitása)*

Legyen $\|.\|$ egy vektornorma $\mathbb{R}^n$-en és $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$ folytonosan differenciálható minden argumentuma szerint. Ha $m \geq \mu(f_y(t, \mathbf{y}(t)))$ minden $t \in [0; 1]$, $\mathbf{y}(t) \in \mathbb{R}^n$-re, akkor a diff. egy. rendszer két tetszőleges $\mathbf{u}(t), \mathbf{v}(t)$ megoldására

$$\|\mathbf{u}(t) - \mathbf{v}(t)\| \leq e^{mt}\|\mathbf{u}(0) - \mathbf{v}(0)\|.$$

Így a diff. egy. rendszer disszipatív, ha $\mu(f_y) \leq m \leq 0$.

Bizonyítás. Legyen $\mathbf{z}(t) = \mathbf{u}(t) - \mathbf{v}(t)$, ekkor a Taylor-formulából

$$\|\mathbf{z}(t+h)\| = \|\mathbf{u}(t+h) - \mathbf{v}(t+h)\| = \|\underbrace{\mathbf{u}(t) + hf(t,\mathbf{u}(t)) - \mathbf{v}(t) - hf(t,\mathbf{v}(t))}_{=\mathbf{g}(1)-\mathbf{g}(0)} + O(h^2)\|.$$

A $\mathbf{v}(t)$, $\mathbf{u}(t)$-t összekötő szakasz pontjaira is fel szeretnénk írni ezt a különbséget, ezért vezessük be a következő jelöléseket:

$$\mathbf{w}(t,s) = \mathbf{w}(s) = \mathbf{v}(t) + s \cdot [\mathbf{u}(t) - \mathbf{v}(t)] = \mathbf{v}(t) + s \cdot \mathbf{z}(t), \quad 0 \le s \le 1,$$

ezzel rögzített t-re a $[\mathbf{v}(t); \mathbf{u}(t)]$ szakasz pontjait jelöltük ki, melyen a $\mathbf{g}$ függvény

$$\mathbf{g}(t,s) = \mathbf{g}(s) = \mathbf{w}(s) + hf(t,\mathbf{w}(s)) \in \mathbb{R}^n, \quad 0 \le s \le 1$$

és deriváltja

$$\begin{aligned}\mathbf{g}'(s) &= \mathbf{w}'(s) + hf_y(t,\mathbf{w}(s)) \cdot \mathbf{w}'(s) = \\ &= \mathbf{z}(t) + hf_y(t,\mathbf{w}(s)) \cdot \mathbf{z}(t).\end{aligned}$$

Vegyük észre, hogy a fenti normában szereplő kifejezés felírható a $\mathbf{g}$ függvénnyel

$$\mathbf{u}(t) + hf(t,\mathbf{u}(t)) - \mathbf{v}(t) - hf(t,\mathbf{v}(t)) = \mathbf{g}(1) - \mathbf{g}(0).$$

Ekkor koordinátánként alkalmazva a Newton-Leibniz-formulát

$$\begin{aligned}\mathbf{g}(1) - \mathbf{g}(0) &= \int_0^1 \mathbf{g}'(s)\,ds = \\ &= \int_0^1 \mathbf{z}(t) + h \cdot f_y(t,\mathbf{w}(s)) \cdot \mathbf{z}(t)\,ds = \\ &= \left(\int_0^1 \mathbf{I} + h \cdot f_y(t,\mathbf{w}(s))\,ds\right) \cdot \mathbf{z}(t).\end{aligned}$$

Becsüljük a $\|\mathbf{z}(t+h)\|$ normáját a mátrixnorma illeszkedését felhasználva és az integrálközelítő összegekre a háromszögegyenlőtlenséget felhasználva:

$$\begin{aligned}\|\mathbf{z}(t+h)\| = \|\mathbf{g}(1) - \mathbf{g}(0) + O(h^2)\| &\le \left\|\int_0^1 \mathbf{I} + h \cdot f_y(t,\mathbf{w}(s))\,ds\right\| \cdot \|\mathbf{z}(t)\| + O(h^2) \le \\ &\le \|\mathbf{z}(t)\| \cdot \int_0^1 \|\mathbf{I} + h \cdot f_y(t,\mathbf{w}(s))\|\,ds + O(h^2) \\ &\le \|\mathbf{z}(t)\| \cdot \max_{0\le s\le 1} \|\mathbf{I} + h \cdot f_y(t,\mathbf{w}(s))\| + O(h^2).\end{aligned}$$

A $\|\mathbf{z}(t)\|$ függvény differenciahányadosát felírva

$$\frac{\|\mathbf{z}(t+h)\| - \|\mathbf{z}(t)\|}{h} \le \|\mathbf{z}(t)\| \max_{0\le s\le 1} \frac{\|\mathbf{I} + h \cdot f_y(t,\mathbf{w}(s))\| - 1}{h} + O(h).$$

A $h \to 0+$ határátmenettel

$$\frac{d}{dt}\|\mathbf{z}(t)\| \le \|\mathbf{z}(t)\| \max_{0\le s\le 1} \mu\left(f_y(t,\mathbf{w}(s))\right) \le m \cdot \|\mathbf{z}(t)\|,$$

A kapott egyenlőtlenséget szorozva e^{-mt}-vel

$$e^{-mt}\frac{d}{dt}\|\mathbf{z}(t)\| \le me^{-mt} \cdot \|\mathbf{z}(t)\|.$$

Átrendezve

$$-me^{-mt}\|\mathbf{z}(t)\| + e^{-mt}\frac{d}{dt}\|\mathbf{z}(t)\| = \left(e^{-mt}\|\mathbf{z}(t)\|\right)' \le 0.$$

Tehát az $(e^{-mt}\|\mathbf{z}(t)\|)$ függvény t-ben fogyó, így

$$e^{-mt}\|\mathbf{z}(t)\| \le \|\mathbf{z}(0)\| \quad \Rightarrow \quad \|\mathbf{z}(t)\| \le e^{mt}\|\mathbf{z}(0)\|.$$

■

Megjegyzések.

1. Csak skaláris szorzatból származó vektornorma esetén a tétel másképp is bizonyítható. Rögzített t-re vezessük be a $F(s) = f(t, \mathbf{v}(t) + s \cdot [\mathbf{u}(t) - \mathbf{v}(t)]) : \mathbb{R} \to \mathbb{R}^n$ függvényt, ekkor

$$F'(s) = f_y(t, \mathbf{v}(t) + s \cdot [\mathbf{u}(t) - \mathbf{v}(t)]) \cdot (\mathbf{u}(t) - \mathbf{v}(t)).$$

Innen a koordináta függvényekre felírva a Newton-Leibniz formulát

$$\begin{aligned}\mathbf{u}'(t) - \mathbf{v}'(t) = f(t, \mathbf{u}(t)) - f(t, \mathbf{v}(t)) = F(1) - F(0) = \int_0^1 F'(s)\, ds = \\ = \underbrace{\left(\int_0^1 f_y(t, \mathbf{v}(t) + s \cdot [\mathbf{u}(t) - \mathbf{v}(t)])\, ds \right)}_{\mathbf{A}(t)} \cdot (\mathbf{u}(t) - \mathbf{v}(t)) = \\ = \mathbf{A}(t)(\mathbf{u}(t) - \mathbf{v}(t)).\end{aligned}$$

A logaritmikus norma c) tulajdonsága alapján

$$(\mathbf{u}'(t) - \mathbf{v}'(t), \mathbf{u}(t) - \mathbf{v}(t)) = (\mathbf{A}(t)(\mathbf{u}(t) - \mathbf{v}(t)), \mathbf{u}(t) - \mathbf{v}(t)) \le \mu(\mathbf{A}(t)).$$

A logaritmikus normára vonatkozó háromszögegyenlőtlenségből (lásd Feladatok az integrálközelítő összegekre alkalmazva) következik, hogy

$$\begin{aligned}\mu(\mathbf{A}(t)) = \mu \left(\int_0^1 f_y(t, \mathbf{v}(t) + s \cdot [\mathbf{u}(t) - \mathbf{v}(t)])\, ds \right) \le \\ \le \int_0^1 \mu \left(f_y(t, \mathbf{v}(t) + s \cdot [\mathbf{u}(t) - \mathbf{v}(t)]) \right)\, ds \le m.\end{aligned}$$

Ezután a logaritmikus norma d) részének bizonyításában láttuk, hogy

$$\frac{1}{2} \cdot \frac{d}{dt} \|\mathbf{u}(t) - \mathbf{v}(t)\|^2 = (\mathbf{u}'(t) - \mathbf{v}'(t), \mathbf{u}(t) - \mathbf{v}(t)) \le m \cdot \|\mathbf{u}(t) - \mathbf{v}(t)\|^2,$$

továbbá

$$\|\mathbf{u}(t) - \mathbf{v}(t)\|^2 \le e^{2mt} \|\mathbf{u}(0) - \mathbf{v}(0)\|^2,$$

ami $m \le 0$ esetén a disszpativitást garantálja.

2. A levezetésből látható, hogy tetszőleges $\mathbf{u}(t), \mathbf{v}(t)$ megoldásokra

$$(\mathbf{u}'(t) - \mathbf{v}'(t), \mathbf{u}(t) - \mathbf{v}(t)) = (f(t, \mathbf{u}(t)) - f(t, \mathbf{v}(t)), \mathbf{u}(t) - \mathbf{v}(t)) \le m \cdot \|\mathbf{u}(t) - \mathbf{v}(t)\|^2,$$

amit $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$-beli vektorokra felírva

$$(f(t, \mathbf{u}) - f(t, \mathbf{v}), \mathbf{u} - \mathbf{v}) \le m \cdot \|\mathbf{u} - \mathbf{v}\|^2.$$

Ez az f függvény egyoldalú Lipschitz-folytonosságát jelenti. A baloldali skaláris szorzatra alkalmazzuk a C-B-S egyenlőtlenséget és a második változójára a Lipschitz-feltételt:

$$(f(t, \mathbf{u}) - f(t, \mathbf{v}), \mathbf{u} - \mathbf{v}) \le \|f(t, \mathbf{u}) - f(t, \mathbf{v})\| \cdot \|\mathbf{u}(t) - \mathbf{v}(t)\| \le L_f \cdot \|\mathbf{u}(t) - \mathbf{v}(t)\|^2.$$

A kétféle eredményt összevetve látjuk, hogy m jó lesz Lipschitz-konstansnak, tehát a tételben szereplő feltétel garantálja az egyoldalú Lipschitz-folytonosságot. De gyakran $m \ll L_f$, ezért a megoldások kontraktivitása szempontjából m a jó állandó.

2-16. Példa. Disszipatív-e az 1. fejezetben megadott

$$\mathbf{c}' = \mathbf{b} - \mathbf{B}\mathbf{c}$$

lineáris csererendszer az euklideszi normában?

Megoldás. Mivel $f_y(t, y(t)) = -\mathbf{B}$, ennek a logaritmikus normáját kell becsülnünk.

$$-\mathbf{B} = -\begin{bmatrix} 2 & -1 & 0 \\ -2 & 2.2 & -0.2 \\ 0 & -1.2 & 1.2 \end{bmatrix} = \begin{bmatrix} -2 & 1 & 0 \\ 2 & -2.2 & 0.2 \\ 0 & 1.2 & -1.2 \end{bmatrix} \quad \Rightarrow \quad \mathbf{C} = \frac{1}{2}(-\mathbf{B} - \mathbf{B}^{\mathbf{T}})$$

A Matlab vagy Maple használatával keressük meg $\mathbf{C}$ sajátértékeit:

$$\lambda_1 \approx -3.88, \quad \lambda_2 \approx -0.96, \quad \lambda_3 \approx -0.56.$$

Innen a logaritmikus norma az euklideszi normában $\mu_2(-\mathbf{B}) = \lambda_3 \approx -0.56 \leq 0$, tehát a rendszer a megadott adatokkal disszipatív.

```
with(linalg):
B:=matrix(3,3,[2,-1,1,-2,2.2,-0.2,0,-1.2,1.2]);
C:=evalm(-(B+transpose(B))/2);
lambda:=eigenvals(C);
```

■

2-17. Példa. Mutassuk meg, hogy kis x-ekre disszipatív az euklideszi normában a következő rendszer! Határozza meg azt a maximális x-intervallumot, melyre a disszipativitás garantálható!

$$y_1' = -y_1 + xy_2,$$
$$y_2' = x^2(y_1 - y_2).$$

Megoldás. A Matlab vagy Maple használatával keressük meg $\mathbf{C} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^{\mathbf{T}})$ sajátértékeit és rajzoljuk ki őket x függvényében. ($\mathbf{A}$ a differenciálegyenlet-rendszer jobboldalát jelöli.)

```
with(linalg):
A:=matrix(2,2,[-1,x,x^2,-x^2]);
C:=evalm((A+transpose(A))/2);
lambda:= eigenvals(C);
plot(max(lambda),x=0..1);
```

Az $\mathbf{A}$ logaritmikus normája az euklideszi normában:

$$\mu_2(\mathbf{A}) = \lambda_{\max}(C) \leq 0, \;\; \Leftrightarrow \;\; x \in [0; 1].$$

■

2-35. T *(A lineáris csererendszer megoldásainak tulajdonságai)*
Legyen a lineáris csererendszer $\mathbf{B}$ *mátrixa konstans.*

a) *Ha* $\mathbf{c}(0), \mathbf{b}(t) \geq 0$, *akkor*

$$\mathbf{c}(t) \geq 0 \quad \forall \; t \geq 0 \quad \Leftrightarrow \quad b_{ij} \leq 0, \;\; i \neq j.$$

b) *Ha* $\mathbf{B}$ *M-mátrix,* $\mathbf{b}(t)$ *folytonos minden* t*-re és létezik a*

$$\lim_{t \to \infty} \mathbf{b}(t) = \mathbf{b}_\infty$$

határérték, akkor a rendszernek $t \to \infty$ esetén is van megoldása. Ez a stacionárius megoldás független $\mathbf{c}(0)$-tól (és ilyen értelemben a rendszer stabil).

c) *Ha $\mathbf{b}(t)$ periodikus és $\mathbf{B}$ M-mátrix, akkor a rendszernek van periodikus megoldása.*

Bizonyítás. Lásd a [10] könyv 23. oldalán. ■

3. fejezet

Numerikus módszerek bevezetése

3.1. Alapfogalmak

Már a bevezetőben említettük, hogy a differenciálegyenlet megoldását a $[0;1]$ intervallumon fogjuk meghatározni. Tekintsük ennek egy egyenletes felosztását h lépésközzel:

$$x_0 = 0, \;\; x_{i+1} = x_i + h, \;\; (i = 0, \ldots, N-1), \;\; h = \frac{1}{N},$$

ahol $N \geq 1$ egész szám. Ezt a felosztást ω_h rácsnak nevezzük.

$$\omega_h = \left\{ x_i = ih, \;\; (i = 0, \ldots, N), \;\; h = \frac{1}{N} \right\}$$

A továbbiakban a diszkrét numerikus megoldás értékeit y_i-vel, a pontos megoldás értékeit $y(x_i)$-vel fogjuk jelölni. Az egylépéses numerikus módszer általános alakja

$$y_{i+1} = y_i + h \cdot \Phi(x_i, y_i; h), \;\; (i = 0, \ldots, N-1), \;\; y_0 = y(x_0),$$

ahol $\Phi(\cdot, \cdot; \cdot)$ folytonos mindegyik argumentumában. Például az explicit Euler-módszer esetén

$$\Phi(x_i, y_i; h) = f(x_i, y_i).$$

3-1. Definíció. A numerikus módszer képlethibájának vagy lokális hibájának nevezzük a

$$g(x_i, h) = g_i = \frac{y(x_{i+1}) - y(x_i)}{h} - \Phi(x_i, y(x_i); h)$$

mennyiséget. Általában a pontos megoldás ismerete nélkül, a Taylor-formula segítségével adjuk meg.

Megjegyezzük, hogy a szakirodalomban a fenti mennyiség h-szorosát is lokális hibaként szokás definiálni. Ekkor a többi definíció és a tételek is ennek megfelelően változnak. Lásd a [2] jegyzetben.

3-1. L *Ha egy egylépéses numerikus módszer a p-edfokú Taylor-polinomot használja a közelítéshez, akkor a lokális hiba becslése*

$$|g_i| \leq \frac{M_{p+1}}{(p+1)!} h^p, \quad (i = 0, \ldots, N-1)$$

ahol $M_{p+1} = \max_{x \in [0;1]} |y^{(p+1)}(x)|$.

Bizonyítás. A Taylor-formulát alkalmazva létezik olyan ξ az x_i és x_{i+1} rácspont között, hogy

$$y(x_{i+1}) = y(x_i) + y'(x_i)h + \frac{y''(x_i)}{2!}h^2 + \ldots + \frac{y^{(p+1)}(\xi)}{(p+1)!}h^{p+1}.$$

$$\frac{y(x_{i+1}) - y(x_i)}{h} = y'(x_i) + \frac{y''(x_i)}{2!}h + \ldots + \frac{y^{(p+1)}(\xi)}{(p+1)!}h^p.$$

Ha az egylépéses numerikus módszer a p-edfokú Taylor-polinomot használja a közelítéshez, akkor

$$\Phi(x_i, y(x_i); h) = y'(x_i) + \frac{y''(x_i)}{2!}h + \ldots + \frac{y^{(p)}(x_i)}{p!}h^{p-1}.$$

Innen a lokális hiba

$$g_i = \frac{y(x_{i+1}) - y(x_i)}{h} - \Phi(x_i, y(x_i); h) = \frac{y^{(p+1)}(\xi)}{(p+1)!}h^p$$

a becslése

$$|g_i| \leq \frac{M_{p+1}}{(p+1)!}h^p.$$

■

3-2. Definíció. A numerikus módszer konzisztens, ha képlethibájára az f adott függvényosztályában

$$g_i = O(h^p) \ \ (i = 0, \ldots, N-1).$$

Ekkor $p \geq 1$ a módszer konzisztencia rendje. (A konstrukciók olyanok, hogy p egész.)

Megjegyzés.

Mivel feltettük, hogy $\Phi(\cdot, \cdot; \cdot)$ és y' is folytonos, ezért bármely $x \in [0; 1]$-re $x_n = nh$, $h \to 0$-ra és $\lim_{n\to\infty} x_n = x$ esetén

$$\lim_{n\to\infty} g_n = y'(x) - \Phi(x, y(x); 0) = 0.$$

[13]-ben a 322. oldalon a következő ekvivalens megfogalmazást találjuk: az egylépéses módszer pontosan akkor konzisztens, ha

$$\Phi(x, y; 0) = f(x, y).$$

3-3. Definíció. A numerikus módszer globális hibájának nevezzük a pontos és számított érték különbségét, az

$$e(x_i, h) = e_i = y(x_i) - y_i$$

mennyiséget.

3-4. Definíció. A numerikus módszer stabil, ha létezik olyan $K > 0$, melyre

$$|e_i| \leq K \left(|e_0| + \sum_{j=0}^{i-1} |g_j| h \right), \ \ (i = 1, \ldots, N)$$

teljesül, vagyis a globális hiba felülről becsülhető a lokális hibák összegének konstansszorosával.

3-5. Definíció. A numerikus módszer konvergens az $x \in [0;1]$ pontban, ha n és h olyan, hogy $h = \frac{x}{n}$, $x_n = nh = x$ bármely n-re és teljesül

$$\lim_{n\to\infty} y_n = y(x).$$

A módszer konvergens, ha az adott függvényosztály bármely f függvényére, bármely kezdeti feltétel mellett, bármely $x \in [0;1]$ pontban konvergens.
A numerikus módszer p-edrendben konvergál, ha n és h olyan, hogy $h = \frac{x}{n}$, $x_n = nh = x$ bármely n-re és létezik olyan M (h-tól és n-től független), melyre

$$|y(x) - y_n| \le Mh^p.$$

3-2. L *(Becslések)*
a) *Ha $x \ge 0$, akkor $1 + x \le e^x$.*
b) *Ha $x \ge 0$, akkor $(1 + x)^i \le e^{xi}$.*

Bizonyítás. Az exponenciális függvény hatványsorából triviális.
■

3-3. T *A kezdetiérték-probléma megoldására tekintsük az általános egylépéses módszert*

$$\begin{aligned} y_0 &= y(x_0) \\ y_{i+1} &= y_i + h \cdot \Phi(x_i, y_i; h), \quad (i = 0, \dots, N-1), \end{aligned}$$

és tegyük fel, hogy a Φ függvény a $D = \{(x,y) |\, x \in [0;1], |y - y_0| \le C\}$ téglalapon a második változójában eleget tesz a Lipschitz-feltételnek:

$$|\Phi(x,u;h) - \Phi(x,v;h)| \le L_\Phi |u - v|, \quad (x,u),(x,v) \in D.$$

Ekkor a módszer stabil, azaz

$$|e_i| \le e^{L_\Phi} \left(|e_0| + \sum_{j=0}^{i-1} |g_j| h \right), \quad (i = 1, \dots, N).$$

Bizonyítás. A lokális hiba definíciójából h-val szorozva és átrendezve, majd kivonva a módszer képletét $(i = 0, \dots, N-1)$-re

$$\begin{aligned} y(x_{i+1}) &= y(x_i) + h\Phi(x_i, y(x_i); h) + g_i h \\ y_{i+1} &= y_i + h\Phi(x_i, y_i; h) \\ \hline y(x_{i+1}) - y_{i+1} &= y(x_i) - y_i + h\left(\Phi(x_i, y(x_i); h) - \Phi(x_i, y_i; h)\right) + g_i h \\ e_{i+1} &= e_i + h\left(\Phi(x_i, y(x_i); h) - \Phi(x_i, y_i; h)\right) + g_i h. \end{aligned}$$

Abszolútértéket véve, felhasználva a háromszög-egyenlőtlenséget és a Lipschitz-feltételt

$$|e_{i+1}| \le |e_i| + h \underbrace{|\Phi(x_i, y(x_i); h) - \Phi(x_i, y_i; h)|}_{\le L_\Phi |y(x_i) - y_i| = L_\Phi |e_i|} + |g_i| h = (1 + L_\Phi h)\,|e_i| + |g_i| h.$$

A rekurziót kibontva

$$\begin{aligned}
|e_{i+1}| &\le (1+L_\Phi h)\left((1+L_\Phi h)|e_{i-1}| + |g_{i-1}|h\right) + |g_i|h = \\
&= (1+L_\Phi h)^2|e_{i-1}| + |g_i|h + (1+L_\Phi h)|g_{i-1}|h \le \\
&\dots \\
&\le (1+L_\Phi h)^{i+1}|e_0| + \sum_{k=0}^{i}(1+L_\Phi h)^{i-k}|g_k|h \le \\
&\le (1+L_\Phi h)^{i+1}\left[|e_0| + \sum_{k=0}^{i}\underbrace{(1+L_\Phi h)^{-(k+1)}}_{\le 1}|g_k|h\right] \le \\
&\le (1+L_\Phi h)^{i+1}\left[|e_0| + \sum_{k=0}^{i}|g_k|h\right],
\end{aligned}$$

Mivel $h(i+1) = x_{i+1} \le 1$ és a Lemma miatt $(1+L_\Phi h)^{i+1} \le e^{L_\Phi h(i+1)} \le e^{L_\Phi}$, ezért

$$|e_{i+1}| \le e^{L_\Phi}\left[|e_0| + \sum_{k=0}^{i}|g_k|h\right]$$

ami épp a módszer stabilitását jelenti. ■

Megjegyzés.

A Lipschitz-folytonosság helyett lokális Lipschitz-folytonosságot feltételezve is beláthatjuk az Euler-módszer stabilitását lásd a [10] jegyzet 31. oldalán.

3-4. T *(Konzisztencia + Stabilitás = Konvergencia)*
A kezdetiérték-probléma megoldására tekintsük az általános egylépéses módszert

$$\begin{aligned}
y_0 &= y(x_0) \\
y_{i+1} &= y_i + h\cdot\Phi(x_i, y_i; h), \quad (i = 0, \dots, N-1),
\end{aligned}$$

mely p-edrendben konzisztens ($p \ge 1$), azaz létezik olyan $K > 0$, hogy

$$|g_i| \le Kh^p \quad (i = 0, \dots, N-1)$$

és stabil, azaz

$$|e_i| \le e^{L_\Phi}\left[|e_0| + \sum_{k=0}^{i-1}|g_k|h\right].$$

Ekkor

$$|e_i| \le e^{L_\Phi}Kh^p \quad (i = 0, \dots, N),$$

vagyis p-edrendben konvergens a numerikus módszer.

Bizonyítás. A konvergencia bizonyításához tegyük fel, hogy $y(x_0) = y_0$, azaz pontos kezdeti feltételből indultunk ki ($e_0 = 0$). Ezek után használjuk fel a stabilitás és a p-edrendű konzisztencia fogalmát.

$$\begin{aligned}
|e_i| &\le e^{L_\Phi}\left[\underbrace{|e_0|}_{=0} + \sum_{k=0}^{i-1}|g_k|h\right] \le e^{L_\Phi}\left[\sum_{k=0}^{i-1}Kh^{p+1}\right] \\
&\le e^{L_\Phi}iKh^{p+1} \le e^{L_\Phi}\underbrace{(ih)}_{\le 1}Kh^p \le e^{L_\Phi}Kh^p \quad (i = 1, \dots, N).
\end{aligned}$$

Ha $x \in [0;1]$ tetszőleges és n és h olyan, hogy $h = \frac{x}{n}$ ($x_n = nh$), akkor

$$|y(x) - y_n| = |e_n| \leq (e^{L_\Phi} K)h^p,$$

ahol $e^{L_\Phi} K$ független h-tól és n-től. Ezzel a p-edrendű konvergenciát beláttuk.
■

Megjegyzések.

1. Ha $e_0 = y(x_0) - y_0 \neq 0$, azaz nem pontos $y(0)$-ból indulunk (pl. a kezdeti érték egy másik feladat eredménye), akkor a fenti bizonyításból látszik, hogy $y_i \nrightarrow y(x_i)$. Ilyenkor a kontraktivitás a döntő, ettől nagyobb x-ekre y_i mégis közel lesz $y(x_i)$-hez.

2. A fenti bizonyítás rendszerekre is ugyanígy elvégezhető, ekkor a 3. és 4. tételben a lokális hibák abszolútértéke helyett a normáját vesszük.

3. A Lipschitz-folytonosság helyett lokális Lipschitz-folytonosságot feltételezve is dolgozhatunk lásd a [4] jegyzet 25. oldalán.

4. A becslésben akkor is az e^{L_f} konstans áll, ha egyoldalú a Lipschitz-feltétel abban az értelemben, hogy $L_f > |\mu(f_y)|$ és

$$-L_f(y_1 - y_2) \leq f(x, y_1) - f(x, y_2) \leq \mu(f_y)(y_1 - y_2) \leq 0, \quad y_1 \geq y_2.$$

Korábban láttuk, hogy diff. egy. rendszerek disszipativitása esetén (a kontraktivitási becslésben) a Lipschitz-konstans helyett a $|\mu(f_y)|$ mennyiség játszik fontos szerepet. Ott L_f (amely jóval nagyobb lehet $|\mu(f_y)|$-nál) közömbös.

3-1. Példa. Tekintsük az

$$y' = \arctan(y), \quad y(0) = y_0$$

kezdetiértékproblémát, ahol y_0 adott valós szám. Adjunk becslést az Euler-módszer (egylépéses módszer lásd később, az elsőfokú Taylor-polinomot használja a közelítéshez) globális hibájára!

Megoldás. Euler-módszer esetén $\Phi(x, y; h) = f(x, y)$, így a Lagrange-középérték tétel miatt létezik olyan ξ az u és v között, hogy

$$|f(x, u) - f(x, v)| \leq |f_y(x, \xi)||u - v| = \frac{1}{1 + \xi^2}|u - v| \leq 1 \cdot |u - v|,$$

így $L_\Phi = 1$. Az 1. Lemma bizonyítását $p = 1$-re felhasználva a lokális hiba

$$g_i = \frac{y''(\xi)}{2}h = \frac{\arctan(y(\xi))}{2(1 + y(\xi)^2)}h,$$

a becslése

$$|g_i| \leq \frac{\pi}{4}h.$$

A tételben szereplő $K = \frac{\pi}{4}$ és $p = 1$.

$$|e_i| \leq \frac{e\pi}{4}h \quad (i = 0, \ldots, N).$$

Ez egy elég pesszimista hibabecslés. Ha a $[0;1]$ intervallumon $\varepsilon = 0.01$ pontosságot akarunk elérni, akkor

$$|e_i| \leq \frac{e\pi}{4}h \leq 0.01 \quad \Leftrightarrow \quad h \leq \frac{4\varepsilon}{e\pi} \approx 0.0047,$$

ami $N = 213$-nak felel meg. A valóságban $N = 19$ is elég a megadott pontossághoz. ■

3.2. A legegyszerűbb numerikus módszerek

3.2.1. Euler-módszer

A legegyszerűbb numerikus módszer Euler nevéhez fűződik. Tekintsük a kezdetiérték feladatot $n = 1$-re. Egy adott x_i pontbeli deriváltat az x_i és x_{i+1} ponthoz tartozó különbségi hányadossal közelítjük.

$$\frac{y(x_{i+1}) - y(x_i)}{h} \approx y'(x_i) = f(x_i, y(x_i))$$

Innen $y(x_{i+1})$-et kifejezve az ω_h felosztásbeli pontokra

$$\begin{aligned} x_0 &:= 0, \quad y_0 := y(0) \\ i = 0, \dots, N-1 &: \\ x_{i+1} &:= x_i + h, \\ y_{i+1} &:= y_i + hf(x_i, y_i). \end{aligned}$$

Az általános egylépéses módszerek

$$y_{i+1} = y_i + h \cdot \Phi(x_i, y_i; h)$$

alakjában

$$\Phi(x_i, y_i; h) = f(x_i, y_i).$$

Fogalmazhattunk volna úgy is az előbb, hogy érintő irányban lépünk tovább az x_i pontból. Ez azt jelenti, hogy a megoldás elsőfokú Taylor-polinomját használjuk a közelítéshez. Ha f folytonosan differenciálható $[0;1] \times \mathbb{R}$-en, akkor az 1. Lemmából a lokális hiba becslésére

$$|g_i| \le \frac{M_2}{2} h, \quad (i = 1, \dots, N)$$

ahol $M_2 = \max_{x \in [0;1]} |y''(x)|$, így a módszer elsőrendben konzisztens.
A 3. Tétel és a 4. Tétel szerint, ha f Lipschitzes a 2. változójában, akkor stabil a módszer és a konzisztenciából következően konvergens is.

3-2. Példa. Alkalmazzuk ugyanarra a kezdetiértékproblémára az elsőrendű Euler-módszert h és $\frac{h}{2}$ lépésközzel egyszerre és becsüljük a hibát az $\widetilde{y}_i \approx 2y_{2i}^{(2)} - y_i^{(1)}$ mennyiségre.

Megoldás. A $h/2$ lépésköz miatt kétszer annyit kell lépnünk, így az összehasonlítható értékek:

$$\begin{aligned} y_i^{(1)} &\approx y(x_i) + c_1 h + O(h^2), \\ y_{2i}^{(2)} &\approx y(x_i) + c_1 \frac{h}{2} + O(h^2). \end{aligned}$$

Készítsük el a következő mennyiséget

$$\widetilde{y}_i \approx 2y_{2i}^{(2)} - y_i^{(1)} = y(x_i) + O(h^2),$$

vagyis ezzel az új értékkel egy másodrendű módszert kaptunk. Ezt a technikát Richardson-féle extrapolációnak nevezik, majd általánosan is foglalkozunk vele az RK-módszereknél.
Készítsünk algoritmust belőle:

$$\begin{aligned} y_{i+1}^{(1)} &:= y_i + hf(x_i, y_i), \\ y_{2i+1}^{(2)} &:= y_i + \frac{h}{2} f(x_i, y_i), \\ y_{2i+2}^{(2)} &:= y_{2i+1}^{(2)} + \frac{h}{2} f(x_i + \frac{h}{2}, y_{2i+1}^{(2)}). \end{aligned}$$

Számítsuk ki belőle az új értéket:

$$\begin{aligned}\widetilde{y}_{i+1} &:= 2y^{(2)}_{2i+2} - y^{(1)}_{i+1} = \\ &= 2y^{(2)}_{2i+1} + hf(x_i + \frac{h}{2}, y^{(2)}_{2i+1}) - y_i - hf(x_i, y_i) = \\ &= 2y_i + hf(x_i, y_i) + hf(x_i + \frac{h}{2}, y^{(2)}_{2i+1}) - y_i - hf(x_i, y_i) = \\ &= y_i + hf(x_i + \frac{h}{2}, y_i + \frac{h}{2} f(x_i, y_i))\end{aligned}$$

Ezzel éppen a javított Euler-módszer algoritmusát kaptuk meg, amit később részletesen tárgyalunk. ■

Feladatok

3-1. Készítsünk programot Matlab-ban az Euler-módszerre! Bemenő paraméterei legyenek a következők: f, N, x_0, x_N, y_0. Rajzoljuk ki a pontos megoldást és a módszerrel kapott közelítést! Tesztként a következő példákkal dolgozzunk:

- $y'(x) = 1, \;\; [0;1], \;\; y(0) = 1$
  A megoldás: $y(x) = x$, a módszerrel a pontos megoldást kapjuk.
- $y'(x) = y, \;\; [0;1], \;\; y(0) = 1$
  A megoldás: $y(x) = e^x$.
- $y'(x) = x + y, \;\; [0;1], \;\; y(0) = 1$
  A megoldás: $y(x) = 2e^x - x - 1$.
- $y'(x) = \cos(x)y, \;\; [0;25], \;\; y(0) = 1$
  A megoldás: $y(x) = \exp(\sin(x))$.

3-2. Készítsünk teszt programot Matlab-ban az Euler-módszerre, mellyel kísérletileg igazoljuk a módszer konvergenciáját! Bemenő paraméterei legyenek a következők: f, x_0, x_N, y_0. Legyen az intervallum végpontja ($x_N = 1$), ahol a konvergenciát igazoljuk, a felosztások számát mindig duplázzuk ($N = 10, 20, 40, 80, 160$, stb.) és irassuk ki y_N értékét minden esetben. Próbáljuk ki, hogy ha nem pontos kezdeti feltételből indulunk ki, akkor nem kapunk konvergenciát!

3-3. Készítsünk teszt programot Matlab-ban az Euler-módszerre, mely megadott kezdetiérték probléma és megoldás esetén a következő táblázathoz készít adatokat! Bemenő paraméter az N értéke legyen.

x_i	y_i	$y(x_i)$	e_i

3-4. Készítsünk teszt programot Matlab-ban az Euler-módszerre, mely megadott kezdetiérték probléma és megoldás esetén a következő táblázathoz készít adatokat! Bemenő paraméter az N értékeket tartalmazó vektor legyen vagy egyetlen N, melyet duplázunk a programban. A kapott adatok elemzésekor az $e_N/e_{N/2}$ hányadosok $h = \frac{1}{2}$-hez konvergálnak, innen látszik a módszer elsőrendű konvergenciája.

N	h	y_N	$y(x_N)$	e_N	$e_N/e_{N/2}$

3.2.2. Implicit Euler-módszer

Egy adott x_{i+1} pontbeli deriváltat az x_i és x_{i+1} ponthoz tartozó különbségi hányadossal közelítjük.

$$\frac{y(x_{i+1}) - y(x_i)}{h} \approx y'(x_{i+1}) = f(x_{i+1}, y(x_{i+1}))$$

Innen $y(x_{i+1})$ közelítésére az

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1})$$

egyenletet kapjuk, amit általában fixpontiterációval tudunk megoldani egy explicit módszerrel kapott x_{i+1} pontbeli $y_{i+1}^{(0)}$ közelítés birtokában.

$$y_{i+1}^{(k+1)} := y_i + hf(x_{i+1}, y_{i+1}^{(k)}) \quad (k = 0, 1, 2, \ldots)$$

Az implicit módszerek részletesebb elemzésével később foglalkozunk.

3-3. Példa. Az iteráció konvergenciájához vizsgáljuk a $\varphi(z) = y_i + hf(x_{i+1}, z)$ függvény kontraktivitását. Adjunk feltételt, mellyel bármely kezdeti feltétel esetén konvergens az iteráció.

Megoldás. Feltételezve, hogy f a 2. változójában Lipschitzes

$$\begin{aligned} |\varphi(z_1) - \varphi(z_2)| &= |y_i + hf(x_{i+1}, z_1) - y_i - hf(x_{i+1}, z_2)| = \\ &= h|f(x_{i+1}, z_1) - f(x_{i+1}, z_2)| \leq \\ &\leq hL|z_1 - z_2|, \end{aligned}$$

vagyis az iteráció csak akkor konvergens bármely kezdeti feltétel esetén, ha $0 < hL \leq q < 1$.
■

3-4. Példa. Nézzük meg, hogy az $y'(x) = a(x)y(x) + b(x)$ elsőrendű lineáris differenciálegyenlet esetén az implicit Euler-módszer milyen megoldandó egyenletet ad. Ekkor nincs szükségünk iterációra.

Megoldás. Az $y_{i+1} = y_i + hf(x_{i+1}, y_{i+1})$ egyenletbe helyettesítsük be a lineáris jobboldalt, így

$$y_{i+1} = y_i + h\left[a(x_{i+1})y_{i+1} + b(x_{i+1})\right].$$

A kapott egyenletünk most lineáris, ezért y_{i+1}-et ki tudjuk fejezni belőle.

$$\begin{aligned} (1 - ha(x_{i+1}))y_{i+1} = y_i + hb(x_{i+1}) \quad &\rightarrow \\ y_{i+1} = \frac{1}{1 - ha(x_{i+1})}y_i + \frac{hb(x_{i+1})}{1 - ha(x_{i+1})} \end{aligned}$$

Tehát az algoritmusunk

$$\begin{aligned} y_0 &:= y(0) \quad \text{adott}, \\ y_{i+1} &:= \frac{1}{1 - ha(x_{i+1})}y_i + \frac{hb(x_{i+1})}{1 - ha(x_{i+1})} \quad (i = 0, 1, \ldots, N-1). \end{aligned}$$

■

3-5. Példa. Nézzük meg az előző példát rendszerekre. Legyen $\mathbf{y}'(x) = \mathbf{A}\mathbf{y}(x) + \mathbf{b}(x)$ elsőrendű lineáris differenciálegyenletrendszer, $\mathbf{A} \in \mathbb{R}^{m\times m}, \mathbf{y}, \mathbf{b} \in \mathbb{R}^n$ és írjuk fel az implicit Euler-módszert. Ekkor sem lesz szükségünk iterációra.

Megoldás. Az $y_{i+1} = y_i + hf(x_{i+1}, y_{i+1})$ egyenletbe helyettesítsük be a lineáris jobboldalt, így

$$\mathbf{y_{i+1}} = \mathbf{y_i} + h\left[\mathbf{A}\mathbf{y_{i+1}} + \mathbf{b}(x_{i+1})\right].$$

Innen ki tudjuk fejezni $\mathbf{y_{i+1}}$-et egy lineáris egyenletrendszerrel.

$$(\mathbf{I} - h\mathbf{A})\mathbf{y_{i+1}} = \mathbf{y_i} + h\mathbf{b}(x_{i+1})$$

Az elején kiszámítjuk az $\mathbf{I} - h\mathbf{A}$ mátrix LU-felbontását és minden lépésben megoldjuk vele a kapott lineáris egyenletrendszert. ■

3.2.3. Javított (módosított) Euler-módszer

Az Euler-módszerhez képest x_i-ből az $x_i + \frac{h}{2}$ pontbeli meredekséggel lépünk tovább.
Ehhez $f\left(x_i + \frac{h}{2}, y(x_i + \frac{h}{2})\right)$-ben az $y(x_i + \frac{h}{2})$ értéket a Taylor-formulával közelítjük

$$y\left(x_i + \frac{h}{2}\right) \approx y(x_i) + \frac{h}{2}f(x_i, y(x_i)) = y_i + \frac{h}{2}f(x_i, y_i)$$

és ezzel lépünk tovább x_i-ből:

$$y_{i+1} := y_i + hf\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}f(x_i, y_i)\right)$$

Algoritmikusan felírva:

$$\begin{aligned}
x_0 &:= 0, \quad y_0 := y(0)\\
i = 0, \ldots, N-1 &:\\
x_{i+1} &:= x_i + h,\\
k_1(x_i, y_i; h) = k_1 &:= f(x_i, y_i),\\
k_2(x_i, y_i; h) = k_2 &:= f\left(\underbrace{x_i + \frac{h}{2}}_{x_{i+1/2}}, \underbrace{y_i + \frac{h}{2}k_1}_{y_{i+1/2}}\right),\\
y_{i+1} &:= y_i + hk_2.
\end{aligned}$$

Ezzel egy egyszerű Runge-Kutta-módszert kaptunk meg, a fenti algoritmikus írásmód is ennek felel meg.

3-5. T *Ha $f \in C^2([0;1] \times \mathbb{R})$, akkor a javított Euler-módszer $p = 2$ rendben konzisztens.*

Bizonyítás. Mielőtt nekilátnánk a bizonyításnak, írjuk fel a kétváltozós f függvényre a Taylor-formulát a harmadrendű maradéktaggal:

$$\begin{aligned}
f(x + \delta x, y + \delta y) = \{f + f_x\delta x + f_y\delta y + \frac{1}{2}[f_{xx}\delta^2 x + 2f_{xy}\delta x\delta y + f_{yy}\delta^2 y]\}(x, y) +\\
+ O(\delta^3 x + \delta^3 y).
\end{aligned}$$

Ezt kell majd alkalmaznunk k_2-re úgy, hogy még egy belső függvényünk is van a 2. változóban. Írjuk fel a módszer lokális hibáját:

$$\begin{aligned}
g(x_i, h) = g_i &:= \frac{y(x_{i+1}) - y(x_i)}{h} - k_2(x_i, y(x_i); h)\\
k_2(x_i, y(x_i); h) &= f\left(x_i + \frac{h}{2}, y(x_i) + \frac{h}{2}f(x_i, y(x_i))\right) \quad \rightarrow\\
hg_i &= y(x_i + h) - y(x_i) - hf\left(x_i + \frac{h}{2}, y(x_i) + \frac{h}{2}f(x_i, y(x_i))\right).
\end{aligned}$$

Írjuk fel hg_i-re az $(x_i, 0)$ pont körül a Taylor-formulát h hatványai szerint a harmadrendű maradéktaggal. Ekkor léteznek olyan $\vartheta_i \in (0;1)$, $\xi_i \in (0; \frac{1}{2})$ és $\eta_i \in (0; \frac{1}{2} f(x_i, y(x_i)))$ melyre

$$\begin{aligned} hg_i = hy'(x_i) + \frac{1}{2}h^2 y''(x_i) + \frac{1}{6}h^3 y'''(x_i + \vartheta_i h) - \\ - h\left(f + \frac{h}{2}(f_x + f_y f)\right)(x_i, y(x_i)) - \\ - h\left(\frac{1}{2}\left(\frac{h}{2}\right)^2 [f_{xx} + 2f_{xy}f + f_{yy}f^2]\right)(x_i + \xi_i h, y(x_i) + \eta_i h). \end{aligned}$$

Figyelembe véve, hogy $y' = f$ és $y'' = f_x + f_y f =: F_1 f$, kapjuk, hogy csak a maradéktagok maradnak meg.

$$hg_i = \frac{1}{6}h^3 y'''(x_i + \vartheta_i h) - \left\{\frac{h^3}{8}[f_{xx} + 2f_{xy}f + f_{yy}f^2]\right\}(x_i + \xi_i h, y(x_i) + \eta_i h)$$

Becsülve

$$|g_i| \le \frac{1}{6}h^2 M_3 + \frac{h^2}{8}\|F_2 f\|_{C[0;1]} = O(h^2),$$

ahol

$$M_3 := \max_{x\in[0;1]} |y'''(x)|, \quad F_2 f := f_{xx} + 2f_{xy}f + f_{yy}f^2.$$

Megjegyezzük, hogy az $F_2 f$-fel jelölt függvény nem azonos y'''-tal, ugyanis

$$y''' = (F_1 f)' = f_{xx} + f_{xy}f + f_{xy}f + f_{yy}f^2 + f_y(f_x + f_y f) = F_2 f + f_y F_1 f.$$

■

Az f kétszer folytonos differenciálhatóságát és a 2. változója szerinti Lipschitz-folytonosságát feltételezve a 3. és 4. Tételből következik a módszer stabilitása és konvergenciája. Ezek a tulajdonságok az általánosabb Runge-Kutta-módszerek tárgyalásánál bizonyított tétellel is megkaphatóak.

Feladatok

3-5. Készítsünk programot Matlab-ban a javított Euler-módszerre! Bemenő paraméterei legyenek f, N, x_0, x_N, y_0. Rajzoljuk ki a pontos megoldást és a módszerrel kapott közelítést! Tesztként a következő példákkal dolgozzunk:

- $y'(x) = 1$, $[0;1]$, $y(0) = 1$
  A megoldás: $y(x) = x$, a módszerrel a pontos megoldást kapjuk.
- $y'(x) = y$, $[0;1]$, $y(0) = 1$
  A megoldás: $y(x) = e^x$.
- $y'(x) = x + y$, $[0;1]$, $y(0) = 1$
  A megoldás: $y(x) = 2e^x - x - 1$.
- $y'(x) = \cos(x)y$, $[0;25]$, $y(0) = 1$
  A megoldás: $y(x) = \exp(\sin(x))$.

3-6. Készítsünk teszt programot Matlab-ban a javított Euler-módszerre, mellyel kísérletileg igazoljuk a módszer konvergenciáját! Bemenő paraméterei legyenek f, x_0, x_N, y_0. Legyen az intervallum végpontja ($x_N = 1$), ahol a konvergenciát igazoljuk, a felosztások számát mindig duplázzuk ($N = 10, 20, 40, 80, 160$, stb.) és irassuk ki y_N értékét minden esetben! Próbáljuk ki, hogy ha nem pontos kezdeti feltételből indulunk ki, akkor nem kapunk konvergenciát!

3-7. Készítsünk teszt programot Matlab-ban a javított Euler-módszerre, mely megadott kezdetiérték probléma és megoldás esetén a következő táblázathoz készít adatokat! Bemenő paraméter az N értéke legyen.

x_i	y_i	$y(x_i)$	e_i

3-8. Készítsünk teszt programot Matlab-ban a javított Euler-módszerre, mely megadott kezdetiérték probléma és megoldás esetén a következő táblázathoz készít adatokat! Bemenő paraméter az N értékeket tartalmazó vektor legyen vagy egyetlen N, melyet duplázunk a programban. A kapott adatok elemzésekor az $e_N/e_{N/2}$ hányadosok $h^2 = \frac{1}{4}$-hez konvergálnak, innen látszik a módszer másodrendű konvergenciája.

N	h	y_N	$y(x_N)$	e_N	$e_N/e_{N/2}$

3.3. A stabilitás általános definíciója

Legyen $\mathbf{F_h} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ invertálható leképezés, ahol $h = \frac{1}{N}$.
(Vagy általánosabban $N = N(h)$ azzal a tulajdonsággal, hogy $h \to 0$ esetén $N(h) \to \infty$.)
Legyen $\|.\|$ és $\|.\|_*$ két vektornorma, melyekre $\|\mathbf{e_h}\| = 1$, $\|\mathbf{e_h}\|_* = 1$ függetlenül N-től, ahol $\mathbf{e_h} = (1, \ldots, 1)^T \in \mathbb{R}^{N+1}$. Megoldandó az

$$\mathbf{F_h}(\mathbf{y_h}) = \mathbf{u_h}$$

egyenlet, ahol $\mathbf{u_h} = (u_0, \ldots, u_N)^T \in \mathbb{R}^{N+1}$ adott vektor és $\mathbf{y_h} = (y_0, \ldots, y_N)^T \in \mathbb{R}^{N+1}$ keresett vektor. Adott egy további egyenlet a $\mathbf{v_h} = (v_0, \ldots, v_N)^T \in \mathbb{R}^{N+1}$ jobboldallal

$$\mathbf{F_h}(\mathbf{z_h}) = \mathbf{v_h}.$$

3-6. Definíció. Az ${\mathbf{F_h}}^{-1}$ leképezést (a megoldási operátort) stabilnak nevezzük, ha létezik olyan h-tól független $M > 0$ konstans, melyre

$$\|{\mathbf{F_h}}^{-1}(\mathbf{u_h}) - {\mathbf{F_h}}^{-1}(\mathbf{v_h})\| \leq M\|\mathbf{u_h} - \mathbf{v_h}\|_*,$$

minden $\mathbf{u_h}, \mathbf{v_h} \in \mathbb{R}^{N+1}$-re. Ekkor a közönséges diff. egyenleteket megoldó módszert 0-stabilnak nevezzük. Ennek ekvivalens felírása:

$$\|\mathbf{y_h} - \mathbf{z_h}\| \leq M\|\mathbf{F_h}(\mathbf{y_h}) - \mathbf{F_h}(\mathbf{z_h})\|_*.$$

3-6. Példa. Mutassuk meg a definíció alkalmazását az Euler-módszerre!

Megoldás. Legyen $\mathbf{F_h}$ az $\mathbf{y} \in \mathbb{R}^{N+1}$-en értelmezett operátor, melyre

$$(\mathbf{F_h}(\mathbf{y}))_k = \begin{cases} y_0 & \text{ha } k = 0 \\ \frac{y_k - y_{k-1}}{h} - f(x_{k-1}, y_{k-1}) & \text{ha } 0 < k \leq N, \end{cases}$$

ahol $f \in Lip(y)$, az $\mathbf{F_h}$ értelmezési tartományán a maximum normát vesszük:

$$\|\mathbf{y}\| := \|\mathbf{y}\|_\infty = \max_{k=0,\ldots,N} |y_k|,$$

a képterében a $\|.\|_*$ norma legyen a h-val súlyozott 1-es norma:

$$\|\mathbf{y}\|_* := \frac{1}{2}\left(|y_0| + \sum_{k=1}^{N} |y_k| h\right).$$

Legyen $\mathbf{y_h}$ az Euler-módszerrel kapott y_k értékek vektora és $\overrightarrow{\mathbf{y}}$ a pontos megoldás $y(x_k)$ értékeiből álljon, ekkor az $\mathbf{F_h}$ leképezés eredménye:

$$\mathbf{u_h} := \mathbf{F_h}(\mathbf{y_h}) = \begin{bmatrix} y_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{és} \quad \mathbf{v_h} := \mathbf{F_h}(\overrightarrow{\mathbf{y}}) = \begin{bmatrix} y(0) \\ g_0 \\ \vdots \\ g_{N-1} \end{bmatrix}.$$

Vegyük észre, hogy a pontos megoldás rácspontbeli értékei megkaphatók az Euler-módszer eredményeként, ha a módszer jobboldalát a lokális hibákkal perturbáljuk. Ekkor a korábbi eredményekből $M = 2e^{L_f}$ konstanssal és $e_0 = 0$-val

$$\|\mathbf{y_h} - \overrightarrow{\mathbf{y}}\| = \max_{i=0,\ldots,N} |y_i - y(x_i)| \le \frac{M}{2}\left(|e_0| + \sum_{k=1}^{N} |g_{k-1}| h\right) =$$
$$= \frac{M}{2}(|e_0| + \|\mathbf{g_h}\|_*) = M\|\mathbf{u_h} - \mathbf{v_h}\|_* = M\|\mathbf{F_h}(\mathbf{y_h}) - \mathbf{F_h}(\overrightarrow{\mathbf{y}})\|_*.$$

■

Megjegyzések.

1. A definíció egybevág a korábban tanult algoritmus numerikus stabilitása definícióval. A kimenő adatok hibája a bemenő adatok konstans-szorosával becsülhető felülről.

2. A fenti fogalom a normáktól függ. Lásd [10]-ben: a Spijker-normában nem stabil a középpont szabály.

3. A stabilitás tetszőleges kezdeti értékekre és a jobboldalak egész osztályára vonatkozik. Ha egyetlen $y(0)$ értékre vagy f függvényre nem igaz a becslés, akkor már instabil a módszer. Fordítva, ha egy módszerről kiderül, hogy instabil, attól még lehet, hogy egy speciális kezdeti értékre vagy függvényre teljesül az egyenlőtlenség.

3.4. Változó lépéstávolság

A modern programcsomagok szinte kizárólag változó lépésközzel dolgoznak, annak érdekében, hogy hatékonyabban állítsák elő a numerikus megoldást.
Legyen

$$\omega_h = \{x_i,\ 0 = x_0 < x_1 < \ldots < x_N = 1,\ h_i = x_{i+1} - x_i,\ (i = 0, \ldots, N-1)\}.$$

Itt $h = h(N) = \frac{1}{N}$ az átlagos lépéstávolság. Ahhoz, hogy h_i nullához tartson minden i-re egyszerre, megköveteljük, hogy létezzen c_1, c_2 konstans $0 < c_1 \le 1 \le c_2$, melyre

$$c_1 h \le h_i \le c_2 h \ \ (i = 0, \ldots, N-1).$$

A képlethibák normáját ezzel az átlagos h lépéstávolsággal definiáljuk:

$$\|\mathbf{g}\| = \sum_{k=0}^{N-1} |g_i| h.$$

3-6. T *(Konzisztencia + Φ Lip = Konvergencia)*
A kezdetiérték-probléma megoldására tekintsük az általános egylépéses módszert változó lépésközzel (mely teljesíti h_i-kre az előző feltételeket)

$$y_{i+1} = y_i + h_i \cdot \Phi(x_i, y_i; h_i), \;\; (i = 0, \ldots, N-1), \;\; y(x_0) = y_0,$$

mely p-edrendben konzisztens ($p \geq 1$), azaz létezik olyan $K > 0$, hogy

$$|g_i| \leq K h^p \quad (i = 0, \ldots, N-1)$$

és Φ a második változójában eleget tesz a Lipschitz-feltételnek. Ekkor

$$|e_i| \leq c_2 e^{L_\Phi} K h^p = O(h^p) \quad (i = 0, \ldots, N-1),$$

vagyis p-edrendben konvergens a numerikus módszer.

Bizonyítás. Nézzük végig a 1. és 4. tétel bizonyítását a változó lépéstávolságnak megfelelően az $|e_i|$ becslésétől kezdődően. Abszolútértéket véve, felhasználva a háromszög-egyenlőtlenséget és a Lipschitz-feltételt

$$|e_{i+1}| \leq (1 + L_\Phi h_i)\, |e_i| + |g_i| h_i.$$

A rekurziót kibontva

$$\begin{aligned}
|e_{i+1}| &\leq e^{L_\Phi h_i} \left(|e_i| + |g_i| h_i\right) \leq \\
&\leq e^{L_\Phi h_i} \left(e^{L_\Phi h_{i-1}} \left[|e_{i-1}| + |g_{i-1}| h_{i-1}\right] + |g_i| h_i\right) \leq \\
&\leq e^{L_\Phi (h_i + h_{i-1})} \left(|e_{i-1}| + |g_{i-1}| h_{i-1} + |g_i| h_i\right) \leq \ldots \\
&\leq e^{L_\Phi x_{i+1}} \left(|e_0| + \sum_{k=0}^{i} |g_k| c_2 h\right) \leq \\
&\leq e^{L_\Phi} \left(|e_0| + c_2 \|\mathbf{g}\|\right) \leq c_2 e^{L_\Phi} \left(|e_0| + \|\mathbf{g}\|\right).
\end{aligned}$$

A konvergencia bizonyításához tegyük fel, hogy $y(x_0) = y_0$, azaz pontos kezdeti feltételből indultunk ki és a konzisztenciából $|g_i| \leq K h^p$, ezért

$$\|\mathbf{g}\| \leq \sum_{k=0}^{N-1} |g_k| c_2 h \leq K h^p c_2 \sum_{k=0}^{N-1} h \leq c_2 K h^p.$$

Ezt beírva a globális hiba becslésbe

$$|e_i| \leq c_2 e^{L_\Phi} \|\mathbf{g}\| \leq c_2 e^{L_\Phi} c_2 K h^p \quad (i = 1, \ldots, N).$$

Ha $x \in [0; 1]$ tetszőleges és n és h olyan, hogy $h = \frac{x}{n}$, így $x_n = nh$, akkor

$$|y(x) - y_n| = |e_n| \leq ((c_2)^2 e^{L_\Phi} K) h^p,$$

ahol $(c_2)^2 e^{L_\Phi} K$ független h-tól és n-től. Ezzel a p-edrendű konvergenciát beláttuk. ■

4. fejezet

Explicit Runge–Kutta-módszerek (ERK)

4.1. Az ERK-módszerek általános jellemzése

A javított Euler-módszernél láttuk, hogy egynél magasabbrendű módszer is konstruálható rekurzív függvényhívásokkal anélkül, hogy az f deriváltjaira szükség lenne. Az ott látott ötletet követve konstruáljuk a Runge-Kutta-módszercsaládot. Először rekurzív módon definiáljuk a k_i számokat (a könnyebb áttekinthetőség kedvéért az argumentumokat elhagyjuk). A k_j számokhoz általában az előző $k_1, k_2, \ldots, k_{j-1}$ számokat használjuk fel:

$$\begin{aligned}
k_1 &= k_1(x_i, y_i, h) := f(x_i, y_i),\\
k_2 &= k_2(x_i, y_i, h) := f(x_i + ha_2, y_i + hb_{21}k_1),\\
&\ldots\\
k_j &= k_j(x_i, y_i, h) := f\left(x_i + ha_j, y_i + h\sum_{l=1}^{j-1} b_{jl}k_l\right), \quad j = 1, \ldots, s
\end{aligned}$$

majd ezeket a c_j súlyokkal kombinálva kapjuk az új y_{i+1} értéket:

$$y_{i+1} = y_i + h\sum_{j=1}^{s} c_j k_j.$$

Az a_j, b_{jl} és c_j számok jellemzik a módszert, ezek függetlenek f, y és h-tól. Továbbá

$$\sum_{j=1}^{s} c_j = 1 \quad \text{és} \quad a_j = \sum_{l=1}^{j-1} b_{jl}.$$

A c_j-kre tett feltétel azért kell, hogy a módszer $f \equiv 1$-re pontos legyen. Az a_j-kre pedig azért, mert általában jóval több a paraméterünk, mint a feltételünk, másrészt az $y_i + h\sum_{l=1}^{j-1} b_{jl}k_l$ pediktorok ekkor pontos értékek. Az s számot a módszer lépcsőszámának nevezzük. Az s növelésével magasabbrendű módszereket lehet szerkeszteni. A Runge-Kutta képleteket a következő áttekinthető alakban szokás megadni, amelyet Butcher-táblázatnak nevezünk, a $\mathbf{B} = (b_{jl})$ mátrixot pedig Butcher-mátrixnak.

$\mathbf{Be}$	$\mathbf{B}$
	$\mathbf{c^T}$

Ha a módszer explicit, akkor a $\mathbf{B}$ mátrix szigorú alsóháromszög mátrix. Az előző fejezetben említett módszerek a következő táblázatokkal írhatóak fel:

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

RK1 Euler, $p = s = 1$ | impl. Euler, $p = s = 1$ | RK2 javított Euler, $p = s = 2$

Megjegyzés.

A RK-módszereket másképp is származtathatjuk. Ha a kezdetiértékproblémát integráljuk $[x_i; x_{i+1}]$-en, majd az $x := x_i + th$ helyettesítést alkalmazzuk, akkor

$$y_{i+1} - y_i = y(x_{i+1}) - y(x_i) = \int_{x_i}^{x_{i+1}} f(x, y(x))\,dx = h\int_0^1 f(x_i + ht, y(x_i + ht))\,dt$$

Az integrált egy kvadratúra-formulával közelítjük, ahol $a_j \in [0; 1]$-k az alappontok és c_j-k a kvadratúra-formula együtthatói.

$$y_{i+1} = y_i + \sum_{j=1}^{s} c_j f(x_i + a_j h, y(x_i + a_j h))$$

Mivel az $f(x_i + a_j h, y(x_i + a_j h))$ értékeket nem ismerjük, ezeket az $y_i, k_1, \ldots, k_{j-1}$ értékekkel közelítjük. Ha $a_1 = 0$, akkor $k_1 = f(x_i, y_i)$. k_j-t a már meglévő k_l-ek lineáris kombinációjaként állítjuk elő, általános alakja:

$$k_j = f(x_i + a_j h, y_i + hb_{j1}k_1 + \ldots + hb_{jj-1}k_{j-1}) = f\left(x_i + a_j h, y_i + h\sum_{l=1}^{j-1} b_{j1}k_1\right).$$

Ha az integrálközelítésre az érintő formulát használjuk, akkor a javított Euler-módszert kapjuk. Ha a trapéz-formulával közelítjük az integrált, akkor a trapéz-módszert kapjuk (implicit módszert). A korábbi algebrai megközelítés előnye, hogy sokkal többféle formulát konstruálhatunk, mint a most említett kvadratúrákkal.

4.2. Harmadrendű ERK-módszerek

A harmadrendű RK-módszer alakja a következő lesz:

$$\begin{aligned} &k_1 = f(x, y), \\ &k_2 = f(x + ha_2, y + hb_{21}k_1) \quad \rightarrow \quad a_2 = b_{21}, \\ &\underline{k_3 = f\left(x + ha_3, y + hb_{31}k_1 + hb_{32}k_2\right) \quad \rightarrow \quad a_3 = b_{31} + b_{32},} \\ &\quad y_{i+1} = y_i + h\sum_{j=1}^{3} c_j k_j \quad \rightarrow \quad 1 = c_1 + c_2 + c_3 \end{aligned}$$

A levezetéshez segítségünkre lesz a következő Taylor-formulára

$$\begin{aligned} f(x + \delta x, y + \delta y) = \{f + f_x\delta x + f_y\delta y + \frac{1}{2}[f_{xx}\delta^2 x + 2f_{xy}\delta x\delta y + f_{yy}\delta^2 y]\}(x, y) + \\ + O(\delta^3 x + \delta^3 y). \end{aligned}$$

Ha $f \in C^3([0;1]\times\mathbb{R})$-en, akkor a k_1, k_2, k_3 kifejezések Taylor-sorba fejthetők h szerint az $(x_i, y(x_i), 0)$ pont körül. A továbbítakban az argumentumokat nem írjuk ki. Alkalmazzuk a fenti kifejezést k_2 felírására a $\delta x = ha_2$ és $\delta y = hb_{21}f$ helyettesítésekkel.

$$\begin{aligned}
k_1 &= f,\\
k_2 &= f + ha_2f_x + hb_{21}ff_y + \frac{1}{2}[(ha_2)^2f_{xx} + 2h^2a_2b_{21}ff_{xy} + (hb_{21}f)^2f_{yy}] + O(h^3) =\\
&= f + ha_2f_x + hb_{21}ff_y + \frac{h^2}{2}[a_2^2f_{xx} + 2a_2b_{21}ff_{xy} + b_{21}^2f^2f_{yy}] + O(h^3) =\\
&= f + hb_{21}[f_x + ff_y] + \frac{(hb_{21})^2}{2}[f_{xx} + 2ff_{xy} + f^2f_{yy}] + O(h^3) =\\
&= f + hb_{21}F_1f + \frac{(hb_{21})^2}{2}F_2f + O(h^3)
\end{aligned}$$

Alkalmazzuk a fenti kifejezést k_3-ban a $\delta x = ha_3$ és a

$$\begin{aligned}
\delta y &= hb_{31}k_1 + hb_{32}k_2 = hb_{31}f + hb_{32}\left(f + hb_{21}F_1f + \frac{(hb_{21})^2}{2}F_2f + O(h^3)\right) =\\
&= ha_3f + h^2b_{21}b_{32}F_1f + O(h^3)
\end{aligned}$$

helyettesítésekkel.

$$\begin{aligned}
k_3 &= f + ha_3f_x + f_y\left[ha_3f + h^2b_{21}b_{32}F_1f + O(h^3)\right] +\\
&\frac{1}{2}\left(f_{xx}(ha_3)^2 + 2f_{xy}(ha_3)\left[ha_3f + h^2b_{21}b_{32}F_1f + O(h^3)\right] +\right.\\
&\left. + f_{yy}\left[ha_3f + h^2b_{21}b_{32}F_1f + O(h^3)\right]^2\right) + O(h^3) =\\
&= f + ha_3(f_x + f_yf) + h^2b_{21}b_{32}f_yF_1f + \frac{1}{2}(ha_3)^2\left(f_{xx} + 2ff_{xy} + f^2f_{yy}\right) + O(h^3) =\\
&= f + ha_3F_1f + h^2b_{21}b_{32}f_yF_1f + \frac{1}{2}(ha_3)^2F_2f + O(h^3)
\end{aligned}$$

Írjuk fel a lokális hibát:

$$\begin{aligned}
hg_i &= y(x_{i+1}) - y(x_i) - h(c_1k_1 + c_2k_2 + c_3k_3) =\\
&= y(x_{i+1}) - y(x_i) - h\left(c_1f + c_2(f + hb_{21}F_1f + \frac{(hb_{21})^2}{2}F_2f) +\right.\\
&\left. + c_3(f + ha_3F_1f + h^2b_{21}b_{32}f_yF_1f + \frac{1}{2}(ha_3)^2F_2f)\right) + O(h^4) =\\
&= y(x_{i+1}) - y(x_i) - h(c_1 + c_2 + c_3)f - h^2(c_2b_{21} + c_3a_3)F_1f -\\
&- \frac{h^3}{2}(c_2b_{21}^2 + c_3a_3^2)F_2f - h^3c_3b_{21}b_{32}f_yF_1f + O(h^4)
\end{aligned}$$

$y(x_{i+1})$ sorfejtését felhasználva:

$$\begin{aligned}
hg_i &= y(x_i) + hf + \frac{h^2}{2}y'' + \frac{h^3}{6}y''' + O(h^4) - y(x_i) - h(c_1 + c_2 + c_3)f - h^2(c_2b_{21} + c_3a_3)F_1f -\\
&- \frac{h^3}{2}(c_2b_{21}^2 + c_3a_3^2)F_2f - h^3c_3b_{21}b_{32}f_yF_1f + O(h^4)
\end{aligned}$$

Írjuk fel a feltételeket, hogy a lokális hiba képletéből kiessenek a h, h^2-es és h^3-s tagok.

$$\begin{aligned}
f &= (c_1+c_2+c_3)f && \to \quad 1 = c_1+c_2+c_3\\
\frac{1}{2}y'' &= (c_2b_{21}+c_3a_3)F_1f && \to \quad \frac{1}{2} = c_2b_{21}+c_3a_3\\
\frac{1}{6}y''' &= \frac{1}{2}(c_2b_{21}^2+c_3a_3^2)F_2f + c_3b_{21}b_{32}f_yF_1f && \to \quad \frac{1}{6} = \frac{1}{2}(c_2b_{21}^2+c_3a_3^2)\\
& && \to \quad \frac{1}{6} = c_3b_{21}b_{32}\\
& && \quad\ a_2 = b_{21}\\
& && \quad\ a_3 = b_{31}+b_{32}
\end{aligned}$$

Emlékezzünk rá, hogy $y''' = F_2f + f_yF_1f$. Ne felejtsük el, hogy az utolsó két egyenletet a levezetés során már használtuk, így a megoldandó nemlineáris egyenletrendszerünk 6 egyenletből áll, amit redukálhatunk 4-re, az ismeretlenek pedig c_1, c_2, c_3, a_2, a_3 és b_{32}.

$$\begin{aligned}
c_1+c_2+c_3 &= 1\\
c_2a_2+c_3a_3 &= \frac{1}{2}\\
c_2a_2^2+c_3a_3^2 &= \frac{1}{3}\\
c_3a_2b_{32} &= \frac{1}{6}
\end{aligned}$$

Mivel 4 egyenletünk van és 6 ismeretlenünk, végtelen sok harmadrendű RK-módszert találunk.

4-1. Példa. Készítsünk (valós) harmadrendű explicit RK-módszert, ha $c_1 = c_3 = \frac{1}{6}$ és $c_2 = \frac{4}{6}$. Adjuk meg az a_j és b_{jl} számokat.

Megoldás. A harmadrendű explicit RK-módszer képleteibe behelyettesítve a következő nemlineáris egyenletrendszert kapjuk.

$$\begin{aligned}
\frac{4}{6}a_2+\frac{1}{6}a_3 &= \frac{1}{2} && \to \quad 4a_2+a_3 = 3\\
\frac{4}{6}a_2^2+\frac{1}{6}a_3^2 &= \frac{1}{3} && \to \quad 4a_2^2+a_3^2 = 2\\
\frac{1}{6}a_2b_{32} &= \frac{1}{6} && \to \quad a_2b_{32} = 1
\end{aligned}$$

Az első egyenletből $a_3 = 3-4a_2$, ezt a második egyenletbe helyettesítve

$$4a_2^2+(3-4a_2)^2 = 20a_2^2-24a_2+9 = 2 \quad \to \quad 20a_2^2-24a_2+7 = 0.$$

Ennek a megoldásai: $a_2 = \frac{1}{2}$ és $a_2 = \frac{7}{10}$, ekkor az a_3 értékek rendre 1 és $\frac{2}{10}$.
Az általános alak és a két konkrét Butcher-táblázat:

0	0	0	0
a_2	a_2	0	0
a_3	$a_3-\frac{1}{a_2}$	$\frac{1}{a_2}$	0
	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$

0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0
1	-1	2	0
	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$

0	0	0	0
$\frac{7}{10}$	$\frac{7}{10}$	0	0
$\frac{2}{10}$	$-\frac{43}{35}$	$\frac{10}{7}$	0
	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$

Az középső táblázat a Simpson-formulának megfelelő RK-képlet. ■

4-2. Példa. Határozzuk meg az összes olyan (valós) harmadrendű explicit RK-módszert, melyre $b_{31}+b_{32} = d > 0$, $c_1 = c_3 = c$, $c_2 = 1-2c$, ahol $c = c(d)$ és d szabad paraméter. Rajzoljuk ki Maple-lel a $c(d)$ függvényt és adjunk példának RK-táblázatokat.

Megoldás. A harmadrendű explicit RK-módszer képleteibe behelyettesítve a következő nemlineáris egyenletrendszert kapjuk. Vezessük még be a következő jelöléseket: $a_3 = d$ és $b_{32} = b$.

$$\begin{aligned}
(1-2c)a_2 + cd = \frac{1}{2} \quad &\rightarrow \quad a_2 = \frac{1-2cd}{2(1-2c)} \\
(1-2c)a_2^2 + cd^2 = \frac{1}{3} \quad &\rightarrow \quad (1-2c)\left(\frac{1-2cd}{2(1-2c)}\right)^2 + cd^2 = \frac{1}{3} \\
ca_2 b = \frac{1}{6} \quad &\rightarrow \quad b = \frac{1-2c}{3c(1-2cd)}
\end{aligned}$$

Egyszerűsítsük a 2. egyenletet:

$$\begin{aligned}
\frac{(1-2cd)^2}{4(1-2c)} + cd^2 &= \frac{1}{3} \\
\frac{3(1-2cd)^2}{(1-2c)} &= 4(1-3cd^2) \\
3(1-2cd)^2 &= 4(1-3cd^2)(1-2c) \\
3 - 12cd + 12c^2d^2 &= 4 - 12cd^2 - 8c + 24c^2d^2
\end{aligned}$$

A kapott egyenletet 0-ra rendezve c hatványai szerint:

$$(12d^2)c^2 + c(-12d^2 + 12d - 8) + 1 = 0.$$

A kapott gyököket érdemes Maple-lel megjeleníteni d függvényében. Ügyeljünk rá, hogy $d > 0$ legyen. Például $d = 1$ esetén a két szép megoldást kapunk c-re: $-\frac{1}{6}$ és $-\frac{1}{2}$. Ekkor az általános alak és a két Butcher-táblázat:

0	0	0	0
a_2	a_2	0	0
d	$d-b$	b	0
	c	$1-2c$	c

0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0
1	1	-2	0
	$-\frac{1}{6}$	$\frac{4}{3}$	$-\frac{1}{6}$

0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0
1	$\frac{5}{3}$	$-\frac{2}{3}$	0
	$-\frac{1}{2}$	2	$-\frac{1}{2}$

Például $d = \frac{2}{3}$ esetén szép megoldások c-re: $-\frac{1}{4}$ és $-\frac{3}{4}$. ■

4-3. Példa. Készítsünk (valós) harmadrendű explicit RK-módszert, amelyben $b_{31} = c_1 = 0$. Oldjuk meg Maple-lel a kapott egyenletrendszert.

Megoldás. A harmadrendű explicit RK-módszer képleteibe behelyettesítve a következő nemlineáris egyenletrendszert kapjuk.

$$\begin{aligned}
c_2 + c_3 &= 1 \\
c_2 a_2 + c_3 a_3 &= \frac{1}{2} \\
c_2 a_2^2 + c_3 a_3^2 &= \frac{1}{3} \\
c_3 a_2 a_3 &= \frac{1}{6}
\end{aligned}$$

Ennek megoldása a Maple jelölésrendszerével:

$$\begin{aligned}
a2 &= (1/6) * y + 1/(6 * y) + 1/2, \\
a3 &= (1/2) * y + 1/(2 * y) + 3/2 - 3 * ((1/6) * y + 1/(6 * y) + 1/2)^2, \\
c2 &= 15/8 + (7/8) * y + 7/(8 * y) - (9/2) * ((1/6) * y + 1/(6 * y) + 1/2)^2, \\
c3 &= -7/8 - (7/8) * y - 7/(8 * y) + (9/2) * ((1/6) * y + 1/(6 * y) + 1/2)^2 \\
y &= (3 + 2 * \sqrt{2})^{1/3};
\end{aligned}$$

Tizedestörttel:

$$a2 = 0.8925502329,\ \ a3 = 0.2877129439,\ \ c2 = 0.3509820905,\ \ c3 = 0.6490179095.$$

■

Feladatok

4-1. Készítsünk $s = 3$ lépcsőszámú $p = 2$ rendű ERK-módszert! Írjuk fel a konstrukcióhoz szükséges nemlineáris egyenletrendszert.

4-2. Készítsünk $s = 2$ lépcsőszámú $p = 2$ rendű ERK-módszert! Írjuk fel a konstrukcióhoz szükséges nemlineáris egyenletrendszert.

4-3. Vezessük le a következő a_2 értékekhez tartozó $s = 2$ lépcsőszámú másodrendű RK-módszereket:

a) $a_2 = \frac{2}{3}$,
b) $a_2 = \frac{1}{2}$,
c) $a_2 = 1$.

Megoldások:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{2}{3} & \frac{2}{3} & 0 \\ \hline & \frac{1}{4} & \frac{1}{4} \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

4.3. Az ERK-módszerek stabilitása és konvergenciája

A következő két tétellel az összes explicit RK-módszer stabilitását és konvergenciáját bizonyítjuk. Látjuk majd, hogy f Lipschitz-folytonosságából következnek ezek a tulajdonságok. Vannak speciális esetek, amikor a Lipschitz-folytonosság hiánya esetén is - megfelelő módszerekkel - elfogadható eredményre jutunk. Lásd [10] 7. feladatát.

4-1. T *(Általános ERK-módszer stabilitása)*
Legyen $f \in Lip(y)$, L_f Lipschitz-állandóval, ekkor az explicit RK-módszer stabil és

$$|e_i| \le e^{L_\Phi} \left(|e_0| + \sum_{k=0}^{i-1} |g_k| h \right),$$

ahol

$$L_\Phi = L_f \left(\sum_{j=1}^{s} |c_j| + Q_{s-1}(hL_f) \right),$$

és a $Q_{s-1}(hL_f)$ a hL_f mennyiség $(s-1)$-edfokú polinomja, melyre $Q_{s-1}(hL_f) = O(hL_f)$ (vagyis nincs konstans tagja).

Bizonyítás. Legyen

$$\Phi(x, y, h) := \sum_{j=1}^{s} c_j k_j, \quad k_j = k_j(x, y, h).$$

Ekkor a lokális hiba

$$hg_i = y(x_{i+1}) - y(x_i) - h\Phi\left(x_i, y(x_i), h\right) \quad \rightarrow \quad y(x_{i+1}) = y(x_i) + h\Phi\left(x_i, y(x_i), h\right) + hg_i.$$
$$y_{i+1} = y_i + h\Phi(x_i, y_i, h),$$

a módszer képlete, tehát a globális hiba

$$e_{i+1} = y(x_{i+1}) - y_{i+1} = e_i + h[\Phi\left(x_i, y(x_i), h\right) - \Phi(x_i, y_i, h)] + g_i h.$$

A 38.-40. feladatokból és k_j, Φ rekurzív definíciójából is következik, hogy $k_j, \Phi \in Lip(y)$. Nézzük meg, mi lesz a Lipschitz-konstans.
$j = 1$ -re

$$|k_1(x,u,h) - k_1(x,v,h)| \le L_f|u-v| \quad \to \quad L_{k_1} = L_f.$$

$j = 2$ -re

$$\begin{aligned}|k_2(x,u,h) - k_2(x,v,h)| &= |f(x+ha_2, u+hb_{21}k_1(x,u,h)) - f(x+ha_2, v+hb_{21}k_1(x,v,h))| \le \\ &\le L_f|u-v+hb_{21}(k_1(x,u,h) - k_1(x,v,h))| \le \\ &\le L_f\left(|u-v| + h|b_{21}|L_f|u-v|\right) \le L_f(1+h|b_{21}|L_f)|u-v| \\ \to \quad L_{k_2} &= L_f(1+h|b_{21}|L_f) = L_f(1+Q_1(hL_f)),\end{aligned}$$

ahol $Q_1(hL_f) = h|b_{21}|L_f$, azaz (hL_f) elsőfokú polinomja. Tegyük fel, kogy $k_1, \ldots, k_{j-1}$ Lipschitz-konstansát ismerjük. Ekkor k_j-re

$$k_j(x,u,h) = f\left(x + ha_j, u + h\sum_{l=1}^{j-1} b_{jl}k_l(x,u,h)\right)$$

a következő becslést kapjuk:

$$\begin{aligned}|k_j(x,u,h) - k_j(x,v,h)| &\le L_f|(u-v) + h\sum_{l=1}^{j-1} b_{jl}(k_l(x,u,h) - k_l(x,v,h))| \le \\ &\le L_f\left(|u-v| + h\sum_{l=1}^{j-1}|b_{jl}||k_l(x,u,h) - k_l(x,v,h)|\right) \le \\ &\le L_f\left(1 + h\sum_{l=1}^{j-1}|b_{jl}|L_{k_l}\right)|u-v| \\ \to \quad L_{k_j} &= L_f\left(1 + h\sum_{l=1}^{j-1}|b_{jl}|L_f(1+Q_{l-1}(hL_f))\right) = L_f\left(1 + Q_{j-1}(hL_f)\right),\end{aligned}$$

ahol $Q_{j-1}(hL_f)$ a (hL_f) $j-1$-edfokú polinomja.

$$|\Phi(x,u,h) - \Phi(x,v,h)| \le \sum_{j=1}^{s}|c_j||k_j(x,u,h) - k_j(x,v,h)| \le \sum_{j=1}^{s}|c_j|L_{k_j}|u-v|$$

Tehát Φ Lipschitz-állandója

$$\begin{aligned}L_\Phi &= L_f\left(\sum_{j=1}^{s}|c_j|\left(1 + h\sum_{l=1}^{j-1}|b_{jl}|L_{k_l}\right)\right) = L_f\left(\sum_{j=1}^{s}|c_j| + h\sum_{j=1}^{s}|c_j|\sum_{l=1}^{j-1}|b_{jl}|L_{k_l}\right) = \\ &= L_f\left(\sum_{j=1}^{s}|c_j| + \sum_{j=1}^{s}|c_j|\sum_{l=1}^{j-1}|b_{jl}|hL_f\left(1+Q_{l-1}(hL_f)\right)\right) = L_f\left(\sum_{j=1}^{s}|c_j| + Q_{s-1}(hL_f)\right),\end{aligned}$$

ahol $Q_{s-1}(hL_f)$ a (hL_f) $s-1$-edfokú polinomja. A Lipschitz-állandó birtokában a bizonyítás ugyanúgy megy, mint a 3. tételnél. ■

4-2. T *(Konzisztencia + Stabilitás = Konvergencia)*
A kezdetiérték-probléma megoldására tekintsük az általános egylépéses módszert

$$y(x_0) = y_0$$
$$y_{i+1} = y_i + h \cdot \Phi(x_i, y_i; h), \quad (i = 0, \ldots, N-1)$$

mely p-edrendben konzisztens ($p \geq 1$), azaz létezik olyan $K > 0$, hogy

$$|g_i| \leq Kh^p \quad (i = 0, \ldots, N-1)$$

és stabil, azaz

$$|e_i| \leq e^{L_\Phi} \left[|e_0| + \sum_{k=0}^{i-1} |g_k| h \right].$$

Ekkor

$$|e_i| \leq e^{L_\Phi} K h^p \quad (i = 0, \ldots, N),$$

vagyis p-edrendben konvergens a numerikus módszer.

Bizonyítás. A módszer Φ Lipschitz-állandójának birtokában a bizonyítás ugyanaz, mint a 4. tételnél. ■

4.4. Ismert ERK-módszerek

Kétlépcsős másodrendű módszerek:

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

RK2 javított Euler, $p = s = 2$ — RK2 általánosított trapéz formula, $p = s = 2$

Háromlépcsős harmadrendű ERK-módszerek:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{2}{3} & 0 & \frac{2}{3} & 0 \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{4}{6} \end{array}$$

RK3 Heun, $p = s = 3$ — RK3 (klasszikus) Simpson, $p = s = 3$ — RK3, $p = s = 3$

Négylépcsős negyedrendű ERK-módszerek:

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{2}{3} & -\frac{1}{3} & 1 & 0 & 0 \\ 1 & 1 & -1 & 1 & 0 \\ \hline & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 1 & 0 & -1 & 2 & 0 \\ \hline & \frac{1}{6} & 0 & \frac{4}{6} & \frac{1}{6} \end{array}$$

RK4 klasszikus 4-edrendű, $p = s = 4$ — RK4 Newton (3/8), $p = s = 4$ — RK4, $p = s = 4$

Vegyük észre, hogy a Heun- és a klasszikus negyedrendű RK-módszer azért is szép, mert az egyes k_j-k kiértékeléséhez csak egyetlen előző k_i-re van szükségünk.

4.5. Összefüggések a lépcsőszám és a rend között

A következő összefüggések ismertek az s lépcsőszám és a $p(s)$ elérhető rend között (az eredmény Butcher-től származik):

$$\begin{aligned} 1 \le s \le 4 &: p(s) = s, \\ s \ge 5 &: p(s) \le s - 1, \\ s \ge 8 &: p(s) \le s - 2, \\ s \ge 10 &: p(s) \le s - 3. \end{aligned}$$

A módszer konstrukciója során nemlineáris egyenletrendszert kell megoldani az a_j, b_{jl} és c_j számokra, melyek függetlenek f-től és h-tól. Ebben az egyenletek száma mindig nagyobb a rendnél. A következő táblázat mutatja, hogy a nemlineáris egyenletek száma hogyan alakul a rend függvényében aszerint, hogy skaláris differenciálegyenletről vagy rendszerről van-e szó.

rend	skaláris	rendszer
4	8	8
6	31	37
8	110	200
10	361	1205
12	1114	7813
14	3259	53272

4.6. Beágyazott módszerek

A közelítő eredmények hibáját csak ritkán tudjuk becsülni, általában csak a konvergenciarendről kapunk információt. Ezért fontos, hogy a számítás során, a numerikus eredmények birtokában azonnal (aposteriori) hibabecslést is tudjunk készíteni. Ennek módját a hibabecslések alfejezetben fogjuk részletesebben tárgyalni. Ehhez szükségünk lesz egy összehasonlítási értékre, melynek előállításához egy beágyazott módszert használunk. Ez azt jelenti, hogy egyszerre használunk két olyan ERK-módszert, melyek $p-1$ és p-edrendűek és a $p-1$-edrendű módszer (ezt nevezzük beágyazott módszernek) ugyanazokat a k_j-ket használja, mint a másik módszer, de a c_j együtthatók már különbözők. Az ötlet Fehlberg-től származik, ezért Runge-Kutta-Fehlberg módszernek is nevezik ezeket a módszereket. A kétféle módszer súlyait $(c_j)_{j=1}^s$ (p-edrendű) illetve $(\widetilde{c}_j)_{j=1}^s$-mal ($p-1$-edrendű) jelölve az $\widetilde{y}_{i+1} - y_{i+1}$ mennyiségből a hiba nagyságára fogunk majd következtetni.
Például az RK3 klasszikus harmadrendű és a javított Euler-módszer ilyen:

0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0
	0	1

RK2 javított Euler, $p = s = 2$

0	0	0	0
1/2	1/2	0	0
.....			
1	−1	2	0
	1/6	4/6	4/6

RK3 (klasszikus) Simpson-RK, $p = s = 3$

0	0	0	0
1/2	1/2	0	0
1	−1	2	0
	1/6	4/6	4/6
	0	1	0

Együtt a két táblázat

Az RK4 klasszikus negyedrendű RK-módszerhez nincs beágyazott harmadrendű módszer. (Lásd [10] 44. oldalán.) Ahhoz, hogy a táblázat olvashatóbb legyen a törteket nem a hagyományos formában, hanem számláló/nevező alakban fogjuk megadni.

A Matlab által használt ode23 azaz **Bogacki-Shampine** 2(3)-rendű módszer táblázata:

0	0	0	0	0
1/2	1/2	0	0	0
3/4	0	3/4	0	0
1	2/9	1/3	4/9	0
$c^{(3)}$	2/9	1/3	4/9	0
$\widetilde{c}^{(2)}$	7/24	1/4	1/3	1/8

A **Cash-Karp** 4(5)-rendű módszer táblázata:

0	0	0	0	0	0	0
1/5	1/5	0	0	0	0	0
3/10	3/40	9/40	0	0	0	0
3/5	3/10	−9/10	6/5	0	0	0
1	−11/54	5/2	−70/27	35/27	0	0
7/8	1631/55296	175/512	575/13824	44275/110592	253/4096	0
$c^{(5)}$	37/378	0	250/621	125/594	0	512/1771
$\widetilde{c}^{(4)}$	2825/27648	0	18575/48384	13525/55296	277/14336	1/4

A **Fehlberg** 4(5)-rendű módszer táblázata:

0	0	0	0	0	0	0
1/4	1/4	0	0	0	0	0
3/8	3/32	9/32	0	0	0	0
12/13	1932/2197	−7200/2197	7296/2197	0	0	0
1	439/216	−8	3680/513	−845/513	0	0
1/2	−8/27	2	−3544/2565	1859/4104	−11/40	0
$c^{(5)}$	16/135	0	6656/12825	28561/56430	−9/50	2/55
$\widetilde{c}^{(4)}$	25/216	0	1408/2565	2197/4104	−1/5	0

Az egyik legtöbbet használt beágyazott módszer **Dormand és Prince**-től származó 4(5)-rendű módszer, a Matlab ode45 programja is ezt használja, együtthatóit a következő Butcher-táblázattal adjuk meg:

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

Itt $c^{(5)}$ az ötödrendű módszer és $\widetilde{c}^{(4)}$ a negyedrendű módszer együtthatóit jelöli. Lépcsőszáma 7, de csak 6 függvénykiértékelésre van szükség lépésenként. Mivel az együtthatókra $c_l^{(5)} = b_{7l}$,

$l = 1, \ldots, 6$ és $c_7^{(5)} = 0$, $k_7^{(i)} = k_1^{(i+1)}$, így ezt nem kell újra számolni. A k_7-et csak a hibabecsléshez szükséges $\widetilde{y}_{i+1}^{(4)}$-hez számoljuk ki. Bár itt 7 szintünk van, de a következő lépés első kiértékelését ezzel már megkaptuk. Dormand és Prince úgy határozták meg a Butcher mátrix együtthatóit, hogy az ötödrendű közelítés hibája legyen kicsi, míg a Fehlberg módszer a negyedrendű hibáját minimalizálja. Ez a két módszer közti fő különbség (lokális extrapolációt tartalmaz), emiatt használják gyakrabban Dormand és Prince módszerét.

4.7. Az ERK-módszerek hatékonysága

4-1. Definíció. Jelöljük $Q(\varepsilon)$-nal azt a műveletigényt, amely az adott ε pontosság eléréséhez szükséges. Egy módszer η abszolút hatékonyságát a következőképpen definiáljuk:

$$\eta := \frac{1}{\varepsilon Q(\varepsilon)}.$$

A hatékonyság fordított arányban áll mind a pontossággal, mind a műveletigénnyel. Látjuk, hogy ha kicsi műveletigénnyel kapunk pontos megoldást, akkor nagy a módszer hatékonysága.

4-2. Definíció. Két módszer relatív hatékonyságát az ugyanazon ε pontosság eléréséhez szükséges műveletigények hányadosával definiáljuk:

$$\eta_{1/2} := \frac{\eta_2}{\eta_1} = \frac{Q_1(\varepsilon)}{Q_2(\varepsilon)}.$$

Az $\eta_{1/2}$ szám adja meg, hogy hányszor több műveletet kell végrehajtani az első módszerrel, mint a másodikkal ugyanahhoz a pontossághoz. Az explicit módszerek osztályában az Euler-módszerhez szokás hasonlítani, mert az nem csak RK-módszer, hanem többlépéses módszer is.

4-4. Példa. Határozzuk meg az s lépcsőszámú és p-edrendű ERK-módszer $Q_{s,p}(\varepsilon)$ műveletigényét!

Megoldás. Ha a $[0;1]$ intervallumon h lépéssel megyünk végig, akkor minden lépésben s függvénykiértékelésre van szükségünk. Ezek adják a lényeges műveletigényt. Ha a módszer p-edrendű, akkor $\varepsilon = O(h^p)$ a pontosság, amit elérünk vele. Az egyszerűség kedvéért $h^p = \varepsilon$-nal számolunk (bár ezzel nem tudunk az azonos rendű módszerek közt különbséget tenni). Ekkor $h(\varepsilon) = \varepsilon^{1/p}$, így

$$Q_{s,p}(\varepsilon) = s \cdot N = \frac{s}{h(\varepsilon)} = \frac{s}{\varepsilon^{1/p}}.$$

Az Euler-módszer elsőrendű és műveletigénye lépésenként egy függvény kiértékelés, így

$$Q_{1,1}(\varepsilon) = \frac{1}{h} = \frac{1}{\varepsilon}.$$

Ebből az abszolút hatékonyság az Euler-módszerhez hasonlítva

$$\eta_{1,1/s,p} = \frac{\eta_{s,p}}{\eta_{1,1}} = \frac{Q_{1,1}(\varepsilon)}{Q_{s,p}(\varepsilon)} = \frac{\frac{1}{\varepsilon}}{\frac{s}{\varepsilon^{1/p}}} = \frac{1}{s \cdot \varepsilon^{1-1/p}} = \frac{1}{s \cdot \varepsilon^{\frac{p-1}{p}}}.$$

■

A különböző ERK-módszerek hatékonysága $\varepsilon = 10^{-4}$ esetén:

s	1	2	3	4	5	6	7	8	9	10	17
p	1	2	3	4	4	5	6	6	7	7	10
η	1	50	154.7	250	200	264.1	307.8	269.3	289.1	268.3	234.2

Látjuk, hogy a 7 lépcsőszámú 6-odrendű ERK-módszer esetén a legnagyobb a hatékonyság.
A relatív hatékonyság definíciója alapján két tetszőleges ERK-módszer hatékonysága összehasonlítható közvetlenül a műveletigényekből. Mivel $Q_{s,p}(\varepsilon) = s \cdot N = \frac{s}{h(\varepsilon)}$, így a következő képlettel értékelhető ki:

$$\eta_{1/2} = \frac{\eta_{s_2,p_2}}{\eta_{s_1,p_1}} = \frac{Q_{s_1,p_1}(\varepsilon)}{Q_{s_2,p_2}(\varepsilon)} = \frac{s_1}{s_2} \cdot \frac{h_2(\varepsilon)}{h_1(\varepsilon)}.$$

4.8. Hibabecslések

A stabilitási tételekben szereplő hibabecslés helyett másik is adható, melyben nincs szükség a Lipschitz-folytonosságra. L_f helyett a f_y Jacobi-mátrix logaritmikus normájának felső becslése szükséges. Nem a pontos megoldáshoz tartozó lokális hibákat adjuk össze, hanem a számított $\mathbf{y_i}$ értékekből kiinduló megoldásokhoz tartozó lokális hibákat.

4-3. T *Tegyük fel, hogy $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$ folytonosan differenciálható, $x_i \in [0;1]$ és $m \geq \mu(f_y)$, ekkor a globális hiba becslése*

$$\|\mathbf{e_i}\| \leq C_p h^p \cdot \begin{cases} \frac{1}{m}(e^{x_i m} - 1), & m > 0 \\ x_i, & m \leq 0. \end{cases}$$

Bizonyítás. Legyen $h_j := x_j - x_{j-1} > 0$ lépéstávolság $(j < i)$, $h := \max_{j=1}^{N}(h_j)$ és $x_0 := 0$.
Legyen $\mathbf{y_j}$ az ERK-módszerrel kapott közelítő érték és $\mathbf{y_j}(x)$ a differenciálegyenlet azon megoldása, melyre

$$\mathbf{y_j}(x_j) = \mathbf{y_j}.$$

Ekkor nyilván $\mathbf{y_0}(x) = \mathbf{y}(x)$ a pontos megoldás és

$$\mathbf{y}(x_i) - \mathbf{y_i} = \sum_{j=1}^{i} (\mathbf{y_{j-1}}(x_i) - \mathbf{y_j}(x_i)).$$

Mivel $\mathbf{y_{j-1}}(x)$ és $\mathbf{y_j}(x)$ a differenciálegyenlet megoldása

$$\mathbf{y'_{j-1}}(x) - \mathbf{y'_j}(x) = \mathbf{f}(x, \mathbf{y_{j-1}}(x)) - \mathbf{f}(x, \mathbf{y_j}(x)).$$

A rendszerekre vonatkozó disszipativitási 34. tétel alapján

$$\|\mathbf{y_{j-1}}(x_i) - \mathbf{y_j}(x_i)\| \leq e^{(x_i - x_j)m} \|\mathbf{y_{j-1}}(x_j) - \mathbf{y_j}(x_j)\|.$$

Vegyük figyelembe, hogy $\mathbf{y_j}(x_j) = \mathbf{y_j}$ a közelítő érték x_j-ben és $\mathbf{y_{j-1}}(x_j)$ a pontos értéke annak a megoldásnak, mely $x = x_{j-1}$-ben $\mathbf{y_{j-1}}$-ből indult ki, ezért

$$\mathbf{y_{j-1}}(x_j) - \mathbf{y_j}(x_j) = h_j \mathbf{g_{j-1}}.$$

Teljesüljön a lokális hibákra

$$h_j \|\mathbf{g_{j-1}}\| \leq C_p h_j^{p+1} \leq C_p h^p h_j.$$

Vegyük elő a korábban felírt globális hibát és becsüljük.

$$\|\mathbf{e_i}\| = \|\mathbf{y}(x_i) - \mathbf{y_i}\| = \sum_{j=1}^{i} e^{(x_i - x_j)m} \|\mathbf{y_{j-1}}(x_j) - \mathbf{y_j}(x_j)\| =$$

$$= \sum_{j=1}^{i} e^{(x_i - x_j)m} \|\mathbf{g_{j-1}}\| h_j \leq C_p h^p \sum_{j=1}^{i} e^{(x_i - x_j)m} h_j.$$

Ha $m \leq 0$ (azaz a rendszer disszipatív), akkor $e^{(x_i - x_j)m} \leq 1$ és így

$$\|\mathbf{e_i}\| \leq C_p h^p x_i.$$

Ha $m > 0$, akkor

$$\sum_{j=1}^{i} e^{(x_i - x_j)m} h_j \leq \int_0^{x_i} e^{(x_i - x)m}\, dx = \left[-\frac{1}{m} e^{(x_i - x)m} \right]_0^{x_i} = \frac{1}{m}(e^{x_i m} - 1)$$

$$\Rightarrow \quad \|\mathbf{e_i}\| \leq \frac{C_p}{m} h^p (e^{x_i m} - 1).$$

■

Ez a becslés a logaritmikus norma megjelenése miatt már élethűbb, de gyakorlatilag még mindig használhatatlan, mert a Jacobi-mátrix logaritmikus normájának felső becslésével (m) kéne rendelkezni, valamint a lokális megoldások lokális hibájának becslésével ($C_p h_j^p$). A C_p konstans tartalmazza az y megoldás $p+1$-edik deriváltjának normáját illetve az f deriváltjainak normáját a p-edrendűekig, amit ritkán tudunk beszerezni. De legalább a pontos megoldás lokális hibájától megszabadultunk. Ezután a következőképpen járhatunk el a gyakorlatilag használható globális hibabecslés érdekében:

- Az aktuális számítási eredményekből (kiegészítő számítások segítségével) kapjuk a lokális hiba $\widetilde{\mathbf{g}}_{\mathbf{j-1}}$ becslését.
- A becsült lokális hibákat összeadjuk ($\widetilde{\mathbf{g}}_{\mathbf{j-1}}$-ból δ_i-ket képezzük).
- A logaritmikus norma becsléséről lemondunk.

A továbbiakban a globális hiba becslése helyett beérjük annak nagyságrendi jellemzésével. Erre a δ_i hibaindikátort használjuk.

$$\delta_i := \sum_{j=1}^{i} \|\widetilde{\mathbf{g}}_{\mathbf{j-1}}\| h_j$$

Ez h^p nagyságrendű lesz (ugyanúgy, mint $\|\mathbf{e_i}\|$), ha $\widetilde{\mathbf{g}}_{\mathbf{j}}$-t megfelelően számítjuk ki. Ezzel foglalkozunk a továbbiakban, többféle módszert is mutatva.

a) Beágyazott módszereknél minden lépésben két értéket számítunk.
$\mathbf{y_j^{(p)}}$ a p-edrendű, $\mathbf{y_j^{(p-1)}}$ a $p-1$-edrendű módszerrel kapott eredmény, a lokális hibák

$$h_j \mathbf{g_{j-1}^{(p-1)}} = \mathbf{y_{j-1}}(x_j) - \mathbf{y_j^{(p-1)}} = C^{(p-1)} h_j^p + O(h_j^{p+1}),$$
$$h_j \mathbf{g_{j-1}^{(p)}} = \mathbf{y_{j-1}}(x_j) - \mathbf{y_j^{(p)}} = C^{(p)} h_j^{p+1} + O(h_j^{p+2}).$$

Mivel az $\mathbf{y_j}(x)$ megoldás ismeretlen készítsük el a tapasztalati lokális hibát

$$h_j \widetilde{\mathbf{g}}_{\mathbf{j-1}} := \mathbf{y_j^{(p)}} - \mathbf{y_j^{(p-1)}}$$

A fenti két egyenletet egymásból kivonva

$$h_j\tilde{\mathbf{g}}_{\mathbf{j-1}} = h_j\mathbf{g}_{\mathbf{j-1}}^{(\mathbf{p-1})} - h_j\mathbf{g}_{\mathbf{j-1}}^{(\mathbf{p})} = h_j\mathbf{g}_{\mathbf{j-1}}^{(\mathbf{p-1})} + O(h_j^{p+1}),$$

amiből látszik, hogy elég kicsi h_j-re elég jól becsüli a $p-1$-edrendű módszer lokális hibáját. Ha $\|\tilde{\mathbf{g}}_{\mathbf{j-1}}\|$ elég kicsi, akkor elfogadjuk az $\mathbf{y}_{\mathbf{j}}^{(\mathbf{p-1})}$ értéket és a globális hibaindikátor újabb értékét kiszámítjuk:

$$\delta_j = \delta_{j-1} + \|\tilde{\mathbf{g}}_{\mathbf{j-1}}\|h_j,$$

és következhet a továbblépés.

b) Ha nincs beágyazott módszer, akkor Runge ötlete nyomán intervallumfelezéssel dolgozunk. A szokásos h_j lépéssel számított értékhez úgy szerzünk jobb közelítést, hogy két lépést teszünk $h_j/2$ lépésközzel, így előállítva az $\widetilde{y}_{j-1/2}$ és $\widetilde{y}_j$ értékeket. A továbbiakban hullámmal jelöljük a $h_j/2$ lépésközzel kapott eredményeket. $\widetilde{y}_j$-t a $2j$. lépés után kapjuk, azt hasonlítjuk y_j-vel.

4-4. T *Ha f $p+1$-szer differenciálható, akkor*

$$\widetilde{y}_j - y_j = (1-2^{-p})h_j g_{j-1} + O(h_j^{p+2})$$

és innen a g_{j-1} lokális hiba becslése

$$\widetilde{g}_{j-1} := \frac{\widetilde{y}_j - y_j}{h_j(1-2^{-p})} = g_{j-1} + O(h_j^{p+1}).$$

Alkalmazhatjuk a Richardson-extrapolációt (lásd Romberg integrálás), mellyel $p+1$-edrendű közelítést kapunk:

$$y_* := \frac{\widetilde{y}_j - 2^{-p}y_j}{1-2^{-p}}.$$

Ennek birtokában a lokális hiba becslése

$$\widetilde{g}_{j-1} = \frac{y_* - y_j}{h_j}$$

alakban is felírható.

Bizonyítás. Lásd [10] 54.oldal.

$$\widetilde{g}_{j-1} = \frac{y_* - y_j}{h_j} = \frac{\frac{\widetilde{y}_j - 2^{-p}y_j}{1-2^{-p}} - y_j}{h_j} =$$
$$= \frac{\widetilde{y}_j - 2^{-p}y_j - y_j + 2^{-p}y_j}{h_j(1-2^{-p})} = \frac{\widetilde{y}_j - y_j}{h_j(1-2^{-p})}$$

■

c) A globális hiba becslése érdekében úgy is eljárhatunk, hogy **a számítási intervallumon többször végigintegrálunk különböző h lépésközökkel**, majd a kapott eredményeket összehasonlítva nyerjük a becslést. A globális hiba becslése lehetőséget ad a lépéshossz megválasztására.

4-5. T *Ha f $p+1$-szer differenciálható és az x_* rögzített helyig h_1 és $h_2 \neq h_1$ lépésközzel integrálva a differenciálegyenletet (Richardson-extrapolációt alkalmazva)*

$$y_* := \frac{h_2^p y_*^{(h_1)} - h_1^p y_*^{(h_2)}}{h_2^p - h_1^p}$$

$p+1$-edrendű közelítést kapunk. Elegendően kicsi h_1, h_2 birtokában a globális hiba becslését kapjuk:

$$y_* - y_*^{(h_2)} = \frac{h_2^p}{h_2^p - h_1^p}(y_*^{(h_1)} - y_*^{(h_2)}) = -e_p(x_*)h_2^p + O(h_2^{p+1}),$$

illetve

$$y_* - y_*^{(h_1)} = \frac{h_1^p}{h_2^p - h_1^p}(y_*^{(h_1)} - y_*^{(h_2)}) = -e_p(x_*)h_1^p + O(h_1^{p+1}).$$

Bizonyítás. Lásd [10] 56. oldal. ■

Ha speciálisan $h_2 = h$ és $h_1 = \frac{h}{2}$, akkor a fenti extrapolációs képletből a **b)** részben adott

$$y_* := \frac{h^p y_*^{(h/2)} - \left(\frac{h}{2}\right)^p y_*^{(h)}}{h^p - \left(\frac{h}{2}\right)^p} = \frac{h^p(y_*^{(h/2)} - 2^{-p}y_*^{(h)})}{h^p(1-2^{-p})} = \frac{y_*^{(h/2)} - 2^{-p}y_*^{(h)}}{1-2^{-p}}$$

képletet kapjuk.

Megjegyzések.

A fenti aszimptotikus képletek formálisan a következő LER megoldással is megkaphatók. Jelölje φ a pontos értéket, amit közelíteni szeretnénk. A φ_1 illetve φ_2 a h_1 illetve h_2 lépésközzel ugyazzal a módszerrel számított értékek. p-edrendű módszer esetén jelölje c_p a hibakonstanst, amire szintén közelítést szeretnénk adni.

$$\varphi - \varphi_1 \approx h_1^p c_p$$
$$\varphi - \varphi_2 \approx h_2^p c_p$$

Írjuk fel mátrixos alakban is.

$$\begin{bmatrix} 1 & -h_1^p \\ 1 & -h_2^p \end{bmatrix} \cdot \begin{bmatrix} \varphi \\ c_p \end{bmatrix} \approx \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix}$$

A mátrix inverzét felírva

$$\begin{bmatrix} \varphi \\ c_p \end{bmatrix} \approx \frac{1}{h_1^p - h_2^p} \begin{bmatrix} -h_2^p & h_1^p \\ -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix}$$

Innen φ-re a Richardson-extrapoláció képletét, c_p-re pedig a Runge-féle hibaképletet kapjuk:

$$\varphi \approx \frac{h_1^p\varphi_2 - h_2^p\varphi_1}{h_1^p - h_2^p} = \frac{(\frac{h_1}{h_2})^p\varphi_2 - \varphi_1}{(\frac{h_1}{h_2})^p - 1} = \varphi_2 + \frac{\varphi_2 - \varphi_1}{(\frac{h_1}{h_2})^p - 1}$$
$$c_p \approx \frac{\varphi_2 - \varphi_1}{h_1^p - h_2^p}.$$

1. Nézzük most a $h_1 = h$ és $h_2 = \frac{h}{2}$ speciális esetet.

$$\varphi \approx \frac{h^p\varphi_2 - \frac{h^p}{2^p}\varphi_1}{h^p - \frac{h^p}{2^p}} = \frac{\varphi_2 - 2^{-p}\varphi_1}{1-2^{-p}} = \varphi_2 + \frac{\varphi_2 - \varphi_1}{2^p - 1}$$
$$c_p \approx \frac{\varphi_2 - \varphi_1}{h^p - \frac{h^p}{2^p}} = \frac{\varphi_2 - \varphi_1}{h^p(1-2^{-p})}$$

2. $p = 1$ esetén az explicit Euler-módszernél megadott h és $\frac{h}{2}$ lépésközű eredményt kapjuk meg.

$$\varphi \approx \frac{h\varphi_2 - \frac{h}{2}\varphi_1}{h - \frac{h}{2}} = 2\left(\varphi_2 - \frac{1}{2}\varphi_1\right) = 2\varphi_2 - \varphi_1$$
$$c_p \approx \frac{\varphi_2 - \varphi_1}{h - \frac{h}{2}} = \frac{2}{h}(\varphi_2 - \varphi_1)$$

Feladatok

4-4. Az előző ötletet felhasználhatjuk többféle lépésközű közelítésekből jobb közelítés illetve pontosabb hibaképlet megadására is. Jelölje φ továbbra is a pontos értéket, amit közelíteni szeretnénk. Legyen φ_1, φ_2 és φ_3 a h_1, h_2 és h_3 lépésközzel ugyanazzal a módszerrel számított érték, c_p és c_{p-1} a pontosabb hibakonstansok.

$$\begin{aligned}\varphi - \varphi_1 &\approx h_1^p c_p + h_1^{p-1} c_{p-1}\\ \varphi - \varphi_2 &\approx h_2^p c_p + h_2^{p-1} c_{p-1}\\ \varphi - \varphi_3 &\approx h_3^p c_p + h_3^{p-1} c_{p-1}\end{aligned}$$

Írjuk fel mátrixos alakban is.

$$\begin{bmatrix}1 & -h_1^p & -h_1^{p-1}\\ 1 & -h_2^p & -h_2^{p-1}\\ 1 & -h_3^p & -h_3^{p-1}\end{bmatrix} \cdot \begin{bmatrix}\varphi\\ c_p\\ c_{p-1}\end{bmatrix} \approx \begin{bmatrix}\varphi_1\\ \varphi_2\\ \varphi_2\end{bmatrix}$$

A kapott LER megoldásával megkaphatjuk a keresett értékek közelítését.

4-5. Általánosítsuk az extrapolációs hibabecslés képletét arra az esetre, hogy h_1 és h_2 különböző lépésközzel teszünk meg egy lépést és ezután a $h_1 + h_2$ hosszú lépéssel megszerezzük az összehasonlítási értéket!

4-6. Próbáljuk ki a tanult hibabecslést a tanult módszerekre és a következő teszt kezdetiérték problémára (írjunk rá programot)

$$\begin{aligned}y'(x) = q \cdot y(x),\ \ x \in [0;1] \quad y(0) = 1,\\ q = -10; 1; 10, \quad h = 1/10; 1/20; 1/40.\end{aligned}$$

4-7. Írjunk programot a 2(3)-os Runge-Kutta módszerre, majd alkalmazzuk az

$$y' = y \cdot \cos(x),\ \ x \in [0;10] \quad y(0) = 1$$

kezdetiérték problémára! A pontos megoldás: $y(x) = e^{\sin(x)}$.
Állapítsuk meg, hogy $h = 0.1$ és $\varepsilon = 10^{-2}, 10^{-4}, 10^{-6}$ esetén melyik lépésválasztási stratégia a legjobb a maximális hiba szempontjából! (Olvassuk le a grafikonokról!)

4-8. Egy egyenes pályán mozog egy m tömegű test. Az l hosszú pálya két végén két rugó van, melyek rugóállandója D. A mozgás során a testre surlódás hat μ surlódási együtthatóval. Modellezzük a test mozgását az idő függvényében!

4.9. Lépésválasztás

Ha tudjuk, hogy adott ε pontosság esetén $\|\mathbf{g_{j-1}}\| \le \varepsilon$ és $\mu(f_y) \le m$, akkor korábbi tételünk alapján

$$\|\mathbf{e_i}\| \le M \sum_{j=1}^{i} \|\mathbf{g_{j-1}}\| h_j \le M\varepsilon \sum_{j=1}^{i} h_j \le M\varepsilon x_i \le M\varepsilon, \quad \text{ahol} \quad M := \begin{cases} e^m, & m > 0\\ 1, & m \le 0.\end{cases}$$

- Tehát, ha $\|\mathbf{g_{j-1}}\| \le \varepsilon$, akkor elfogadjuk a lépést és x_i-ből továbblépünk. Ezzel biztosítjuk az eredményben az $O(\varepsilon)$ pontosságot.

- Ha az előző feltétel nem teljesül, akkor csökkentjük a lépéstávolságot.
  $\|\mathbf{g_{j-1}}(h)\| = C_p h^p + O(h^{p+1})$ miatt elég kicsi h-ra $\|\mathbf{g_{j-1}}(h)\| \le \varepsilon$ elérhető.
  A jó h lépésközt a $\|\mathbf{g_{j-1}}(h)\| = \varepsilon$ képletből szeretnénk számolni, de a $\mathbf{g_{j-1}}(h)$ pontos lokális hibát nem ismerjük, csak becsülni tudjuk. Ehhez egy próbalépést alkalmazunk h_j-vel. Ekkor a lokális hiba tapasztalati becslése

$$\|\widetilde{\mathbf{g}}_{\mathbf{j-1}}(h_j)\| = C_p h_j^p + O(h_j^{p+1}).$$

  Innen h^p/h_j^p-vel szorozva

$$\frac{\|\widetilde{\mathbf{g}}_{\mathbf{j-1}}(h_j)\|}{h_j^p} h^p = C_p h^p + O(h_j h^p) = \|\mathbf{g_{j-1}}(h)\| + h^p O(h_j + h) = \varepsilon + h^p O(h_j + h).$$

  Elhanyagolva a második tagot kiszámíthatjuk a jó h értékét:

$$\frac{\|\widetilde{\mathbf{g}}_{\mathbf{j-1}}(h_j)\|}{h_j^p} h^p = \varepsilon, \quad \text{azaz} \quad h^* := h_j \left(\frac{\varepsilon}{\|\widetilde{\mathbf{g}}_{\mathbf{j-1}}(h_j)\|} \right)^{1/p}$$

  amivel lépnünk kell az ε pontosság eléréséhez. A számítási gyakorlat igazolta, hogy érdemes egy c_0 „biztonsági szorzót” beiktatni:

$$h^* := c_0 h_j \left(\frac{\varepsilon}{\|\widetilde{\mathbf{g}}_{\mathbf{j-1}}(h_j)\|} \right)^{1/p}, \quad c_0 \in [0.8; 1).$$

A kapott képletet a következő **stratégiákban** hasznosíthatjuk:

1. A h_j és $\|\widetilde{\mathbf{g}}_{\mathbf{j-1}}(h_j)\|$ adatok alapján meghatározzuk h^* értékét, ez lesz a következő lépés hossza

$$h_{j+1} := h^*.$$

2. Az eredeti elképzelést részben megvalósítjuk:

- ha $h^* < 0.1\, h_j$, akkor $h_j := 0.1\, h_j$ és újra lépjük a j. lépést,
- ha $0.1\, h_j \le h^* \le h_j$, akkor $h_j := h^*$ és újra lépjük a j. lépést.

A fenti két lépés összevonható: a $h^* \le h_j$ feltétel mellett elvetjük az előző lépést és $h_j := \max(0.1\, h_j, h^*)$ lépésközzel újra lépjük a j. lépést.

- Ha $h_j < h^* \le 2\, h_j$, akkor bár lehetett volna nagyobb hosszal lépni, de ismétlése saját magunk büntetése lenne. Ekkor az új h^*-t csak a következő lépésben alkalmazzuk: $h_{j+1} := h^*$,
- ha $h^* > 2\, h_j$, akkor $h_{j+1} := 2\, h_j$-vel lépünk tovább.

A fenti két lépés összevonható: a $h_j \le h^*$ feltétel mellett $h_{j+1} := \min(2\, h_j, h^*)$ lépésközzel lépjük a következő $j+1$. lépést.

Megjegyzések.

1. A legelső lépésben érdemes megengedni a lépéshossz többszöri újraválasztását is, míg a kezdeti lépés hibája elfogadható nem lesz.

2. Ha a pontosabb módszer eredményével lépünk tovább (pl. beágyazott RK-módszer esetén a p-edrendűvel, felezés esetén az extrapolációs értékkel), akkor gyakorlatilag használható a fenti lépésválasztás, de a hibabecslés nem használható, mert az a kevésbé pontos eredményre vonatkozott. Ugyanakkor az eljárás hatékonyabb, egyenlő műveletigénnyel pontosabb eredményt kapunk.

3. Az eddigi stratégiák hátránya, hogy lassan változó megoldások ellenére erősen oszcilláló lépéstávolságokat eredményez, pl. merev differenciálegyenletek esetén, ami a visszautasított lépések és a gyakran szükségtelenül kicsi lépéstávolságok miatt az eljárás hatékonyságát csökkenti. Ekkor előnyös, ha az előző lépést jellemző mennyiségeket is figyelembe vesszük pl.

$$h := \sqrt{h_j h_{j-1} \left(\frac{\varepsilon}{\|\widetilde{g}_{j-1}(h_j)\|} \right)^{1/p}}.$$

5. fejezet

Lineáris többlépéses módszerek (LTM)

5.1. Általános lineáris többlépéses módszerek

Ebben a fejezetben egy olyan módszercsaláddal foglalkozunk, mely magasabb rend esetén kevesebb függvénykiértékelést igényel, mint a RK-módszerek. Induljunk ki újra az Euler-módszerből

$$y_{i+1} = y_i + hf_i, \quad \text{ahol} \quad f_i := f(x_i, y_i).$$

Ez egy egylépéses módszer, mivel csak a megelőző y_i-t használjuk az új y_{i+1} közelítéséhez. Ennek megfelelően az l-lépéses módszereket úgy általánosíthatjuk, hogy l régi értéket, az $y_i, y_{i+1}, \ldots, y_{i+l-1}$ értékeket használjuk fel az új y_{i+l} közelítéshez.

5-1. Definíció. Ha $\alpha_l = 1$ és $|\alpha_0| + |\beta_0| \neq 0$, akkor a

$$\sum_{k=0}^{l} \alpha_k y_{i+k} = h \sum_{k=0}^{l} \beta_k f_{i+k}, \quad f_{i+k} := f(x_{i+k}, y_{i+k})$$

alakban felírt módszert lineáris l-lépéses módszernek nevezzük.

Az $\alpha_l = 1$ feltételt azért követeltük meg, mert azon módszerek közt, ahol az α_k, β_k és $\overline{\alpha}_k, \overline{\beta}_k$ együtthatók csak egy szorzóban különböznek, valójában ugyanazt a megoldást generálják. A kezdetiértékproblémából csak y_0-at ismerjük, ezért a módszer alkalmazásához egy explicit módszerrel (pl. RK) a hiányzó $y_1, y_2, \ldots, y_{l-1}$ értékeket ki kell számolnunk.

a) Ha $\beta_l = 0$, akkor explicit módszerről beszélünk. Az i. lépésben az $y_i, y_{i+1}, \ldots, y_{i+l-1}$ és $f_i, f_{i+1}, \ldots, f_{i+l-2}$ értékekből számítjuk f_{i+l-1}-et, majd y_{i+l}-et:

$$f_{i+l-1} := f(x_{i+l-1}, y_{i+l-1})$$
$$y_{i+l} := \sum_{k=0}^{l-1} (h\beta_k f_{i+k} - \alpha_k y_{i+k})$$

b) Ha $\beta_l \neq 0$, akkor implicit módszerről beszélünk. Átrendezett alakja:

$$y_{i+l} - h\beta_l f_{i+l} = \sum_{k=0}^{l-1} (h\beta_k f_{i+k} - \alpha_k y_{i+k}),$$

ahol a jobboldal nem tartalmaz y_{i+l}-es tagot. Az i. lépésben az $y_i, y_{i+1}, \dots, y_{i+l-1}$ értékekből először kiszámítjuk f_{i+l-1}-et, majd az F_i^l mennyiséget, majd alkalmazzuk a fixpontiterációt y_{i+l} közelítésére.

$$\begin{aligned} f_{i+l-1} &:= f(x_{i+l-1}, y_{i+l-1}) \\ F_i^l &:= \sum_{k=0}^{l-1} (h\beta_k f_{i+k} - \alpha_k y_{i+k}) \\ y_{i+l}^{(0)} &\quad \text{adott egy explicit módszerből vagy pl.} \\ y_{i+l}^{(0)} &:= y_{i+l-1} \\ m &= 0, 1, \dots \\ y_{i+l}^{(m+1)} &:= h\beta_l f(x_{i+l}, y_{i+l}^{(m)}) + F_i^l. \end{aligned}$$

Az i. lépésben a

$$\varphi(y) = h\beta_l f(x_{i+l}, y) + F_i^l$$

operátor fixpontját közelítjük. A konvergencia bizonyításához a Banach-féle fixponttételt használjuk.

5-1. T *Ha f a második változójában kielégíti a Lipschitz-feltételt az L_f állandóval és a fixpontiteráció kontrakciós állandója*

$$q = h|\beta_l|L_f < 1,$$

akkor a fenti iteráció konvergens.

Bizonyítás. Vizsgáljuk meg, hogy φ kontrakció-e $\mathbb{R}^n$-en. A Lipschitz-felételt felhasználva

$$\begin{aligned} \|\varphi(y) - \varphi(z)\| &= \|(h\beta_l f(x_{i+l}, y) + F_i^l) - (h\beta_l f(x_{i+l}, z) + F_i^l)\| = \\ &= h|\beta_l| \cdot \|f(x_{i+l}, y) - f(x_{i+l}, z)\| \le h|\beta_l|L_f \cdot \|y - z\|, \end{aligned}$$

$q = h|\beta_l|L_f < 1$ esetén φ kontrakció, ezért az iteráció konvergens. ■

Megjegyzés.
A fenti feltétel a $h < \frac{1}{|\beta_l|L_f}$ feltételt jelenti a lépésközre. Azonban, ha q 1-hez közeli, akkor a konvergencia lassú, sok függvénykiértékelésre van szükségünk, ezért nem lesz hatékonyabb az RK-módszereknél. Érdemes feltennünk, hogy

$$h \le \frac{1}{2|\beta_l|L_f} \quad \Rightarrow \quad q \le \frac{1}{2},$$

így a konvergencia gyors. A gyakorlatban jó kezdeti érték esetén legtöbbször egy-két iterációt végzünk. Ekkor lépésenként két-három függvénykiértékelés szükséges: f_{i+l-1}, $f(x_{i+l}, y_{i+l}^{(0)})$, $f(x_{i+l}, y_{i+l}^{(1)})$. Összehasonlításul a Dormand–Prince-féle 5(4) RK-képlet lépésenként 6 függvénykiértékelést igényel.

5.2. Adams-módszerek

5-2. Definíció. Azokat a többlépéses módszereket, melyekre

$$y_i = y_{i-1} + h \sum_{k=0}^{l} \beta_k f_{i-k},$$

Adams-féle módszereknek nevezik.
Ha $\beta_0 = 0$, akkor explicit a módszer és Adams-Bashforth-módszernek nevezzük. Ha $\beta_0 \neq 0$, akkor implicit a módszer és Adams-Moulton-módszernek nevezzük.

Megjegyzések.

1. A módszer indexelése eltér az általános többlépés módszerekétől. Az y_i értékhez y_{i-1}-re és az $f_i, f_{i-1}, \ldots, f_{i-l}$ értékekre van szükségünk. Vagyis az $y_{i-2}, \ldots, y_{i-l}$ értékekre csak közvetett módon f-en keresztül.

2. Az explicit Adams-módszereket stabilitási okok miatt már régóta nem használják önállóan, csak implicit módszerrel kombinálva.

3. Az Adams-féle módszereket Lagrange-interpolációval is levezethetjük. Nézzük az $n = 1$ (valós) esetre a konstrukciót implicit Adams-féle módszerek esetén. A kezdetiértékproblémát integrálva

$$y(x_i) = y(x_{i-1}) + \int_{x_{i-1}}^{x_i} f(x, y(x))\,dx.$$

Itt az $f(x, y(x))$ függvényt közelítjük az $(x_{i-k}, f_{i-k})_{k=0}^{l}$ pontokra felírt $L_l(x)$ l-edfokú Lagrange-interpolációs polinomjával.

$$y(x_i) \approx y(x_{i-1}) + \int_{x_{i-1}}^{x_i} L_l(x)\,dx = y(x_{i-1}) + \int_{x_{i-1}}^{x_i} \sum_{k=0}^{l} f_{i-k}\ell_{i-k}(x)\,dx =$$

$$= y(x_{i-1}) + \sum_{k=0}^{l} f_{i-k} \int_{x_{i-1}}^{x_i} \ell_{i-k}(x)\,dx,$$

ahol $\ell_{i-k}(x)$ az $i-k$. alapponthoz tartozó Lagrange-alappolinom. Ezzel a választással implicit eljárást kapunk, mert y_i előállításához f_i-t is felhasználjuk.

$$y_i = y_{i-1} + \sum_{k=0}^{l} \overline{\beta}_{i-k} f_{i-k}, \quad \overline{\beta}_{i-k} = \int_{x_{i-1}}^{x_i} \ell_{i-k}(x)\,dx$$

Ha a felosztásunk egyenletes, akkor a $\overline{\beta}_{i-k}$ értékeink i-től függetlenek lesznek. Nézzük meg ezt az esetet. Tegyük fel, hogy $x_i = ih$ és végezzünk el egy $x = x_i - sh$ helyettesítést.

$$\overline{\beta}_{i-k} = \int_{x_{i-1}}^{x_i} \ell_{i-k}(x)\,dx = \int_{x_{i-1}}^{x_i} \prod_{j=0, j\neq k}^{l} \frac{x - x_{i-j}}{x_{i-k} - x_{i-j}}\,dx =$$

$$= \int_0^1 h \prod_{j=0, j\neq k}^{l} \frac{x_i - sh - x_i + jh}{x_i - kh - x_i + jh}\,ds = h \int_0^1 \prod_{j=0, j\neq k}^{l} \frac{j-s}{j-k}\,ds$$

A kvadratúra formuláknál tanultakhoz hasonlóan bevezethetünk i-től és h-tól független együtthatókat.

$$\beta_k = \int_0^1 \prod_{j=0, j\neq k}^{l} \frac{j-s}{j-k}\,ds \quad (k = 0, \ldots, l)$$

4. Nézzük végig az explicit módszerek konstrukcióját az $n = 1$ esetben. A kezdetiértékproblémát integrálva

$$y(x_i) = y(x_{i-1}) + \int_{x_{i-1}}^{x_i} f(x, y(x))\,dx.$$

Itt az $f(x, y(x))$ függvényt közelítjük az $(x_{i-k}, f_{i-k})_{k=1}^{l}$ pontokra felírt $L_{l-1}(x)$ $l-1$-edfokú Lagrange-interpolációs polinomjával.

$$y(x_i) \approx y(x_{i-1}) + \int_{x_{i-1}}^{x_i} L_{l-1}(x)\,dx = y(x_{i-1}) + \int_{x_{i-1}}^{x_i} \sum_{k=1}^{l} f_{i-k}\ell_{i-k}(x)\,dx =$$
$$= y(x_{i-1}) + \sum_{k=1}^{l} f_{i-k} \int_{x_{i-1}}^{x_i} \ell_{i-k}(x)\,dx,$$

ahol $\ell_{i-k}(x)$ az $i-k$. alapponthoz tartozó Lagrange-alappolinom.

$$y_i = y_{i-1} + \sum_{k=1}^{l} \overline{\beta}^*_{i-k} f_{i-k}, \quad \overline{\beta}^*_{i-k} = \int_{x_{i-1}}^{x_i} \ell_{i-k}(x)\,dx$$

Ha a felosztásunk egyenletes, akkor a $\overline{\beta}^*_{i-k}$ értékeink i-től függetlenek lesznek. Nézzük meg ezt az esetet. Tegyük fel, hogy $x_i = ih$ és végezzünk el egy $x = x_i - sh$ helyettesítést.

$$\overline{\beta}^*_{i-k} = \int_{x_{i-1}}^{x_i} \ell_{i-k}(x)\,dx = \int_{x_{i-1}}^{x_i} \prod_{j=1, j\neq k}^{l} \frac{x - x_{i-j}}{x_{i-k} - x_{i-j}}\,dx =$$
$$= \int_0^1 h \prod_{j=1, j\neq k}^{l} \frac{x_i - sh - x_i + jh}{x_i - kh - x_i + jh}\,ds = h \int_0^1 \prod_{j=1, j\neq k}^{l} \frac{j-s}{j-k}\,ds$$

A kvadratúra formuláknál tanultakhoz hasonlóan bevezethetünk i-től és h-tól független együtthatókat.

$$\beta^*_k = \int_0^1 \prod_{j=1, j\neq k}^{l} \frac{j-s}{j-k}\,ds \quad (k = 1, \ldots, l)$$

5. Ellenőrzésképpen érdemes az $f \equiv c$ konstanst helyettesíteni a módszerek képletébe. Az interpoláció miatt $L(x) \equiv c$, ekkor a módszerek a pontos megoldást adják:

$$y_i = y_{i-1} + hc \equiv y_{i-1} + hc \sum_{k=0}^{l} \beta_k,$$

amiből az implicit illetve explicit esetre

$$\sum_{k=0}^{l} \beta_k = 1 \quad \text{és} \quad \sum_{k=1}^{l} \beta^*_k = 1$$

adódik.

5-1. Példa. Készítsük el az 1-lépéses Adams-Moulton módszert!

Megoldás. A fenti implicit képletek alapján

$$\beta_k = \int_0^1 \prod_{j=0, j\neq k}^{1} \frac{j-s}{j-k}\,ds \quad (k = 0, 1).$$

De gondolkodhatunk úgy is, hogy a $0, 1$ alappontokhoz tartozó alappolinomokat kell integrálnunk $[0; 1]$-en.

$$\beta_0 = \int_0^1 (1-s)\,ds = \frac{1}{2}, \quad \beta_1 = \int_0^1 s\,ds = \frac{1}{2}$$

Így az 1-lépéses Adams-Moulton módszer alakja

$$y_i = y_{i-1} + \frac{h}{2}(f_i + f_{i-1}) \quad \Leftrightarrow \quad y_{i+1} = y_i + \frac{h}{2}(f_i + f_{i+1}),$$

ami a trapéz-módszert adja (lásd később az implicit RK-módszereknél). ■

5-2. Példa. Készítsük el az 1-lépéses Adams-Bashforth módszert!

Megoldás. A $l = 1$ esetben az egyetlen (x_{i-1}, f_{i-1}) pontra írjuk fel a konstans Lagrange interpolációs polinomot, ami $L_0(x) \equiv f_{i-1}$. Így ennek integrálja $[0; 1]$-en hf_{i-1}.
A módszer tehát

$$y_i = y_{i-1} + hf_{i-1},$$

vagyis az Euler-módszert kaptuk.

■

Feladatok

5-1. Igazoljuk, hogy az implicit Euler-módszer is Adams-Moulton-módszer!

$$y_i = y_{i-1} + hf_i$$

5-2. Készítsük el a 2-lépéses Adams-Bashforth módszert!

$$y_i = y_{i-1} + \frac{h}{2}\left(3f_{i-1} - f_{i-2}\right)$$

5-3. Készítsük el a 3-lépéses Adams-Bashforth módszert!

$$y_i = y_{i-1} + \frac{h}{12}\left(23f_{i-1} - 16f_{i-2} + 5f_{i-3}\right)$$

5-4. Készítsük el a 4-lépéses Adams-Bashforth módszert!

$$y_i = y_{i-1} + \frac{h}{24}\left(55f_{i-1} - 59f_{i-2} + 37f_{i-3} - 9f_{i-4}\right)$$

5-5. Készítsük el a 2-lépéses Adams-Moulton módszert!

$$y_i = y_{i-1} + \frac{h}{12}\left(5f_i + 8f_{i-1} - f_{i-2}\right)$$

5-6. Készítsük el a 3-lépéses Adams-Moulton módszert!

$$y_i = y_{i-1} + \frac{h}{24}\left(9f_i + 19f_{i-1} - 5f_{i-2} + f_{i-3}\right)$$

5-7. Készítsük el az 4-lépéses Adams-Moulton módszert!

$$y_i = y_{i-1} + \frac{h}{720}\left(251f_i + 646f_{i-1} - 264f_{i-2} - 126f_{i-3} - 19f_{i-4}\right)$$

5.3. A középpont szabály

Az egyik legegyszerűbb többlépéses módszer a középpontszabály, mely stabil és másodrendű, de vannak egyéb problémái, ami miatt a gyakorlatban nem alkalmazzák. Arra azonban alkalmas, hogy az új fogalmakat szemléltessük rajta. Az Euler-módszerhez képest az x_{i+1}-beli közelítést most az x_{i-1}-beli közelítésből kapjuk: az x_i pontbeli meredekséggel ($y'(x_i) = f(x_i, y_i)$-vel) $2h$-t lépünk x_{i-1}-ből.

5-3. Definíció. A középpont szabály alakja

$$y_{i+1} = y_{i-1} + 2hf_i,$$

ahol y_0 adott, y_1 például a javított Euler-módszerrel előállított közelítés.

5-4. Definíció. A korábbi definícióval összhangban a középpont szabály képlethibája (vagy lokális hibája) a

$$g(x_i, h) = g_i = \frac{y(x_{i+1}) - y(x_{i-1})}{h} - 2f(x_i, y(x_i))$$

mennyiség, ahol $y(x)$ a differenciálegyenlet pontos megoldása.

5-2. T *(A középpontszabály tulajdonságai)*

a) *Legyen $f \in C^2([0;1] \times \mathbb{R})$, akkor a középpont szabály másodrendben konzisztens.*

b) *Ha $f \in Lip_y$, akkor a középpont szabály stabil.*

c) *A középpont szabály az előző f-re tett feltételekkel konvergens.*

Bizonyítás. a) Írjuk fel a Taylor-formulát hg_i-re x_i középponttal, harmadrendű maradéktaggal.

$$\begin{aligned} hg_i &= y(x_i) + hy'(x_i) + \frac{h^2}{2}y''(x_i) + \frac{h^3}{6}y'''(\mu_i) \\ &\quad - \left(y(x_i) - hy'(x_i) + \frac{h^2}{2}y''(x_i) - \frac{h^3}{6}y'''(\nu_i)\right) - 2hy'(x_i) = \\ &= \frac{h^3}{3}\frac{(y'''(\mu_i) + y'''(\nu_i))}{2} = \frac{h^3}{3}y'''(\xi_i) \end{aligned}$$

Innen

$$|g_i| \le \frac{h^2}{3}M_3, \quad \text{ahol} \quad M_3 = \max_{x\in[0;1]} |y'''(x)|,$$

ami a másodrendű konzisztenciát jelenti.

b) A stabilitásvizsgálathoz a globális hibát kell becsülnünk a lokális hibákkal

$$\begin{aligned} e_{i+1} &= y(x_{i+1}) - y_{i+1} = [y(x_{i-1}) + 2hf(x_i, y(x_i)) + hg_i] - (y_{i-1} + 2hf_i) = \\ &= (y(x_{i-1}) - y_{i-1}) + 2h(f(x_i, y(x_i)) - f_i) + hg_i = e_{i-1} + 2h\phi_i e_i + hg_i \end{aligned}$$

a hibaegyenlet, a

$$\phi_i = \begin{cases} \frac{f(x_i,y(x_i)) - f(x_i,y_i)}{y(x_i) - y_i}, & \text{ha } y(x_i) \ne y_i \\ L_f, & \text{ha } y(x_i) = y_i \end{cases}$$

segédfüggvénnyel. Így $|\phi_i| \le L_f$. A globális hibára kapott rekurziót e_0-ra kellene visszafejtenünk, azonban itt kétlépéses rekurziónk van, ezért áttérünk mátrixos felírásra. Vektorosan már egy mélységű lesz a rekurzió.

$$\begin{bmatrix} e_i \\ e_{i+1} \end{bmatrix} = \left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + h\begin{bmatrix} 0 & 0 \\ 0 & 2\phi_i \end{bmatrix}\right)\begin{bmatrix} e_{i-1} \\ e_i \end{bmatrix} + h\begin{bmatrix} 0 \\ g_i \end{bmatrix},$$

$$\mathbf{u_{i+1}} = (\mathbf{A} + h\mathbf{B_i})\mathbf{u_i} + h\mathbf{v_i},$$

ahol

$$\mathbf{u_i} = \begin{bmatrix} e_{i-1} \\ e_i \end{bmatrix}, \quad \mathbf{v_i} = \begin{bmatrix} 0 \\ g_i \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{B_i} = \begin{bmatrix} 0 & 0 \\ 0 & 2\phi_i \end{bmatrix}$$

Ezután maximum normában felírva a becslést:

$$\begin{aligned} \|\mathbf{u_{i+1}}\|_\infty &\le (\|\mathbf{A}\|_\infty + h\|\mathbf{B_i}\|_\infty)\,\|\mathbf{u_i}\|_\infty + h\|\mathbf{v_i}\|_\infty \le \\ &\le (1 + 2hL_f)\|\mathbf{u_i}\|_\infty + h|g_i|, \end{aligned}$$

azaz

$$\begin{aligned} &\|\mathbf{u_2}\|_\infty \le (1 + 2hL_f)\|\mathbf{u_1}\|_\infty + h|g_1|, \\ &\|\mathbf{u_3}\|_\infty \le (1 + 2hL_f)\|\mathbf{u_2}\|_\infty + h|g_2| \le (1 + 2hL_f)^2\,(\|\mathbf{u_1}\|_\infty + h|g_1| + h|g_2|)\,. \end{aligned}$$

Visszafejtve

$$\begin{aligned} |e_{i+1}| \le \|\mathbf{u_{i+1}}\|_\infty &\le (1 + 2hL_f)^i \left(\|\mathbf{u_1}\|_\infty + \sum_{k=1}^{i} |g_k|h\right) \le \\ &\le (1 + 2hL_f)^i \left(\max(|e_0|, |e_1|) + \sum_{k=1}^{i} |g_k|h\right) \le \\ &\le e^{2L_f x_i} \left(\max(|e_0|, |e_1|) + \sum_{k=1}^{i} |g_k|h\right) \end{aligned}$$

Ami a stabilitást jelenti a középpont szabályra.

c) Ha teljesülnek az f-re tett feltételek és $e_0 = 0$, $e_1 = O(h^2)$, akkor az előző becslésből:

$$|e_i| \le e^{2L_f} \left(|e_1| + x_i \frac{h^2}{3} M_3\right) = O(h^2).$$

■

Megjegyzések.

1. Ha y_1-et az Euler-módszerrel számítjuk ki, akkor

$$|e_1| \le h|g_0| = \frac{h^2}{2} M_2,$$

ami teljesíti a feltételeket, de más szempontból nem jó.

2. Ha y_1-et a javított Euler-módszerrel számítjuk ki, akkor

$$|e_1| \le h|g_0| = \frac{h^3}{6} M_3 + \frac{h^3}{8} \|F_2 f\|_{C[0;1]},$$

ami túl pontosnak tűnik, azonban a számítógépes program jobb lesz. Érdemes az első lépést pontosabban megtenni, mert az ekkor okozott hiba az egész megoldásra kihat.

5-3. Példa. Alkalmazzuk a középpont szabályt az $f(x, y) = qy$ esetben, ahol q konstans. Elemezzük a viselkedését.

Megoldás. Ekkor az

$$y_{i+1} - 2hqy_i - y_{i-1} = 0, \quad i = 1, 2, \dots$$

differenciaegyenletet kapjuk. Mivel $y(x) = e^{qx}$ az egzakt megoldása a differenciálegyenletnek, ezért tekintsük az

$$y_0 = 1, \quad y_1 = e^{hq}$$

pontos kezdeti feltételeket. A differenciaegyenlet megoldását a karakterisztikus egyenlet gyökeivel írjuk fel.

$$\varrho(z) = z^2 - 2hqz - 1 = 0 \quad \rightarrow \quad z_{1,2} = \frac{2hq \pm \sqrt{4h^2q^2 + 4}}{2} = hq \pm \sqrt{1 + (hq)^2}$$

Mivel $hq \ll 1$, ezért a gyököket a

$$z_1 = hq + \sqrt{1 + (hq)^2}, \quad z_2 = -\frac{1}{z_1}$$

stabilabb alakban használjuk. A differenciaegyenlet általános megoldása:

$$y_i = c_1 z_1^i + c_2 \left(-\frac{1}{z_1}\right)^i, \quad i = 0, 1, \dots, N$$

A kezdeti feltételt felhasználva határozzuk meg c_1 és c_2 értékét. A Maple segítségével oldjuk meg a kapott 2×2-es LER-t, majd vizsgáljuk a pontos megoldás ($y(x_i) = y(ih) = e^{ihq}$) és a középpont szabállyal kapott megoldás (y_i) eltérését $q < 0$ illetve $q > 0$ esetben. Rajzoljuk ki az eltérést konkrét h ($h \to 0$) és q esetén. Láthatjuk, hogy a $q < 0$ esetben, amikor a pontos megoldás monoton fogyó, a közelítésünk oszcilláló lesz, vagyis a kapott közelítésünk a megoldás viselkedését nem tartja meg. A $q > 0$ esetben ez a probléma nem jelentkezik, a megoldás és a közelítés is monoton növő. A pontos elméleti elemzést lásd [10] 67. oldalán. ■

A fenti tanulságos példa mutatja, hogyan lehet egy numerikus módszer konzisztens, stabil és konvergens, de gyakorlatilag használhatatlan. Az ilyen kellemetlen példák miatt kézenfekvő, hogy a stabilitási fogalmat szigorítsuk vagy megköveteljük, hogy a 0-stabilitás mellett bizonyos alapvetően fontos differenciálegyenleteken megfelelően viselkedjen bármely h lépésköz esetén.

5-5. Definíció. A $\mathbf{z_h} = (z_0, \dots, z_N)^T \in \mathbb{R}^{N+1}$ vektor Spijker-normája a következő mennyiség

$$\|\mathbf{z_h}\|_S = \max(|z_0|, \dots, |z_{l-1}|) + \max_{i=l}^{N} \left| \sum_{j=l}^{i} z_j h \right|,$$

ahol $1 \leq l$ rögzített érték. A középpont szabály esetén $l = 2$.

A következő tétel megmutatja, hogy ha a stabilitás általános definíciójában a jobboldali $\|.\|_*$ normát speciálisan választjuk, akkor a középpont szabály nem stabil. Vagyis a stabilitás függ a normától.

5-3. T *(Spijker, a középpont szabály instabil)*
A középpont szabály nem stabil a Spijker-normában, ha y_1 számítására az Euler-módszert alkalmazzuk. Ezt formálisan is megfogalmazzuk a következőkben. A középpont szabályt írjuk fel egy $F : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ leképezéssel:

$$\begin{aligned}
(F(\mathbf{y_h}))_0 &:= y_0, \\
(F(\mathbf{y_h}))_1 &:= y_1 - y_0 - hf(x_0, y_0), \\
(F(\mathbf{y_h}))_i &:= \frac{y_i - y_{i-2}}{2h} - f(x_i, y_i), \quad i = 2, \dots, N,
\end{aligned}$$

ahol $\mathbf{y_h} = (y_0, \ldots, y_N)^T \in \mathbb{R}^{N+1}$. *Ekkor nincs az* N*-től független olyan* M *konstans, melyre a középpont szabály két tetszőleges* $\mathbf{y_h^{(1)}}, \mathbf{y_h^{(2)}}$ *megoldása esetén*

$$\|\mathbf{y_h^{(1)}} - \mathbf{y_h^{(2)}}\|_\infty \leq M\|F(\mathbf{y_h^{(1)}}) - F(\mathbf{y_h^{(2)}})\|_S.$$

Bizonyítás. Lásd [10] 68. oldalán. ■

5.4. Többlépéses módszerek 0-stabilitása

Lineáris többlépéses módszerek esetén az $y_1, \ldots, y_{l-1}$ kezdeti értékek hibát tartalmaznak. Kérdés, hogy érzékeny-e a módszer a kezdeti értékek változásaira.

5-6. Definíció. Egy lineáris l-lépéses módszer 0-stabil (zero-stabil), ha létezik egy $K \in \mathbb{R}$ konstans, hogy bármely, a módszerrel generált $\mathbf{y_h}$ és $\mathbf{z_h}$ sorozatra $h \to 0$ esetén teljesül minden $i = l, \ldots, N$-re, hogy

$$|y_i - z_i| \leq K \max\{|y_0 - z_0|, |y_1 - z_1|, \ldots, |y_{l-1} - z_{l-1}|\},$$

ahol $y_0, \ldots, y_{l-1}$, $z_0, \ldots, z_{l-1}$ a sorozatokhoz tartozó kezdőértékek.

5-7. Definíció. Egy lineáris l-lépéses módszer első stabilitási (karakterisztikus) polinomja a

$$\varrho(z) = \sum_{k=0}^{l} \alpha_k z^k,$$

a második stabilitási (karakterisztikus) polinomja a

$$\sigma(z) = \sum_{k=0}^{l} \beta_k z^k.$$

5-8. Definíció. Azt mondjuk, hogy ϱ *teljesíti a gyökfeltételt*, ha az első stabilitási polinom gyökei a zárt egységkörön belül vannak és az egységkörön lévő gyökei egyszeresek.

5-4. T *(LTM 0-stabilitása)*
Az LTM pontosan akkor 0*-stabil, ha a módszer első stabilitási polinomja kielégíti a gyökfeltételt.*

Bizonyítás. $\Rightarrow$: Tekintsük az $y' = 0$-ra felírt LTM-et.

$$\alpha_l y_{i+l} + \ldots + \alpha_0 y_i = 0$$

Ennek megoldása a differenciaegyenletek elméletéből a következő alakú:

$$y_i = \sum_{j=1}^{k} p_j(i) z_j^i,$$

ahol z_j egy $m_j \geq 1$ multiplicitású gyöke a $\varrho(z)$ első stabilitási polinomnak. p_j egy legfeljebb $m_j - 1$-edfokú polinom és $k \leq l$ a különböző gyökök száma. Tegyük fel indirekt, hogy a módszer 0-stabil, de a gyökök nem teljesítik a gyökfeltételt.

a) Ha valamelyik gyökre $|z_j| > 1$, akkor vannak olyan $y_0, \ldots, y_{l-1}$ kezdeti értékek, melyekre a megoldás $|z_j|^i$ szerint nő. Ha $h \to 0$ ($Nh = 1$), így $N \to \infty$, akkor a megoldás nem korlátos. Vegyük mellé a $z_0 = \ldots = z_{l-1} = 0$ kezdeti feltételeket a $z_i = 0 \; \forall\, i$-re megoldással. Ekkor a 0-stabilitás nem teljesül, ellentmondásra jutottunk. Tehát nem lehet egynél nagyobb abszolút értékű gyök.

b) Ha valamelyik gyökre $|z_j| = 1$ és $m_j > 1$ (a körvonalon lévő többszörös gyök), akkor vannak olyan kezdeti értékek, hogy a megoldásban a $p_k(i)z_j^i$ tag legalább $i^{m_j-1}z_j^i$-nel arányosan nő, így nem korlátos. A 0-stabilitás ekkor sem teljesül, ellentmondásra jutottunk.

$\Leftarrow$: Hosszadalmas, lásd [10] 75. oldalán. ■

5-4. Példa. Az Euler- és implicit Euler-módszer 0-stabil, mert $\varrho(z) = z - 1$. Ez kielégíti a gyökfeltételt, mivel $z = 1$ az egyetlen gyöke. Ugyanez vonatkozik a trapéz módszerre is.

5-5. Példa. Az l-lépéses Adams-Bashforth és Adams-Moulton módszerek is 0-stabilak, mert az első stabilitási polinomjuk $\varrho(z) = z^l - z^{l-1} = z^{l-1}(z-1)$. A $z = 0$ $l-1$-szeres gyök, $z = 1$ pedig egyszeres, ezzel teljesíti a gyökfeltételt.

5-6. Példa. A következő 3-lépéses módszer nem 0-stabil.

$$11y_{i+3} + 27y_{i+2} - 27y_{i+1} - 11y_i = 3h(f_{i+3} + 9f_{i+2} + f_i)$$

Az első stabilitási polinomja $\varrho(z) = 11z^3 + 27z^2 - 27z - 11$. Ennek gyökei:

$$z_1 = 1, \quad z_2 \approx -0.32, \quad z_3 = -3.14,$$

így $|z_3| > 1$, a gyökfeltétel nem teljesül.

5-7. Példa. A következő 3-lépéses módszer nem 0-stabil.

$$y_{i+3} + y_{i+2} - y_{i+1} - y_i = 2h(f_{i+2} + f_{i+1})$$

Az első stabilitási polinomja

$$\varrho(z) = z^3 + z^2 - z - 1 = (z^2 - 1)(z + 1) = (z + 1)^2(z - 1).$$

Ennek gyökei: $z_{1,2} = -1$, $z_3 = 1$, így az egységkörön lévő kétszeres gyök miatt a gyökfeltétel nem teljesül.

5-8. Példa. Az $y_i - y_{i-2}$ baloldalú módszerek is 0-stabilak (ilyenek a Nyström és Milne-Simpson módszer), mert az első stabilitási polinomjuk

$$\varrho(z) = z^l - z^{l-2} = z^{l-2}(z^2 - 1) = z^{l-2}(z + 1)(z - 1).$$

A $z = 0$ $l-2$-szeres gyök, $z = 1$ és $z = -1$ pedig egyszeres, ezzel teljesíti a gyökfeltételt.

5.5. Többlépéses módszerek konzisztenciája

5-9. Definíció. Az LTM képlethibája vagy lokális hibája az

$$L(x_i, y(x_i), h) := \frac{1}{h}\sum_{k=0}^{l}(\alpha_k y(x_{i+k}) - h\beta_k y'(x_{i+k}))$$

kifejezés, melyben y a pontos megoldás. Az LTM-et konzisztensnek nevezzük, ha elegendően sima $y(x)$ esetén létezik olyan h_0, hogy

$$L(x_i, y(x_i), h) = O(h^p), \;\; 0 < h \leq h_0 \;\text{ és }\; p \geq 1.$$

Ekkor a módszer konzisztencia rendje p.

5-5. L *(LTM konzisztenciája)*
Tegyük fel, hogy a differenciálegyenlet megoldása kétszer folytonosan differenciálható $[0;1]$*-en. Ha*

$$\varrho(1) = 0 \quad \textit{és} \quad (0 \neq)\, \varrho'(1) = \sigma(1),$$

akkor a módszer konzisztens.

Bizonyítás. A lokális hiba h-szorosára felírva a Taylor-formulát a másodrendű hibataggal

$$\begin{aligned} hL(x_i, y(x_i), h) &= \sum_{k=0}^{l}\{\alpha_k y(x_i + kh) - h\beta_k y'(x_i + kh)\} = \\ &= \sum_{k=0}^{l}\{\alpha_k \left(y(x_i) + khy'(x_i) + O(h^2)\right) - h\beta_k \left(y'(x_i) + O(h)\right)\} = \\ &= y(x_i)\sum_{k=0}^{l}\alpha_k + hy'(x_i)\left(\sum_{k=0}^{l}k\alpha_k - \sum_{k=0}^{l}\beta_k\right) + O(h^2) = \\ &= y(x_i)\sum_{k=0}^{l}\alpha_k + hy'(x_i)\sum_{k=0}^{l}(k\alpha_k - \beta_k) + O(h^2) \end{aligned}$$

$$hL(x_i, y(x_i), h) = C_0 y(x_i) + C_1 hy'(x_i) + O(h^2),$$

ahol

$$\begin{aligned} C_0 &= \sum_{k=0}^{l}\alpha_k = \varrho(1) \\ C_1 &= \sum_{k=0}^{l}(k\alpha_k - \beta_k) = \varrho'(1) - \sigma(1). \end{aligned}$$

A stabilitási polinomokra tett feltételből következik, hogy $C_0 = C_1 = 0$, ekkor

$$L(x_i, y(x_i), h) = O(h),$$

ami a konzisztenciát jelenti. A $(0 \neq)\, \varrho'(1) = \sigma(1)$ feltétel azt biztosítja, hogy a $z = 1$ gyök egyszeres gyöke legyen $\varrho(z)$-nek, ezzel teljesül a gyökfeltétel. ■

5-6. L *(LTM p-edrendű konzisztenciája)*
Tegyük fel, hogy a differenciálegyenlet megoldása $l+2$-szer folytonosan differenciálható. Ha az $\{\alpha_k\}_{k=0}^l$ számokra

$$\alpha_0 + \alpha_1 + \ldots + \alpha_l = 0,$$

akkor egyértelműen léteznek olyan $\{\beta_k\}_{k=0}^l$ számok, hogy az ehhez tartozó módszer konzisztencia rendje $p = l+1$ legyen.

Bizonyítás. A lokális hiba h-szorosára felírva a Taylor-formulát a $p+1$-edrendű hibataggal

$$\begin{aligned}
hL(x_i, y(x_i), h) &= \sum_{k=0}^{l} \{\alpha_k y(x_i + kh) - h\beta_k y'(x_i + kh)\} = \\
&= \sum_{k=0}^{l} \{\alpha_k \left(\sum_{j=0}^{p} \frac{(kh)^j}{j!} y^{(j)}(x_i) + \frac{(kh)^{p+1}}{(p+1)!} y^{(p+1)}(\xi_i) \right) - \\
&- h\beta_k \left(\sum_{j=0}^{p-1} \frac{(kh)^j}{j!} y^{(j+1)}(x_i) + \frac{(kh)^p}{p!} y^{(p+1)}(\eta_i) \right) \} = \\
&= y(x_i) \sum_{k=0}^{l} \alpha_k + hy'(x_i) \sum_{k=0}^{l} (k\alpha_k - \beta_k) + \ldots \\
&+ \frac{h^j}{j!} y^{(j)}(x_i) \sum_{k=0}^{l} (k^j \alpha_k - jk^{j-1}\beta_k) + \ldots + O(h^{p+1}) = \\
&= \sum_{j=0}^{p} \gamma_j \frac{h^j}{j!} y^{(j)}(x_i) + O(h^{p+1}) = \sum_{j=0}^{p} C_j y^{(j)}(x_i) h^j + O(h^{p+1}),
\end{aligned}$$

ahol $C_j = \frac{\gamma_j}{j!}$ és

$$\begin{aligned}
\gamma_0 &= \sum_{k=0}^{l} \alpha_k = \varrho(1) \\
\gamma_1 &= \sum_{k=0}^{l} (k\alpha_k - \beta_k) = \varrho'(1) - \sigma(1) \\
\gamma_2 &= \sum_{k=0}^{l} (k^2\alpha_k - 2k\beta_k) \\
\gamma_3 &= \sum_{k=0}^{l} (k^3\alpha_k - 3k^2\beta_k) \\
&\vdots \\
\gamma_j &= \sum_{k=0}^{l} (k^j\alpha_k - jk^{j-1}\beta_k), \quad j = 0, 1, \ldots, p.
\end{aligned}$$

A lemma feltétele $\gamma_0 = 0$-t biztosít. Rögzített α_k-k és $p = l+1$ mellett a fenti egyenletek a β_k-kra a következő LER megoldását jelentik.

$$\begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ 0 & 1 & 2 & & l \\ 0 & 1^2 & 2^2 & \ldots & l^2 \\ \vdots & \vdots & & & \vdots \\ 0 & 1^l & 2^l & \ldots & l^l \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_l \end{bmatrix} = \begin{bmatrix} \sum_{k=0}^{l} k\alpha_k \\ \frac{1}{2}\sum_{k=0}^{l} k^2\alpha_k \\ \frac{1}{3}\sum_{k=0}^{l} k^3\alpha_k \\ \vdots \\ \frac{1}{l+1}\sum_{k=0}^{l} k^{l+1}\alpha_k \end{bmatrix}$$

A mátrix Vandermonde-mátrix, a determinánsa nem nulla, a LER megoldása egyértelmű. Tehát a β_k-k egyértelműen megadhatók és a lokális hiba $O(h^{l+1})$. ■

Megjegyzések.

1. Ha explicit módszert akarunk konstruálni, akkor $\beta_l = 0$ miatt egy egyenlettel kevesebb áll rendelkezésre, ezért a konzisztencia rendje csak l lesz.

2. Ha rögzített l-hez $l+1$-nél magasabb rendű módszert akarunk konstruálni, akkor ez már csak az α_k-k megfelelő választásával érhető el. Így is csak $l+2$ rendig juthatunk el (lásd később), ha a 0-stabilitás követelményét figyelembe vesszük.

3. Az l-lépéses Adams-Moulton módszer lokális hibája a felhasznált Lagrange-interpoláció miatt $l+1$-edrendű lesz és hibatagja

$$\left\| \frac{f^{(l+1)}}{(l+1)!}\omega_{l+1} \right\|_\infty \leq M_{l+1}h^{l+1}.$$

Hasonlóan az Adams-Bashforth módszer l-edrendű.

4. Ha az $\alpha_l = 1$, $\alpha_{l-1} = -1$, $\alpha_k = 0$ $(k = 0, 1, \ldots, l-2)$ és megköveteljük, hogy implicit módszer esetén a rendje $l+1$, explicit módszer esetén l legyen, akkor a lemma miatt csak Adams-módszerekről lehet szó. A fenti egyenletrendszerrel a β_k számokat meghatározhatjuk, nem kell integrálokat számolnunk a konstrukcióhoz.

5-9. Példa. Keressük az explicit 3-lépéses maximális rendű módszert

$$\alpha_0 y_{i-2} + \alpha_1 y_{i-1} + \alpha_2 y_i + \alpha_3 y_{i+1} = h(\beta_0 f_{i-2} + \beta_1 f_{i-1} + \beta_2 f_i),$$

melyben

$$\alpha_0 = \alpha_2 = 0, \;\; \alpha_3 = 1.$$

Megoldás. Összesen 4 szabad paraméterünk van, ehhez 4 egyenletet kell felírnunk a lemma szerint.

$$\begin{aligned}
\gamma_0 = 0 \quad &\Leftrightarrow \quad \sum_{k=0}^{3} \alpha_k = 0 \\
\alpha_1 + \alpha_3 = 0 \quad &\Rightarrow \quad \alpha_1 + 1 = 0 \qquad\qquad \Rightarrow \quad \alpha_1 = -1 \\
\gamma_1 = 0 \quad &\Leftrightarrow \quad \sum_{k=0}^{3} k\alpha_k = \sum_{k=0}^{3} \beta_k \\
\alpha_1 + 2\alpha_2 + 3\alpha_3 = \beta_0 + \beta_1 + \beta_2 \quad &\Rightarrow \quad -1 + 3 = \beta_0 + \beta_1 + \beta_2 \\
\gamma_2 = 0 \Leftrightarrow \quad &\sum_{k=0}^{3} k^2\alpha_k = \sum_{k=0}^{3} 2k\beta_k \\
\alpha_1 + 4\alpha_2 + 9\alpha_3 = 2\beta_1 + 4\beta_2 \quad &\Rightarrow \quad -1 + 9 = 2\beta_1 + 4\beta_2 \\
\gamma_3 = \sum_{k=0}^{3} k^3\alpha_k = \sum_{k=0}^{3} 3k^2\beta_k \\
\alpha_1 + 8\alpha_2 + 27\alpha_3 = 3\beta_1 + 12\beta_2 \quad &\Rightarrow \quad -1 + 27 = 3\beta_1 + 12\beta_2
\end{aligned}$$

$\alpha_1 = -1$-et behelyettesítve

$$\begin{aligned} \beta_0 + \beta_1 + \beta_2 &= 2 \\ 2\beta_1 + 4\beta_2 &= 8 \\ 3\beta_1 + 12\beta_2 &= 26. \end{aligned}$$

A 3. egyenletből vonjuk ki a 2. egyenlet 3-szorosát.

$$-3\beta_1 = 2 \quad \rightarrow \quad \beta_1 = -\frac{2}{3}$$

A 2. egyenletből

$$4\beta_2 = 8 + \frac{4}{3} = \frac{28}{3} \quad \rightarrow \quad \beta_2 = \frac{7}{3}.$$

Az 1. egyenletből

$$\beta_0 = 2 - \beta_1 - \beta_2 = 2 + \frac{2}{3} - \frac{7}{3} = \frac{1}{3}.$$

A kapott módszerünk a Nyström-módszer, melynek rendje 3

$$y_{i+1} = y_{i-1} + \frac{1}{3}\, h\,(7f_i - 2f_{i-1} + f_{i-2}).$$

■

5-10. Példa. Keressük az implicit 3-lépéses maximális rendű módszert

$$\alpha_0 y_{i-2} + \alpha_1 y_{i-1} + \alpha_2 y_i + \alpha_3 y_{i+1} = h(\beta_0 f_{i-2} + \beta_1 f_{i-1} + \beta_2 f_i + \beta_3 f_{i+1}),$$

melyben

$$\alpha_0 = \alpha_2 = 0, \;\; \alpha_3 = 1.$$

Megoldás. Összesen 5 szabad paraméterünk van, ehhez 5 egyenletet kell felírnunk a lemma szerint.

$$\begin{aligned} \gamma_0 = 0 \quad &\Leftrightarrow \quad \sum_{k=0}^{3} \alpha_k = 0 \\ \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 = 0 \quad &\Rightarrow \quad \alpha_1 + 1 = 0 \qquad\qquad \Rightarrow \quad \alpha_1 = -1 \\ \gamma_1 = 0 \quad &\Leftrightarrow \quad \sum_{k=0}^{3} k\alpha_k = \sum_{k=0}^{3} \beta_k \\ \alpha_1 + 2\alpha_2 + 3\alpha_3 = \beta_0 + \beta_1 + \beta_2 + \beta_3 \quad &\Rightarrow \quad -1 + 3 = \beta_0 + \beta_1 + \beta_2 + \beta_3 \\ \gamma_2 = 0 \quad &\Leftrightarrow \quad \sum_{k=0}^{3} k^2\alpha_k = \sum_{k=0}^{3} 2k\beta_k \\ \alpha_1 + 4\alpha_2 + 9\alpha_3 = 2\beta_1 + 4\beta_2 + 6\beta_3 \quad &\Rightarrow \quad -1 + 9 + 2 = \beta_1 + 4\beta_2 + 6\beta_3 \\ \gamma_3 = 0 \quad &\Leftrightarrow \quad \sum_{k=0}^{3} k^3\alpha_k = \sum_{k=0}^{3} 3k^2\beta_k \\ \alpha_1 + 8\alpha_2 + 27\alpha_3 = 3\beta_1 + 12\beta_2 + 27\beta_3 \quad &\Rightarrow \quad -1 + 27 = 3\beta_1 + 12\beta_2 + 27\beta_3 \\ \gamma_4 = 0 \quad &\Leftrightarrow \quad \sum_{k=0}^{3} k^4\alpha_k = \sum_{k=0}^{3} 4k^3\beta_k \\ \alpha_1 + 16\alpha_2 + 81\alpha_3 = 4\beta_1 + 32\beta_2 + 108\beta_3 \quad &\Rightarrow \quad -1 + 81 = 4\beta_1 + 32\beta_2 + 108\beta_3 \end{aligned}$$

A Matlab vagy Maple segítségével oldjuk meg a LER-t. A kapott β_k értékek:

$$\beta_0 = 0,\ \ \beta_1 = \frac{1}{3},\ \ \beta_2 = \frac{4}{3},\ \ \beta_3 = \frac{1}{3},$$

A kapott módszerünk az implicit 3-lépéses Milne-Simpson módszer, melynek rendje 4

$$y_{i+1} = y_{i-1} + \frac{1}{3}\,h\,(f(x_{i+1}, y_{i+1}) + 4f_i + f_{i-1}).$$

■

Többlépéses módszerek készíthetők a Fejér-Hermite interpoláció felhasználásával is, erre mutatunk példát a következőkben.

5-11. Példa. Tekintsük a következő „lyukas" Fejér-Hermite interpolációt. Adottak az x_0, x_1 alappontok és az y_0, y_0', y_1' függvény és derivált értékek. Keressük azt a P másodfokú polinomot, mely az alappontokban a megadott függvény és derivált értéket adja. Ezután $y_1 = P_2(x_1)$-ből felírható a rekurzió, amit numerikus módszerként használhatunk.

Megoldás. A polinomot a Newton-alak rekurziója alapján

$$P(x) = H_1(x) + c(x - x_0)^2 = y_0 + y_0'(x - x_0) + c(x - x_0)^2$$

alakban keressük. Ennek deriváltja ismert az x_1 pontban.

$$y_1' = P'(x_1) = y_0' + 2c(x_1 - x_0) = y_0' + 2ch \quad \to \quad c = \frac{1}{2h}(y_1' - y_0')$$

Így az $y(x)$ megoldás közelítésére a

$$y(x) \approx P(x) = y_0 + y_0'(x - x_0) + \frac{1}{2h}(y_1' - y_0')(x - x_0)^2$$

polinomot kapjuk. Tehát az $y(x_1)$ közelítése

$$y(x_1) \approx y_1 := P(x_1) = y_0 + y_0'h + \frac{h}{2}(y_1' - y_0') = y_0 + \frac{h}{2}(y_1' + y_0').$$

Innen a rekurzió

$$y_{i+1} := y_i + \frac{h}{2}(f_{i+1} + f_i),$$

vagyis a trapéz-szabályt kaptuk. Látjuk, hogy y_1' megadása miatt implicit módszert kaptunk.
■

Feladatok

5-8. Miért nem alkalmazható a 6. lemma a középpont szabályra, abban az értelemben, hogy kétlépéses módszer lévén a konzisztencia rendje 3 lehetne? Konstruáljuk meg azt a 2-lépéses módszert, melynek α-számai megegyeznek a középpont szabályéval, de harmadrendű.

5-9. Készítsük el az explicit 2-lépéses maximális rendű módszert, melynek alakja

$$\alpha_0 y_{i-1} + \alpha_1 y_i + \alpha_2 y_{i+1} = h(\beta_0 f_{i-1} + \beta_1 f_i),$$

ahol $\alpha_0 = -1$, $\alpha_2 = 1$. (A középpont szabályt kapjuk.)

5-10. Készítsünk maximális rendű explicit 2-lépéses módszert, melyben az α_k-kra nem teszünk megkötéseket. Alakja

$$\alpha_0 y_{i-1} + \alpha_1 y_i + \alpha_2 y_{i+1} = h(\beta_0 f_{i-1} + \beta_1 f_i).$$

0-stabil-e a módszer? Most 4 szabad paraméterünk van, ehhez 4 egyenletet kell felírnunk és 3 lesz a módszer rendje. A következő módszert kapjuk:

$$y_{i+1} = -4y_i + 5y_{i-1} + h(2f_{i-1} + 4f_i).$$

5-11. Határozzuk meg a következő LTM módszer szabad paramétereit úgy, hogy a konzisztencia rendje maximális legyen! (f elég sima függvény.)

$$y_{i+1} - y_{i-1} = h(af_i + bf_{i-1})$$

Vizsgálja a módszer stabilitását és konvergenciáját is!

5-12. Határozzuk meg a következő LTM módszer szabad paramétereit úgy, hogy a konzisztencia rendje maximális legyen! (f elég sima függvény.)

$$y_{i+1} - y_{i-1} = h(af_{i+1} + bf_i + cf_{i-1})$$

Vizsgálja a módszer stabilitását és konvergenciáját is!

5-13. Határozzuk meg a következő LTM módszer szabad paramétereit úgy, hogy a konzisztencia rendje maximális legyen! (f elég sima függvény.)

$$ay_i + by_{i-1} + cy_{i-2} = hf_i$$

Vizsgálja a módszer stabilitását és konvergenciáját is!

5-14. Készítsünk maximális rendű implicit 2-lépéses módszert, melyben $\alpha_0 = \alpha_2 = 1$ és $\alpha_1 = -2$. Adjuk meg a $\beta_0, \beta_1, \beta_2$ értékeket. 0-stabil-e a módszer?

$$y_{i-1} - 2y_i + y_{i+1} = h(\beta_0 f_{i-1} + \beta_1 f_i + \beta_2 f_{i+1}).$$

A következő módszert kapjuk: ($\beta_1 = 0, \beta_2 = \frac{1}{2}$ és $\beta_0 = -\frac{1}{2}$, nem 0-stabil)

$$y_{i+1} = 2y_i - y_{i-1} + \frac{1}{2}\,h\,(f_{i+1} - f_{i-1}).$$

5-15. Adjuk meg α és β értékét úgy, hogy a következő 3-lépéses módszer negyedrendű legyen.

$$y_{i+1} + \alpha(y_i - y_{i-1}) - y_{i-2} = h\beta(f_{i-1} + f_i)$$

Mutassuk meg, hogy nem 0-stabil.

5-16. A következő LTM-k közül melyek 0-stabilak?

a) $y_{i+1} + y_i - 2y_{i-1} = h(f_{i+1} + f_i + f_{i-1})$

b) $y_{i+1} - y_{i-1} = \frac{1}{3}\,h\,(f_{i+1} + 4f_i + f_{i-1})$

c) $y_{i+1} - y_i = \frac{1}{2}\,h\,(3f_i - f_{i-1})$

d) $y_{i+1} - y_i = \frac{1}{12}\,h\,(5f_{i+1} + 8f_i - f_{i-1})$

5-17. Készítsük el az x_i, x_{i-1} és x_{i+1} egyenletes felosztású pontokhoz tartozó legfeljebb másodfokú interpolációs polinomot (P), majd deriváljuk. Mutassuk meg, hogy ha $y \in C^3[0;1]$, akkor

$$y'(x_{i+1}) = P'(x_{i+1}) + O(h^2) = \frac{1}{2h}(3y(x_{i+1}) - 4y(x_i) + y(x_{i-1}) + O(h^2)).$$

A kapott módszer:

$$3y_{i+1} - 4y_i + y_{i-1} = 2hf_{i+1}.$$

5-18. Készítsünk interpoláció segítségével többlépéses módszert, ha x_0, x_1, x_2 alappontok és y_0, y_1, y_2 a megadott függvényértékek. A kapott polinomot deriváljuk és értékeljük ki az x_1 pontban, vagyis az y_1' közelítését kapjuk. Rendezzük át y_2-re, majd készítsük el a numerikus módszer rekurzióját.

5-19. Készítsünk Fejér-Hermite interpoláció segítségével többlépéses módszert, ha x_0, x_1 alappontok és y_0, y_0', y_1, y_1' a megadott függvény és deriváltértékek. A kapott polinomot értékeljük ki az x_2 pontban, vagyis az y_2 közelítését kapjuk. Ebből készítsük el a numerikus módszer rekurzióját.

5-20. Vizsgáljuk meg a következő LTM módszer konzisztenciáját, stabilitását, konvergenciáját! (f elég sima függvény.)

$$y_{i+4} - y_i = \frac{1}{3}h(8f_{i+3} - 4f_{i+2} + 8f_i)$$

5-21. Vizsgáljuk meg a következő LTM módszer konzisztenciáját, stabilitását, konvergenciáját! (f elég sima függvény.)

$$y_{i+2} + y_{i+1} - 2y_i = \frac{1}{2}h(5f_{i+1} + f_i)$$

Számítsuk ki az y_i megoldást rögzített $x_* = x_i = ih$ helyen az $f = 0$, $y_0 = 1$, $y_1 = 1 + h$ speciális esetben! Hogyan viselkedik ez a megoldás $i \to \infty$ esetén?

5-22. Vizsgáljuk meg a következő LTM módszer konzisztenciáját, stabilitását, konvergenciáját! (f elég sima függvény.)

$$y_{i+2} - 6y_{i+1} + 5y_i = -4hf_i$$

Számítsuk ki az y_i megoldást rögzített $x_* = x_i = ih$ helyen az $f = 0$, $y_0 = 1$, $y_1 = 1 + h$ speciális esetben! Hogyan viselkedik ez a megoldás $i \to \infty$ esetén?

5-23. Vizsgáljuk meg a következő LTM módszer konzisztenciáját, stabilitását, konvergenciáját! (f elég sima függvény.)

$$y_{i+2} - 6y_{i+1} + 5y_i = -h(f_{i+1} + 3f_i)$$

Számítsuk ki az y_i megoldást rögzített $x_* = x_i = ih$ helyen az $f = 0$, $y_0 = 1$, $y_1 = 1 + h$ speciális esetben! Hogyan viselkedik ez a megoldás $i \to \infty$ esetén?

5-24. Vizsgáljuk meg a következő LTM módszer konzisztenciáját, stabilitását, konvergenciáját! (f elég sima függvény.)

$$y_{i+2} + 4y_{i+1} - 5y_i = h(4f_{i+1} + 2f_i)$$

Számítsuk ki az y_i megoldást rögzített $x_* = x_i = ih$ helyen az $f = 0$, $y_0 = 1$, $y_1 = 1 + h$ speciális esetben! Hogyan viselkedik ez a megoldás $i \to \infty$ esetén?

5-25. Vizsgáljuk a következő Milne-Simpson módszer

$$y_{i+1} - y_{i-1} = \frac{1}{3}h(f_{i+1} + 4f_i + f_{i-1})$$

konzisztenciáját, stabilitását! Javasoljunk alkalmas egylépéses módszert a módszer beindítására! Milyen numerikus viselkedés várható $f = qy$ esetben?

5.6. 0-stabil többlépéses módszerek maximális rendje

5-7. T *(Dahlquist ekvivalencia tétele)*
Tekintsünk egy ℓ-lépéses LTM-et, mely konzisztens és f teljesíti a Lipschitz-feltételt. Ekkor a 0-stabilitás szükséges és elégséges feltétele a konvergenciának. Továbbá, ha $y \in C^{p+1}$ és a lokális hiba $O(h^p)$, akkor a módszer globális hibája is $O(h^p)$.

5-8. T *(Dahlquist 1. tétele)*
Tegyük fel, hogy az ℓ-lépéses LTM konzisztens ($\varrho(1) = 0$, $\varrho'(1) = \sigma(1)$) és ϱ teljesíti a gyökfeltételt.

a) *Ha a módszer explicit, akkor a rendje maximálisan ℓ.*

b) *Ha a módszer implicit, akkor a rendje maximálisan*

$$\begin{cases} \ell + 1, & \text{ha } \ell \text{ páratlan} \\ \ell + 2, & \text{ha } \ell \text{ páros} \end{cases}$$

c) *Az ℓ páros esetben az $\ell + 2$-rendű módszer minden gyöke rajta van az egységkörön.*

Bizonyítás. A bizonyítást lásd [10] 94. oldalán. ■

Megjegyzések.

1. A 0-stabilitás gyökfeltétele ugyan korlátozza a lehetséges módszerek osztályát, viszont nem biztosítja azt, hogy a módszer az alkalmazás szempontjából is jó legyen. Lásd a középpont szabály esetét.

2. Várható volt, hogy az explicit és implicit módszerek között különbség adódik a stabilitás szempontjából, hiszen implicit esetben interpolációról, explicit esetben extrapolációról van szó.

5-12. Példa. A Dahlquist tétel értelmében, ha $\ell = 1$, akkor a 0-stabil módszer rendje nem lehet 2-nél nagyobb. A trapéz-módszer 0-stabil és másodrendű.

5-13. Példa. A kétlépéses

$$y_{i+1} - y_{i-1} = \frac{1}{3} h \left(f_{i+1} + 4f_i + f_{i-1} \right)$$

módszer 0-stabil, mert az első stabilitási polinomja $\varrho(z) = z^2 - 1$ és ennek gyökei: $z_{1,2} = \pm 1$, így tejesíti a gyökfeltételt. Megmutatható, hogy a rendje 4. Ez a legmagasabb rend, ami a Dahlquist tétel értelmében előállítható.

5-14. Példa. A következő 3-lépéses módszer rendje 6.

$$11y_{i+1} + 27y_i - 27y_{i-1} - 11y_{i-2} = 3h \left(f_{i+1} + 9f_i + 9f_{i-1} + f_{i-2} \right)$$

A Dahlquist tétellel is bizonyítható (de a gyökfeltétellel is), hogy nem 0-stabil.

6. fejezet

Implicit Runge–Kutta-módszerek (IRK)

6.1. IRK-módszerek

A későbbiekben az abszolút stabilitási tartomány és a merev rendszerek tárgyalásakor látni fogjuk, hogy az implicit Runge–Kutta (IRK)-módszereknek fontos szerepük van. Tetszőlegesen magas rendű A-stabil képletek konstruálhatók. Az s lépcsőszámú módszer általános alakja:

$$k_1 = k_1(x_i, y_i, h) := f\left(x_i + ha_1, y_i + h\sum_{l=1}^{s} b_{1l}k_l\right),$$

$$\dots$$

$$k_j = k_j(x_i, y_i, h) := f\left(x_i + ha_j, y_i + h\sum_{l=1}^{s} b_{jl}k_l\right), \quad j = 1, \dots, s,$$

majd ezeket a c_j súlyokkal kombinálva kapjuk az új y_{i+1} értéket:

$$y_{i+1} = y_i + h\sum_{j=1}^{s} c_j k_j.$$

Az a_j, b_{jl} és c_j számok jellemzik a módszert, ezek függetlenek f, y és h-tól. Továbbá

$$\sum_{j=1}^{s} c_j = 1 \quad \text{és} \quad a_j = \sum_{l=1}^{s} b_{jl}.$$

Ha az eredeti differenciálegyenlet-rendszer n egyenletből áll, akkor a k_j vektorok számítása $n \cdot s$ méretű nemlineáris egyenletrendszer megoldását követelné minden lépésben. Emiatt ezen az úton nem jutunk a merev rendszerek hatékony megoldásához. Módosításuk azonban elvezet a gyakorlatban előnyös eljárásokhoz, melyek kitüntetett stabilitási tulajdonságokkal rendelkeznek.

6.2. IRK-módszerek konstrukciója

Nézzük az $s = 1$ lépcsőszámú IRK módszerek levezetését. Formailag a következőképpen adhatók meg:

$$k_1 = f(x_i + ha_1, y_i + hb_{11}k_1),$$
$$y_{i+1} = y_i + hc_1k_1,$$

ahol a_1, b_{11} és c_1 paraméterek. Ezeket úgy szeretnénk megadni, hogy a módszer maximális rendig pontos legyen.

A továbbiakban feltesszük az ERK módszereknél is használt $a_1 = b_{11}$ összefüggést. A lokális hiba Taylor-sorba fejtése h szerint a k_1-re felírt implicit egyenlet miatt problémásabb. A továbbiakban az egyszerűbb írásmód kedvéért elhagyjuk az $(x_i, y(x_i))$ argumentum feltüntetését.

$$\begin{aligned} k_1 &= f(x_i + a_1 h, y(x_i) + h b_{11} k_1) = \\ &= f + a_1 h (f_x + k_1 f_y) + \\ &+ \frac{1}{2} a_1^2 h^2 (f_{xx} + 2k_1 f_{xy} + k_1^2 f_{yy}) + O(h^3) \end{aligned}$$

Összevetve

$$k_1 = d_0 + d_1 h + d_2 h^2 + O(h^3)$$

sorfejtésével kapjuk, hogy

$$\begin{aligned} d_0 + d_1 h + d_2 h^2 + O(h^3) = f + a_1 h (f_x + (d_0 + d_1 h) f_y) + \\ + \frac{1}{2} a_1^2 h^2 (f_{xx} + 2d_0 f_{xy} + d_0^2 f_{yy}) + O(h^3). \end{aligned}$$

Innen h együtthatóit egyeztetve kapjuk a sorfejtés együtthatóit.

$$\begin{aligned} d_0 &= f \\ d_1 &= a_1 (f_x + d_0 f_y) = a_1 (f_x + f f_y) = a_1 F_1 f \\ d_2 &= a_1 d_1 f_y + \frac{1}{2} a_1^2 (f_{xx} + 2 f f_{xy} + f^2 f_{yy}) = a_1^2 \left(f_y F_1 f + \frac{1}{2} F_2 f \right) \end{aligned}$$

Felírva a lokális hibát

$$\begin{aligned} h g_i &= y(x_{i+1}) - y(x_i) - h c_1 k_1 = \\ &= h f + \frac{1}{2} h^2 F_1 f + \frac{1}{6} h^3 (F_2 f + f_y F_1 f) - \\ &- h c_1 f - h^2 c_1 a_1 F_1 f - h^3 a_1^2 c_1 \left(f_y F_1 f + \frac{1}{2} F_2 f \right) + O(h^4) = \\ &= (1 - c_1) h f + \left(\frac{1}{2} - a_1 c_1 \right) h^2 F_1 f + \\ &+ \left[\left(\frac{1}{6} - a_1^2 c_1 \right) f_y F_f + \left(\frac{1}{6} - a_1^2 c_1 \right) F_2 f \right] h^3 + O(h^4) \end{aligned}$$

Ha $c_1 = 1$ és $a_1 = \frac{1}{2}$, akkor h és h^2 együtthatói eltűnnek, míg h^3 együtthatója

$$-\frac{1}{12} f_y F_f + \frac{1}{24} F_2 f$$

lesz. Így az IRK módszer rendje 2, képletei:

$$\begin{aligned} k_1 &= f\left(x_i + \frac{1}{2} h, y_i + \frac{1}{2} h k_1 \right), \\ y_{i+1} &= y_i + h k_1 \end{aligned}$$

Butcher-mátrixa

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

$p = 2,\ s = 1$

Az $s = 2$ lépcsőszámú IRK képlet a következő alakú:

$$\begin{aligned} k_1 &= f(x_i + a_1 h, y_i + h b_{11} k_1 + h b_{12} k_2), \\ k_2 &= f(x_i + a_2 h, y_i + h b_{21} k_1 + h b_{22} k_2), \\ y_{i+1} &= y_i + h(c_1 k_1 + c_2 k_2), \end{aligned}$$

ahol $a_1 = b_{11} + b_{12}$ és $a_2 = b_{21} + b_{22}$. Részletes levezetését lásd [10] 112. oldalán. Ahhoz, hogy legalább harmadrendű módszert kapjunk a következő egyenletrendszert kell megoldanunk.

$$\begin{aligned} c_1 + c_2 &= 1 \\ c_1(b_{11} + b_{12}) + c_2(b_{21} + b_{22}) &= \frac{1}{2} \\ c_1(b_{11} + b_{12})^2 + c_2(b_{21} + b_{22})^2 &= \frac{1}{3} \\ c_1[b_{11}(b_{11} + b_{12}) + b_{12}(b_{21} + b_{22})] + c_2[b_{21}(b_{11} + b_{12}) + b_{22}(b_{21} + b_{22})] &= \frac{1}{6} \end{aligned}$$

A $c_1 = c_2 = \frac{1}{2}$ esetén a következő $\mathbf{B}$ mátrixot kapjuk:

$$\mathbf{B} = \begin{bmatrix} b & \frac{1}{2} \mp \frac{\sqrt{3}}{6} - b \\ \frac{1}{2} \pm \frac{\sqrt{3}}{6} - b & b \end{bmatrix},$$

stabilitási függvénye (lásd később)

$$F(z) = \frac{1 - (2b-1)z + (\frac{1}{3} - b)z^2}{1 - 2bz - (b - \frac{1}{6}z^2)}.$$

Speciális esetek:

a) $b = \frac{1}{4}$, ekkor

$$\mathbf{B} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \mp \frac{\sqrt{3}}{6} \\ \frac{1}{4} \pm \frac{\sqrt{3}}{6} & \frac{1}{4} \end{bmatrix}$$

$$F(z) = \frac{1 + \frac{1}{2}z + \frac{1}{12}z^2}{1 - \frac{1}{2}z + \frac{1}{12}z^2} = \Pi_{22}(z), \quad \Pi_{22}(z) - e^z = O(z^5).$$

Ez egy a Butcher által bevezetett Gauss-képletek közül, valójában negyedrendű módszer. A kapott módszer A-stabil, de nem erősen A-stabil, sem L-stabil, emiatt merev rendszerekkel kapcsolatban nem érdekes.

b) $b = \frac{1}{3}$, ekkor

$$\mathbf{B} = \begin{bmatrix} \frac{1}{3} & \frac{1}{6}(1 \mp \sqrt{3}) \\ \frac{1}{6}(1 \pm \sqrt{3}) & \frac{1}{3} \end{bmatrix}$$

$$F(z) = \frac{1 + \frac{1}{3}z}{1 - \frac{2}{3}z + \frac{1}{6}z^2} = \Pi_{12}(z).$$

A képlet L-stabil és harmadrendű.

c) $b = \frac{1}{2} \mp \frac{\sqrt{3}}{6}$, ekkor

$$\mathbf{B} = \begin{bmatrix} \frac{1}{2} \mp \frac{\sqrt{3}}{6} & 0 \\ -\frac{\sqrt{3}}{3} & \frac{1}{2} \mp \frac{\sqrt{3}}{6} \end{bmatrix}$$

$$F(z) = \frac{1 \pm \frac{\sqrt{3}}{3}z + (-\frac{1}{6} \pm \frac{\sqrt{3}}{6})z^2}{1 - (1 \mp \frac{\sqrt{3}}{3})z + (\frac{1}{3} \mp \frac{\sqrt{3}}{6})z^2}.$$

A stabilitási függvény nem az exponenciális függvény valamelyik Padé approximációja. (Azok P_r és Q_s polinomjai mindig racionális együtthatójúak.) A b előjelétől függően erősen különböznek a módszer stabilitási tulajdonságai. A kapott eljárás harmadrendű diagonálisan implicit Runge-Kutta módszer. A Butcher mátrix alsó háromszögű és diagonális elemei azonosak. Ez a módszer használata során fellépő nemlineáris egyenletek megoldása során lesz előnyös.

További kétlépcsős IRK-módszerek:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|cc} \frac{3}{4} & \frac{1}{4} & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

trapéz, $p = s = 2$ | harmadrendű, $p = 3, s = 2$ | optimális, $p = 4, s = 2$

6-1. Definíció. A módszert diagonálisan implicit Runge-Kutta módszernek nevezzük (angol rövidítése DIRK), ha $\mathbf{B}$ mátrixa alsóháromszög mátrix és diagonális elemei azonosak.

6.3. IRK-módszerek konvergenciája

Nézzük meg egy kétlépcsős módszeren, mely feltételekkel biztosítható az x_{i+1} pontbeli belső iteráció konvergenciája. Vezessük be a következő jelöléseket:

$$\mathbf{k} := \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}, \quad \mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2,$$

ahol

$$\mathbf{F}(\mathbf{k}) = \begin{bmatrix} f(x_i + a_1 h, y_i + h b_{11} k_1 + h b_{12} k_2) \\ f(x_i + a_2 h, y_i + h b_{21} k_1 + h b_{22} k_2) \end{bmatrix}$$

A $\mathbf{k}$ értékek számítása a $\mathbf{k_{j+1}} = \mathbf{F}(\mathbf{k_j})$ fixpontiterációval történik. Ennek konvergenciája azon múlik, hogy $\mathbf{F}$ kontrakció-e. Az $\mathbf{F}$ sorfejtése:

$$\mathbf{F}(\mathbf{k} + \triangle\mathbf{k}) \approx \mathbf{F}(\mathbf{k}) + \mathbf{F}'(\mathbf{k}) \cdot \triangle\mathbf{k}.$$

$$\|\mathbf{F}(\mathbf{k} + \triangle\mathbf{k}) - \mathbf{F}(\mathbf{k})\| \approx \|\mathbf{F}(\mathbf{k}) + \mathbf{F}'(\mathbf{k}) \cdot \triangle\mathbf{k} - \mathbf{F}(\mathbf{k})\| = \|\mathbf{F}'(\mathbf{k}) \cdot \triangle\mathbf{k}\| \leq \underbrace{\|\mathbf{F}'(\mathbf{k})\|}_{=q<1} \cdot \|\triangle\mathbf{k}\|$$

$\mathbf{F}'(\mathbf{k})$ a módszer Butcher-mátrixa segítségével felírható:

$$\mathbf{F}'(\mathbf{k}) = h f_y \mathbf{B}.$$

Itt f_y a Lipschitz konstanssal becsülhető és így

$$\|\mathbf{F}'(\mathbf{k})\| \leq h L_f \|\mathbf{B}\|.$$

Tehát a kontrakció feltétele

$$h L_f \|\mathbf{B}\| < 1 \quad \Leftrightarrow \quad h < \frac{1}{L_f \|\mathbf{B}\|}.$$

6-1. Példa. Az optimális kétlépcsős negyedrendű IRK-módszer esetén

$$\|\mathbf{B}\|_\infty = \frac{1}{4} + \frac{3 + 2\sqrt{3}}{12} = \frac{3 + \sqrt{3}}{6} \quad \Rightarrow \quad h < \frac{6}{(3 + \sqrt{3}) L_f}.$$

6.4. Rosenbrock-módszerek

Reakció-kinetikai differenciálegyenletekkel való foglalkozásakor Rosenbrock arra az ötletre jutott, hogy az IRK-módszereket az $\mathbf{f_y}$ Jacobi-mátrix segítségével fogalmazza meg. IRK-módszerek esetén a nemlineáris egyenlet megoldásánál a Newton-módszer használatakor $\mathbf{f_y}$-ra úgyis szükség van rá. Így jutott olyan módszercsaládhoz, mely iteráció mentes, egyszerűen programozható (az ERK-programot lényegében egy LU-felbontással kell kiegészíteni). Általános elismerést vívott ki magának a merev rendszerek sikeres kezelésével. A következőkben ezeket mutatjuk be arra az esetre, amikor f független x-től. Ezeket autonóm rendszereknek hívjuk. Minden differenciálegyenlet felírható ilyen alakban a következő módon:

$$y' = f(x,y) \quad \Leftrightarrow \quad \begin{bmatrix} y \\ x \end{bmatrix}' = \begin{bmatrix} f(x,y) \\ 1 \end{bmatrix}.$$

Mivel merev rendszereket konstans lépésköz esetén csak implicit módszerekkel lehet hatékonyan megoldani, ezért a következő IRK-képletekből indulunk ki:

$$\mathbf{y_{i+1}} = \mathbf{y_i} + h\sum_{j=1}^{s} c_j \mathbf{k_j},$$

$$\mathbf{k_j} = f\left(\mathbf{y_i} + h\sum_{l=1}^{j-1} b_{jl}\mathbf{k_l} + h\gamma\mathbf{k_j}\right).$$

$\mathbf{k_j}$ képletét helyettesítsük a Taylor-sorfejtése elejével:

$$\mathbf{k_j} = f\left(\mathbf{y_i} + h\sum_{l=1}^{j-1} b_{jl}\mathbf{k_l}\right) + h\gamma\mathbf{J_i}\mathbf{k_j},$$

ahol $\mathbf{J_i} = \mathbf{J}(\mathbf{y_i}) = \mathbf{f_y}(\mathbf{y_i}) \in \mathbb{R}^{n\times n}$ a Jacobi-mátrix. A $h\gamma\mathbf{J_i}\mathbf{k_j}$ vektort balra rendezve megkapjuk a Rosenbrock-módszer $\mathbf{k}$-képletét.

$$(\mathbf{I} - h\gamma\mathbf{J_i})\mathbf{k_j} = f\left(\mathbf{y_i} + h\sum_{l=1}^{j-1} b_{jl}\mathbf{k_l}\right)$$

Ezt a módszert később Wanner úgy módosította, hogy a jobb oldalon további szabad paraméterekkel rendelkező tagokat tett hozzá, mert így elérhető, hogy magasabbrendű A-stabil módszereket is be lehessen vezetni. Emiatt **ROW-módszerek**ről is beszélnek. A $\mathbf{k}$-képlet ezután:

$$(\mathbf{I} - h\gamma\mathbf{J_i})\mathbf{k_j} = f\left(\mathbf{y_i} + h\sum_{l=1}^{j-1} b_{jl}\mathbf{k_l}\right) + h\mathbf{J_i}\sum_{l=1}^{j-1}\gamma_{jl}\mathbf{k_l} \quad (j = 1,\ldots,s).$$

Itt csak lineáris egyenletrendszereket kell megoldanunk, amelyek mátrixa nem változik j-vel. Egy LU-felbontással és s visszahelyettesítéssel megkapjuk az új $\mathbf{y}$ értéket. A műveletigény tovább csökkenthető, ha a Jacobi-mátrixot nem minden lépésben, hanem csak néha számítjuk ki (mint a módosított Newton-módszernél). Ez utóbbi nem tűnik a gyakorlatban túl sikeresnek.

Az előző képletet a számítógépes megvalósítás szempontjából átalakítjuk, hogy a telt $\mathbf{J_i}$ mátrixot ne szorozzuk feleslegesen a $h\gamma$ számmal. Osszuk γ-val az egyenletet és h-t rendezzük a k_l-ek mellé.

$$\left(\frac{1}{h\gamma}\mathbf{I} - \mathbf{J_i}\right) h\mathbf{k_j} = \frac{1}{\gamma} f\left(\mathbf{y_i} + \sum_{l=1}^{j-1} b_{jl} h\mathbf{k_l}\right) + \mathbf{J_i}\sum_{l=1}^{j-1}\frac{\gamma_{jl}}{\gamma} h\mathbf{k_l}$$

Vezessünk be új ismeretleneket $\widetilde{\mathbf{k_j}} := h\mathbf{k_j}$ és $\mathbf{K_{j-1}} := \sum_{l=1}^{j-1} \frac{\gamma_{jl}}{\gamma} \widetilde{\mathbf{k_l}}$:

$$\left(\frac{1}{h\gamma}\mathbf{I} - \mathbf{J_i}\right)\widetilde{\mathbf{k_j}} - \mathbf{J_i}\mathbf{K_{j-1}} = \frac{1}{\gamma} f\left(\mathbf{y_i} + \sum_{l=1}^{j-1} b_{jl}\widetilde{\mathbf{k_l}}\right)$$

$$\left(\frac{1}{h\gamma}\mathbf{I} - \mathbf{J_i}\right)\widetilde{\mathbf{k_j}} - \mathbf{J_i}\mathbf{K_{j-1}} + \frac{1}{h\gamma}\mathbf{K_{j-1}} = \frac{1}{\gamma} f\left(\mathbf{y_i} + \sum_{l=1}^{j-1} b_{jl}\widetilde{\mathbf{k_l}}\right) + \frac{1}{h\gamma}\mathbf{K_{j-1}}$$

$$\left(\frac{1}{h\gamma}\mathbf{I} - \mathbf{J_i}\right)(\widetilde{\mathbf{k_j}} + \mathbf{K_{j-1}}) = \frac{1}{\gamma} f\left(\mathbf{y_i} + \sum_{l=1}^{j-1} b_{jl}\widetilde{\mathbf{k_l}}\right) + \frac{1}{h\gamma}\mathbf{K_{j-1}} \quad (j = 1, \ldots, s).$$

A továbblépéshez az új képlet:

$$\mathbf{y_{i+1}} = \mathbf{y_i} + \sum_{j=1}^{s} c_j \widetilde{\mathbf{k_j}}.$$

Kaps és Rentrop konstruált egy egyszerű struktúrájú, hatékony és lépésköz kontrollal ellátott módszert, melyet

$$S = \frac{|\lambda_{max}|}{|\lambda_{min}|} > 10^7$$

merevségi hányadossal teszteltek. (Lásd merev rendszerek.) Az alábbi képletekből látszik, hogy ez két ROW-módszer beágyazott módszerként kezelve:

$$y_{i+1} = y_i + h\Phi_1(y_i; h), \quad \widehat{y}_{i+1} = y_i + h\Phi_2(y_i; h)$$

$$\Phi_1(y_i; h) = \sum_{j=1}^{3} c_j k_j, \quad \Phi_2(y_i; h) = \sum_{j=1}^{4} \widehat{c}_j k_j$$

$$k_j = f\left(y_i + h\sum_{l=1}^{4} b_{jl} k_l\right) + h\mathbf{J_i}\sum_{l=1}^{4} \gamma_{jl} k_l \ \ (j = 1, 2, 3, 4).$$

Az első módszer harmadrendű, a második negyedrendű. A hiányzó együtthatók:

$$\begin{aligned}
\gamma &= 0.220428410, \quad \gamma_{ii} = \gamma \ \ (i = 1, 2, 3, 4), \quad \gamma_{ij} = 0 \ \ (i < j) \\
\gamma_{21} &= 0.822867461, \\
\gamma_{31} &= 0.695700194, \ \ \gamma_{32} = 0, \\
\gamma_{41} &= 3.90481342, \ \ \gamma_{42} = 0, \ \ \gamma_{43} = 1 \\
b_{21} &= -0.554591416, \quad b_{ij} = 0 \ \ (i \leq j) \\
b_{31} &= 0.252787696, \ \ b_{32} = 1, \\
b_{41} &= \beta_{31}, \ \ b_{42} = b_{32}, \ \ b_{43} = 0, \\
\widehat{c}_1 &= 0.545211088, \ \ \widehat{c}_2 = 0.301486480, \\
\widehat{c}_3 &= 0.1770640668, \ \ \widehat{c}_2 = -0.0237622363, \\
c_1 &= -0.162871035, \ \ c_2 = 1.18215360, \ \ c_3 = -0.0192825995,
\end{aligned}$$

A beágyazott módszereknél alkalmazott lépéskontroll

$$h_{\text{új}} = 0.9h\sqrt[3]{\frac{\varepsilon h}{|\widehat{y}_{i+1} - y_{i+1}|}}.$$

Lásd bővebben a [8] 491. oldalán.

7. fejezet

Stabilitás

7.1. Belső, lényegi instabilitás

Tekintsük a következő instabil differenciálegyenletet a megadott kezdetiértékkel:

$$y'(x) = \lambda(y(x) - F(x)) + F'(x)$$
$$y(x_0) = y_0,$$

ahol $F(x)$ folytonosan differenciálható egy x_0-t tartalmazó intervallumon. A differenciálegyenletnek létezik explicit megoldása, mely a homogén egyenlet megoldásából és egy partikuláris megoldásból előállítható $y(x) = y_{hom}(x) + y_{part}(x)$ alakban. A homogén egyenlet megoldása: $y_{hom}(x) = Ce^{\lambda x}$, egy partikuláris megoldás: $y_{part}(x) = F(x)$, így a megoldás:

$$y(x) = (y_0 - F(x_0)) \exp(\lambda(x - x_0)) + F(x).$$

mely kielégíti a kezdeti feltételt. A speciális $y_0 = F(x_0)$ kezdeti feltétel esetén $y(x) = F(x)$, ezzel az exponenciális tag eltűnik. Vegyünk egy perturbált kezdőértéket és vizsgáljuk, hogyan változik a megoldás. Legyen $\widehat{y}_0 = F(x_0) + \varepsilon$, ahol $\varepsilon > 0$ kicsi érték, ekkor a megoldás

$$\widehat{y}(x) = \varepsilon \exp(\lambda(x - x_0)) + F(x).$$

Legyen $\lambda > 0$, ekkor az első tag exponenciálisan nő, vagyis az $\widehat{y}(x)$ perturbált megoldás egyre távolabb lesz az $y(x)$ pontos megoldástól. A megoldás tehát nagyon érzékeny a kezdeti feltétel változására. Az ilyen típusú feladatot rosszul kondicionáltnak nevezzük. Ez egy belső, lényegi (inherent) instabilitás. Az ilyen feladatok csak magas rendű, nagy pontosságú módszerekkel kezelhetők.

7-1. Példa. Tekintsük a következő feladatot:

$$y'(x) = 10\left(y(x) - \frac{x^2}{1+x^2}\right) + \frac{2x}{(1+x^2)^2}$$
$$y(x_0) = y_0 = 0,$$

melynek megoldása $y(x) = \frac{x^2}{1+x^2}$. A pontos megoldás az 1-hez közelít a $[0, 2.2]$ intervallumon, míg a klasszikus RK-módszerrel $h = 0.01$ lépésközzel kapott megoldás $-\infty$-hez. Érdemes programot írni ennek szemléltetésére.

7.2. Aszimptotikus stabilitás

Láttuk a korábbiakban, hogy az eddigi stabilitás fogalom nem volt elegendő, hogy jól használható módszereket kapjunk. A numerikus megoldás viselkedése gyakran eltérő, erre mutatunk példát.

7-2. Példa. Tekintsük a következő kezdetiértékproblémát:

$$y'(x) = -2xy^2(x)$$
$$y(0) = 1,$$

és alkalmazzuk a középpont szabályt $h = 0.1$ lépésközzel a $[0; 5]$ intervallumon. Írjunk rá programot! Tudjuk, hogy a középpont szabály konzisztens és 0-stabil, így konvergens, de a pontos megoldás lecsengő, míg a numerikus megoldás oszcillál. Éppen akkor, amikor a valódi megoldás lesimul.

A továbbiakban a következő lineáris teszt feladatra vizsgáljuk a módszerek viselkedését:

$$y'(x) = \lambda y(x)$$
$$y(0) = 1,$$

ahol $\lambda \in \mathbb{R}$ vagy $\lambda \in \mathbb{C}$. Ennek megoldása: $y(x) = e^{\lambda x}$. A további megfontolások azonban nemlineáris egyenletekre is érvényesek maradnak, mert lokálisan ezek is lineárisaknak tekinthetők. Kis h lépésköz esetén az approximáció viselkedése hasonló.

7-3. Példa. Alkalmazzuk a teszt feladatra az Euler-módszert és nézzük meg a viselkedését!

$$y_{i+1} = y_i + h(\lambda y_i) = (1 + h\lambda)y_i = (1 + h\lambda)^{i+1} y_0$$

a) A $\lambda > 0$ esetben a pontos megoldás monoton módon tart végtelenhez, vagyis a megoldás instabil. Ebben az esetben $1 + h\lambda > 1$, vagyis a numerikus megoldás szintén végtelenhez tart. Tehát az Euler-módszer viselkedése helyes lesz.

b) A $\lambda < 0$ esetben a megoldás 0-hoz tart. Ugyanezt a numerikus módszer csak $|1 + h\lambda| < 1$, vagyis $h < \frac{-2}{\lambda} = \frac{2}{|\lambda|}$ esetben teljesíti. Ráadásul $-1 < 1 + h\lambda \leq 0$ esetén, vagyis $h \geq \frac{-1}{\lambda} = \frac{1}{|\lambda|}$ esetben a numerikus megoldás oszcillál, míg a valódi nem.

7-4. Példa. Alkalmazzuk az RK4 klasszikus 4-edrendű Runge-Kutta módszert a teszt feladatra és vizsgáljuk annak viselkedését!

$$\begin{aligned}
k_1 &= \lambda y_i \\
k_2 &= \lambda\left(y_i + \frac{1}{2}hk_1\right) = \left(\lambda + \frac{1}{2}h\lambda^2\right) y_i \\
k_3 &= \lambda\left(y_i + \frac{1}{2}hk_2\right) = \left(\lambda + \frac{1}{2}h\lambda^2 + \frac{1}{4}h^2\lambda^3\right) y_i \\
k_4 &= \lambda(y_i + hk_3) = \left(\lambda + h\lambda^2 + \frac{1}{2}h^2\lambda^3 + \frac{1}{4}h^3\lambda^4\right) y_i \\
y_{i+1} &= y_i + \frac{1}{6}h(k_1 + 2k_2 + 2k_3 + k_4) = \\
&= y_i \left[1 + \frac{1}{6}h\lambda + \left(\frac{1}{3}h\lambda + \frac{1}{6}h^2\lambda^2\right) + \left(\frac{1}{3}h\lambda + \frac{1}{6}h^2\lambda^2 + \frac{1}{12}h^3\lambda^3\right) + \right. \\
&+ \left. \left(\frac{1}{6}h\lambda + \frac{1}{6}h^2\lambda^2 + \frac{1}{12}h^3\lambda^3 + \frac{1}{24}h^3\lambda^4\right)\right] = \\
&= \left(1 + h\lambda + \frac{1}{2}h^2\lambda^2 + \frac{1}{6}h^3\lambda^3 + \frac{1}{24}h^4\lambda^4\right) y_i
\end{aligned}$$

A $h\lambda$-tól függő polinomot jelöljük F-fel.

$$F(h\lambda) = 1 + h\lambda + \frac{1}{2}h^2\lambda^2 + \frac{1}{6}h^3\lambda^3 + \frac{1}{24}h^4\lambda^4 = \sum_{j=0}^{4} \frac{(h\lambda)^j}{j!},$$

ami nem más, mint az exponenciális függvény 4-edfokú Taylor-polinomja. Tehát kis h-kra $F(h\lambda)$ jól közelíti $e^{h\lambda}$-t. Az összes explicit p-szintű módszerre $p \leq 4$ esetén teljesül, hogy $F(h\lambda)$ az $e^{h\lambda}$ p-edfokú Taylor-polinomja lesz.

a) A $\lambda > 0$ esetben a pontos megoldás monoton módon tart végtelenhez, vagyis a megoldás instabil. Ebben az esetben $h\lambda > 0$, így $F(h\lambda) > 1$. Ez azt jelenti, hogy a numerikus megoldás szintén végtelenhez tart. Tehát a módszer viselkedése helyes lesz. Ez a kevésbé fontos eset, mert a gyakorlatban előforduló feladatok tartalmaznak exponenciálisan csökkenő komponenst.

b) A $\lambda < 0$ esetben a megoldás 0-hoz tart. A numerikus módszer pontosan $|F(h\lambda)| < 1$ esetén tart nullához. $\lambda < 0$ esetén $F(h\lambda) < 1$, (ugyanis $F(h\lambda)$ az exponencális függvény közelítése). Mivel $F(h\lambda)$ negyedfokú polinom $h\lambda$-ban, ezért $|F(h\lambda)| < 1$ nem fog minden negatív λ-ra teljesülni.

7-1. Definíció. A numerikus módszer aszimptotikusan stabil, ha $\lambda < 0$ esetben a teszt feladatra alkalmazva bármely $h > 0$ lépésköz esetén $(y_n) \to 0$, ha $n \to \infty$. A nemzetközi szakirodalomban A_0 stabilitásnak is nevezik.

7-2. Definíció. Ha egylépéses módszerre alkalmazzuk a tesztfeladatot, akkor az

$$y_{i+1} = F(h\lambda)\, y_i$$

sorozatot kapjuk. Az $F(h\lambda)$ függvényt stabilitási függvénynek nevezzük. A

$$B := \{\mu \in \mathbb{C} \mid |F(\mu)| < 1\}$$

halmazt a módszer abszolút stabilitási tartományának nevezzük.

7-3. Definíció. A 0-stabil módszert A-stabilnak nevezzük, ha stabilitási tartománya az egész baloldali félsíkot tartalmazza, azaz

$$\forall\, \mu: \ \ \mathrm{Re}\,(\mu) < 0 \ \ \Rightarrow \ \ \mu \in B.$$

Megjegyzések.

1. A tartomány a valós tengelyre szimmetrikus. Az abszolút stabilitási tartománynak a valós tengellyel vett metszetét stabilitási intervallumnak nevezzük. Ezzel a stabilitási tartomány nagyságáról kaphatunk információt.

2. A [6] 419.oldalán szép ábrákat találunk az egyes módszerek abszolút stabilitási tartományairól. A rend növelésével a stabilitási tartomány nő. Az $F(\mu)$ függvényt stabilitási függvénynek nevezzük. A definíció ekvivalens a [10] könyvben leírttal, hiszen pontosan olyan h lépésközökre kapunk korlátos sorozatot, ha $\mu = h\lambda$ benne van a stabilitási tartományban. Tehát a tartomány korlátozza a lépésköz választását. Az előző feladatokból kapott stabilitási függvények:
explicit Euler-módszer:

$$F(\mu) = 1 + \mu$$

A klasszikus negyedrendű RK4-módszerre:

$$F(\mu) = 1 + \mu + \frac{1}{2}\mu^2 + \frac{1}{6}\mu^3 + \frac{1}{24}\mu^4.$$

3. Az ERK-módszerek stabilitási intervallumai

s	1	2	3	4	5
intervallum	$(-2;0)$	$(-2;0)$	$(-2.51;0)$	$(-2.78;0)$	$(-3.21;0)$

7-5. Példa. Számítsuk ki a trapéz módszer stabilitási függvényét!

$$y_{i+1} = y_i + \frac{h}{2}\left(f(x_i, y_i) + f(x_{i+1}, y_{i+1})\right)$$

Megoldás. Helyettesítsük a módszerbe az $f(x, y) = \lambda y$ függvényt!

$$y_{i+1} = y_i + \frac{h}{2}\left(\lambda y_i + \lambda y_{i+1})\right)$$

Fejezzük ki y_{i+1}-et:

$$y_{i+1}\left(1 - \frac{h}{2}\lambda\right) = y_i\left(1 + \frac{h}{2}\lambda\right)$$

$$y_{i+1} = \frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}\, y_i = \frac{2 + h\lambda}{2 - h\lambda}\, y_i.$$

Tehát a stabilitási függvény:

$$F(\mu) = \frac{2 + \mu}{2 - \mu}.$$

A módszer stabilitási tartománya azon μ értékek halmaza, melyekre $|F(\mu)| < 1$.

$$\frac{|2 + \mu|}{|2 - \mu|} < 1 \quad \Leftrightarrow \quad |2 + \mu| < |2 - \mu|$$

A $\mu = a + b \cdot i$ algebrai alakra áttérve $(2 + a)^2 + b^2 < (2 - a)^2 + b^2$, ami $a < 0$ esetén teljesül. Tehát a $\mathrm{Re}\,(\mu) < 0$ feltételt kaptuk μ-re, vagyis a teljes baloldali félsík a stabilitási tartomány, tehát a módszerünk A-stabil. ■

7-6. Példa. Számítsuk ki a következő IRK módszer stabilitási függvényét!

$$k_1 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}hk_1\right),$$

$$y_{i+1} = y_i + hk_1$$

Megoldás. Helyettesítsük a módszerbe az $f(x, y) = \lambda y$ függvényt és fejezzük ki k_1-et!

$$k_1 = \lambda\left(y_i + \frac{1}{2}hk_1\right)$$

$$k_1\left(1 - \frac{h\lambda}{2}\right) = \lambda y_i \quad \rightarrow \quad k_1 = \frac{\lambda}{1 - \frac{h\lambda}{2}}\, y_i = \frac{2\lambda}{2 - h\lambda}\, y_i$$

Helyettesítsük be k_1 értékét a módszerbe:

$$y_{i+1} = y_i + hk_1 = y_i + \frac{2h\lambda}{2 - h\lambda}\, y_i = \left(1 + \frac{2h\lambda}{2 - h\lambda}\right) y_i =$$

$$= \left(\frac{2 - h\lambda + 2h\lambda}{2 - h\lambda}\right) y_i = \left(\frac{2 + h\lambda}{2 - h\lambda}\right) y_i.$$

Tehát a stabilitási függvénye

$$F(\mu) = \frac{2 + \mu}{2 - \mu}.$$

■

7-4. Definíció. 1. A 0-stabil módszert erősen A-stabilnak nevezzük, ha A-stabil és létezik $c < 1$ konstans, hogy

$$|F(\mu)| \to c \quad (\mu \to -\infty).$$

2. A 0-stabil módszert L-stabilnak nevezzük, ha A-stabil és

$$|F(\mu)| \to 0 \quad (\mu \to -\infty).$$

7-7. Példa. Nézzük meg többlépéses módszerek esetén is az abszolút stabilitási tartományt. Alkalmazzuk a teszt feladatra a többlépéses módszert.

Megoldás. A formulánk

$$\sum_{k=0}^{l} \alpha_k y_{i+k} = h \sum_{k=0}^{l} \beta_k \lambda y_{i+k} \quad \Leftrightarrow \quad \sum_{k=0}^{l} (\alpha_k - h\lambda\beta_k)\, y_{i+k} = 0$$

Egy l-edrendű homogén lineáris differencia egyenletet kaptunk, melynek megoldását a karakterisztikus egyenlet megoldásával kapjuk meg. A tanult stabilitási (karakterisztikus) polinomokat felhasználva $\mu \in \mathbb{C}$-re

$$\varphi(\mu) := \varrho(\mu) - h\lambda\sigma(\mu) = 0.$$

φ-t teljes stabilitási (karakterisztikus) polinomnak nevezzük.
Ha adott $z := h\lambda$-re teljesül a gyökfeltétel, azaz a φ polinom bármely μ_j gyökére

$$|\mu_j(z)| \leq 1 \text{ és a többszörös gyökökre } |\mu_j(z)| < 1,$$

akkor a tesztfeladaton a megoldás korlátos, így z az abszolút stabilitási tartomány eleme.
Ha minden j-re $|\mu_j(z)| < 1$, akkor a megoldás 0-hoz konvergál. ■

7-5. Definíció. Többlépéses módszerek esetén az abszolút stabilitási tartomány azon $z = h\lambda$ komplex számok halmaza, melyre a $\varphi(\mu) = \varrho(\mu) - z\sigma(\mu) = 0$ polinom gyökei az egységkör belsejében vannak.

A példák és a feladatok mutatják, hogy a stabilitási függvények az e^x polinomiális vagy racionális közelítései. Az e^x megfelelő approximációját adja az (r, s)-Padé-közelítése. Ez egy olyan $\Pi_{r,s}(z)$ racionális függvény, melyre

$$\Pi_{r,s}(z) := \frac{P_r(z)}{Q_s(z)} = e^z + O(z^{r+s+1}),$$

ahol $P_r(z)$ r-edfokú, $Q_s(z)$ s-edfokú polinom. P_r-nek és Q_s-nek összesen $r+s+2$ szabad paramétere van, de a normálás után már csak $r + s + 1$ marad az e^x $(r + s)$. Taylor-polinomjának meghatározásához.
Például a legismertebb módszerek esetén:

- explicit Euler-módszer: $r = 1$, $s = 0$
- implicit Euler-módszer: $r = 0$, $s = 1$
- trapéz-szabály: $r = s = 1$
- harmadrendű RK-módszer: $r = 3$, $s = 0$

Kérdés, hogy milyen feltételek esetén igaz minden $\mathrm{Re}\,(z) < 0$ komplex számra, hogy

$$\left|\frac{P_r(z)}{Q_s(z)}\right| < 1?$$

A következő tétel erre ad választ.

7-1. T *(Az exponenciális függvény Padé-approximációja)*
Az exponenciális függvény Padé-approximációja pontosan akkor rendelkezik a

$$\forall\, z:\ \mathrm{Re}\,(z) < 0 \quad \Rightarrow \quad \left|\frac{P_r(z)}{Q_s(z)}\right| < 1$$

tulajdonsággal, ha $s - 2 \le r \le s$.

Megjegyzés.
Az $s > r$ esetben $z \to -\infty$ esetén $F(z) = \frac{P_r(z)}{Q_s(z)} \to 0$, így az $s = r + 1$ és $s = r + 2$ esetek különösen előnyösek.

7-2. L *(Dalhquist 1. tétele: LTM A-stabilitásának szükséges feltétele)*
Ha a LTM A-stabil, akkor tetszőleges $|\mu| > 1$*-re*

$$\mathrm{Re}\left(\frac{\varrho(\mu)}{\sigma(\mu)}\right) > 0.$$

Bizonyítás. Mivel a módszer A-stabil, ezért a

$$\varphi(\mu) = \varrho(\mu) - z\sigma(\mu) = 0$$

karakterisztikus egyenlet bármely μ gyökére teljesül, hogy ha $\mathrm{Re}\,(z) \le 0$, akkor $|\mu| \le 1$. Innen

$$z = \frac{\varrho(\mu)}{\sigma(\mu)} \quad \to \quad \mathrm{Re}\,(z) = \mathrm{Re}\left(\frac{\varrho(\mu)}{\sigma(\mu)}\right) \le 0 \quad \Rightarrow \quad |\mu| \le 1.$$

Tegyük fel indirekt, hogy μ_0-ra $|\mu_0| > 1$ és

$$\mathrm{Re}\left(\frac{\varrho(\mu_0)}{\sigma(\mu_0)}\right) \le 0.$$

Ekkor $z = z_0 := \frac{\varrho(\mu_0)}{\sigma(\mu_0)}$ esetén μ_0 a karakterisztikus egyenlet gyöke, de $\mathrm{Re}\,(z_0) \le 0$ és $|\mu_0| > 1$, ami ellentmond az A-stabilitásnak. ■

7-3. L *(Dalhquist 2. tétele: A-stabil többlépéses módszerek jellemzése)*
a) *Explicit többlépéses módszer nem lehet A-stabil.*
b) *Implicit A-stabil többlépéses módszer rendje nem lehet 2-nél nagyobb. Ezek közül a trapéz formula normált hibakonstansa abszolút értékben a legkisebb,* $C_3^* = -\frac{1}{12}$.

Megjegyzések.

1. Ezek után ne gondoljunk arra, hogy nincs olyan explicit módszereken alapuló eljárás, amelyet hatékonyan alkalmazhatnánk merev rendszerek megoldására. Elegendő változó lépésközt megengedni. Akkor néhány kicsi, a stabilitási feltételeket kielégítő lépésköz után következhet egy nagy

lépés is. A kérdés csupán az, hogy a nagy lépéstávolságok megengedhetőek-e a pontosság szempontjából.

2. Explicit RK módszerek esetén az $F(z)$ stabilitási függvény s-edfokú polinom, emiatt nem teljesítheti $\mathrm{Re}\,(z) \leq 0$ esetén $|F(z)| \leq 1$-et. Tehát az explicit RK módszerek sem lehetnek A-stabilak.

3. Implicit A-stabil RK módszerek rendje viszont lehet 2-nél tetszőlegesen nagyobb. Stabilitásuk ellenőrzésére hasznos a következő eredmény.

7-4. L *(RK módszerek A-stabilitása)*
Legyen az RK módszer $F(z)$ stabilitási függvénye analitikus $\mathrm{Re}\,(z) < 0$-ra és teljesítse

$$|F(iy)| \leq 1 \ \textit{tetszőleges valós } y\textit{-ra.}$$

Ekkor a módszer A-stabil.

Feladatok

7-1. Írjuk fel az implicit Euler-módszer abszolút stabilitási tartományát!

$$F(\mu) = \frac{1}{1-\mu}$$

7-2. Írjuk fel az explicit harmadrendű RK3-módszer abszolút stabilitási tartományát!

$$F(\mu) = 1 + \mu + \frac{1}{2}\mu^2 + \frac{1}{6}\mu^3$$

7-3. Írjunk programot, mely konkrét módszerek esetén kirajzolja a stabilitási függvény ismeretében az abszolút stabilitási tartományt! Az $|F(\mu)| = 1$ egyenletet kell megoldani

$$F(\mu) = e^{i\theta}, \ \ 0 \leq \theta \leq 2\pi$$

és kirajzolni. Általában csak a valós tengely feletti részt szokás megjeleníteni a szimmetria miatt.

7-4. Határozzuk meg a kétlépéses (explicit és implicit) Adams-módszerek abszolút stabilitási tartományait!

7-5. Írjuk fel az általános RK-módszerek stabilitási függvényét!

$$F(\mu) = 1 + \mu \mathbf{c^T}(\mathbf{I} + \mu\mathbf{B})^{-1}\mathbf{e} = \frac{|\mathbf{I} - \mu\mathbf{B} + \mu\mathbf{e}\mathbf{c^T}|}{|\mathbf{I} - \mu\mathbf{B}|},$$

ahol $\mathbf{c}$ és $\mathbf{B}$ a Butcher táblázat megfelelő elemeit jelenti. A jobboldali alakhoz a Shermann-Morrison-formulát és a determinánsra vonatkozó összefüggést használtuk:

$$(\mathbf{A} + \mathbf{u}\mathbf{v^T})^{-1} = \mathbf{A^{-1}} - \frac{\mathbf{A^{-1}}\mathbf{u}\mathbf{v^T}\mathbf{A^{-1}}}{1 + \mathbf{v^T}\mathbf{A^{-1}}\mathbf{u}} \quad \text{és} \quad |\mathbf{A} + \mathbf{u}\mathbf{v^T}| = |\mathbf{A}|(1 + \mathbf{v^T}\mathbf{A^{-1}}\mathbf{u}).$$

7.3. Merev (stiff) differenciálegyenletek

Az alkalmazásokban fontos szerepet játszanak az olyan differenciálegyenletek, melyek tompított oszcillációkat és más, stacionárius állapotba való átmeneteket írnak le (ezek fizikailag a rendszer energiaveszteségére vezethetők vissza). Gyakran előfordul, hogy egyidejűleg olyan folyamatok illetve komponensek is vannak, melyek nagyon gyorsan stacionáriussá válnak és olyanok is, amelyek csak lassan teszik meg azt. Például az olyan - eléggé különböző reakciósebességű - vegyi folyamatok, mint például amelyek a városi szmogot előidézik, vagy az enzimreakciók az emberi testben. Nézzük a következő példát!

7-8. Példa. Tekintsük a következő homogén lineáris differenciálegyenletrendszert.

$$\begin{aligned} y_1' &= -0.5\,y_1 \quad +32.6\,y_2 + 35.7\,y_3 \\ y_2' &= \qquad\quad -48\,y_2 + 9\,y_3 \\ y_3' &= \qquad\quad\;\; 9\,y_2 - 72\,y_3 \\ y_1(0) &= 4 \\ y_2(0) &= 13 \\ y_3(0) &= 1 \end{aligned}$$

kezdeti feltételekkel. A rendszer $\mathbf{A}$ mátrixának

$$\lambda_1 = -0.5, \;\; \lambda_2 = -45, \;\; \lambda_3 = -75$$

sajátértékei segítségével felírható a megoldás:

$$\begin{aligned} y_1(x) &= 15\,e^{-0.5x} \quad -12\,e^{-45x} + e^{-75x} \\ y_2(x) &= \qquad\quad\;\; 12\,e^{-45x} + e^{-75x} \\ y_1(x) &= \qquad\quad\;\;\; 4\,e^{-45x} - 3\,e^{-75x}, \end{aligned}$$

mely kielégíti a kezdeti feltételeket. Alkalmazzuk a negyedrendű RK-módszert a feladatra. Látjuk, hogy a megoldás egyes komponensei más-más mértékben csökkennek. A módszer abszolút stabilitási intervalluma $(-2.78; 0)$, nézzük milyen lépésköz felel meg:

$$\begin{aligned} -2.78 < h\cdot(-75) \quad &\Rightarrow \quad h < \frac{2.78}{75} \approx 0.0371 \\ -2.78 < h\cdot(-45) \quad &\Rightarrow \quad h < \frac{2.78}{45} \approx 0.0618, \\ -2.78 < h\cdot(-0.5) \quad &\Rightarrow \quad h < \frac{2.78}{0.5} \approx 5.56. \end{aligned}$$

Csak a legkisebb lépésközt választhatnánk $h < 0.0371$-t, vagyis a leggyorsabban fogyó e^{-75x}-hez kell választanunk a h lépésközt, hogy ne lépjünk ki a stabilitási intervallumból. A példán megmutatjuk, hogy a lépésköz megfelelő változtatásával elérhető nagyobb lépésköz is magasabb pontosság mellett.

Ahhoz, hogy az e^{-75x} komponenst 4 jegy pontossággal megkapjuk $h_1 = 0.0025$ lépésközre van szükségünk. Ezt az $F(-75h_1)$-nek (az exp függvény 4-edfokú Taylor polinomja a $-75h_1$ helyen) az 5 jegyre pontos közelítésével kaptuk:

$$\frac{(-75\cdot 0.0025)^5}{5!} \approx -0.2\cdot 10^5.$$

$x = 0$-ból $h_1 = 0.0025$ lépésközzel 60 lépést teszünk meg, így az $x = 0.0025\cdot 60 = 0.15$-höz értünk. Hasonlítsuk össze a két nagy abszolútértékű sajátértékhez tartozó komponens értékét.

$$e^{-75\cdot 0.15} = e^{-11.25} \approx -0.000013 \ll e^{-45\cdot 0.15} = e^{-6.75} \approx -0.00171$$

Látjuk, hogy az e^{-75x} komponens lecsengett, növelhetjük a lépésközt.
A $h_2 = 0.005$ lépésközzel az e^{-45x}-nek az 5 jegyre pontos közelítést kaptuk:

$$\frac{(-45 \cdot 0.005)^5}{5!} \approx -0.5 \cdot 10^5.$$

Ezzel a lépésközzel 30 lépést teszünk meg $x = 0.15 + 30 \cdot 0.005 = 0.3$.
Hasonlítsuk össze a két kisebb abszolútértékű sajátértékhez tartozó komponens értékét.

$$e^{-45 \cdot 0.3} = e^{-13.5} \approx -0.0000014 \ll e^{-0.5 \cdot 0.3} = e^{-0.15} \approx -0.8607$$

Látjuk, hogy az $e^{-0.5x}$ függvény pontossága nem elegendő, ezt a $h_3 = 0.4$ lépésközzel érnénk el:

$$\frac{(-0.5 \cdot 0.4)^5}{5!} \approx -0.26 \cdot 10^5.$$

Ezzel az értékkel megszegjük a stabilitási feltételt, $h < 0.0371$-t.
Ha maradunk a még engedélyezett $h_3 = 0.035$-nél, akkor $x = 1$-et 200 lépéssel érjük el. Ha a megengedettnél nagyobbat választunk, akkor az RK-módszer instabil lesz. Lásd a [6] könyv ábráját a 419. oldalon az ERK-módszerek stabilitási tartományairól.

7-6. Definíció. A lineáris differenciálegyenlet rendszert merevnek nevezzük, ha mátrixának

a) léteznek olyan λ_i sajátértékei, melyekre $\mathrm{Re}\,(\lambda_i) \ll 0$, tehát $\varrho(\mathbf{A}) \gg 1$,

b) léteznek $|\lambda_j| \ll |\lambda_i|$ sajátértékei, tehát $cond(\mathbf{A}) = |\lambda_{\max}|/|\lambda_{\min}| \gg 1$ (vagy $\mathbf{A}$ szinguláris),

c) nincsenek nagy pozitív valós résszel rendelkező sajátértékek és nincsenek nagy képzetes résszel rendelkező sajátértékek (kivéve amikor $\mathrm{Re}\,(\lambda_i) \ll 0$).

A merevség mértékét az

$$S = \frac{\max |\mathrm{Re}\,(\lambda_j)|}{\min |\mathrm{Re}\,(\lambda_j)|}$$

hányadossal adjuk meg.

Megjegyzések.

1. Az előző példában ez a hányados $S = 150$, ami azt jelenti, hogy ez nem is nagyon merev rendszer. A gyakori értéke 10^3 és 10^6 közötti. A lépésköz választása azoknál a módszereknél problémás, ahol az abszolút stabilitási tartomány nem a $\mathrm{Re}\,(z) < 0$ félsík.

2. Nemlineáris rendszer esetén az

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}(x)), \quad \mathbf{y}(x) \in \mathbb{R}^n$$

egyenlet megoldásának lokális viselkedését az x_k közelében az $\mathbf{y}(x_k) = \mathbf{y_k}$ kezdeti feltétellel vizsgáljuk. Legyen

$$\mathbf{y}(x) = \mathbf{y_k} + \mathbf{z}(x), \quad x_k \le x \le x_k + h$$

és tegyük fel, hogy h és $\|\mathbf{z}(x)\|$ kicsi. Az egyenletet linearizálva a következő lineáris feladatot kapjuk:

$$\mathbf{z}'(x) \approx \mathbf{J}(x_k)(\mathbf{y}(x) - \mathbf{y_k}) + \mathbf{f_k} + \mathbf{g_k}(x - x_k) = \mathbf{J}(x_k)\mathbf{z}(x) + \mathbf{f_k} + (x - x_k)\mathbf{g_k},$$

ahol $\mathbf{J}$ az $\mathbf{f}$ $\mathbf{y}$ szerinti Jacobi mátrixa és $\mathbf{g_k}$ az $\mathbf{f}$ x szerinti deriváltja.

$$\mathbf{J}(x_k) = \left(\frac{\partial f_i}{\partial y_j}\right)_{i,j=1}^{n}, \quad \mathbf{g_k} = \left(\frac{\partial f_i}{\partial x_i}\right)_{i=1}^{n}$$

Ezért nemlineáris esetben a merevséget a Jacobi mátrix sajátértékeivel definiáljuk.

3. Ha ilyen rendszereket akarunk numerikusan megoldani, akkor több probléma is van. Lehetséges, hogy a numerikus megoldás nem is tart 0-hoz, hanem oszcillál vagy divergál. A módszer 0-stabilitása nem garantálja a numerikus megoldás helyes viselkedését rögzített h-ra (lásd középpont szabály példája). Mindössze a korlátosságot garantálja a 0-stabilitás.

4. Ha a folyamat kezdeti (nem stacionárius) részét pontosan akarjuk végigszámítani, akkor RK vagy LTM módszerrel tehetjük, de mindkét esetben szükséges egy $hL_f < 1$ feltétel betartása a pontosság érdekében. ($L_f = \|\mathbf{A}\| \geq |\lambda_{\max}|$ nagy.) A stacionárius (lassan változó) állapot elérése után a merevség („stiffness") jelensége lép fel. Ha alkalmatlan a módszerünk, akkor a $hL_f < 1$ feltétel még mindig szükséges, de most inkább stabilitási okokból és kevésbé a pontosság miatt, hiszen a $|\lambda_{\max}|$-tól függő komponensek már régen jelentéktelenek a megoldásban. A feltétel megszegése esetén a numerikus módszer instabillá válik. Így kénytelenek vagyunk kis lépésközt választani, jóllehet a megoldás olyan lassan változik, hogy lényegesen nagyobb lépés is lehetséges. Lásd a bevezető példát.

5. A sajátértékek alapján adott meghatározásunk az olyan feladatoknál problémás, ahol az $\mathbf{A}$ mátrix függ t-től. Emiatt azt is mondhatjuk, hogy a merevség leginkább a számítás során derül ki.

Feladatok

7-6. Döntsük el, hogy vajon merev-e a következő differenciálegyenlet?

$$\dot{y} = -(10^5 \exp(-10^4 t) + 1) \cdot (y - 1), \quad t \in [0; 1] \quad y(0) = 0$$

A megoldás:

$$y(t) = \exp(10(-10^4 t) - 1) \cdot \exp(-t) + 1.$$

7-7. Döntsük el, hogy vajon merev-e a következő differenciálegyenlet-rendszer?

$$\begin{aligned} \dot{y}_1 &= -0.5\, y_1 + 0.501\, y_2, & y_1(0) &= 1.1 \\ \dot{y}_2 &= 0.501\, y_1 - 0.5\, y_2, & y_2(0) &= -0.9 \end{aligned}$$

7-8. Döntsük el, hogy vajon merev-e a következő differenciálegyenlet-rendszer?

$$\dot{\mathbf{y}} = \mathbf{A}(t)\mathbf{y}, \quad t \in [0; 1]$$

$$\mathbf{A} = \begin{bmatrix} -1 + 100\cos(200t) & 100(1 - \sin(200t)) & 0 \\ -100(1 + \sin(200t)) & -(1 + 100\cos(200t)) & 0 \\ 1200(\cos(100t) + \sin(100t)) & 1200(\cos(100t) - \sin(100t)) & -501 \end{bmatrix}$$

8. fejezet

Közönséges differenciálegyenletek peremérték feladatai

8.1. Peremérték feladatok

Az

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \;\; 0 \le x \le L$$

alakú differenciálegyenlet-rendszerrel kapcsolatban gyakran olyan feladatok lépnek fel, ahol az $x = 0$ pontban az $\mathbf{y}(0)$ kezdővektorból csak j komponens ismert ($0 < j < n$), de ezenkívül az intervallum végpontjában $n - j$ komponens adott az $\mathbf{y}(L)$ vektorból. Két ilyen típusú példát mutatunk.

8-1. Példa. A célba lövés: Legyen y a golyó magassága a föld felett, $x(t)$ a távolság a cél felé a t. időpillanatban, $L > 0$ a cél távolsága és g a Föld gravitációs gyorsulása.

Matematikai modellünk a következő: a golyó x irányban konstans $u \neq 0$ sebességgel repül, y irányban csak a gravitáció hat rá. Mozgását a következő egyenletek írják le:

$$\dot{x} = u, \;\; \ddot{y} = -g,$$

ahol $\dot{x} := \frac{dx}{dt}$ a sebesség, $\ddot{y} := \frac{d^2y}{dt^2}$ a gyorsulás. Az $x(0) = 0$ kezdeti feltételt figyelembe véve az $x(t) = ut$ megoldást kapjuk, melynek segítségével az y-ra feírt egyenletet átalakíthatjuk:

$$\begin{aligned} \dot{y} &= \frac{dy}{dx}\dot{x} \\ -g = \ddot{y} &= \frac{d^2y}{dx^2}\dot{x}^2 + \frac{dy}{dx}\underbrace{\ddot{x}}_{=0} = y''u^2. \end{aligned}$$

Ezután a következő peremérték feladattal állunk szemben:

$$\begin{aligned} y''(x) &= -\frac{g}{u^2} \;\; (0 \le x \le L) \\ y(0) = 0, &\;\; y(L) = 0. \end{aligned}$$

A baloldali $y(0) = 0$ peremfeltétel segítségével készíthetjük el a feladat analitikus megoldását egy c szabad paraméterrel:

$$y(x) = -\frac{g}{2u^2}x^2 + cx.$$

A paraméter értékét megkapjuk a jobboldali peremfeltétellel:

$$y(L) = -\frac{g}{2u^2}L^2 + cL = 0,$$

tehát $c = \frac{gL}{2u^2}$. Ekkor a megoldás

$$y(x) = -\frac{g}{2u^2}x^2 + \frac{gL}{2u^2}x = \frac{gx}{u^2}(L - x).$$

Ennek következménye, hogy

$$y'(x) = -\frac{g}{u^2}x + \frac{gL}{2u^2}$$
$$y'(0) = \frac{gL}{2u^2} = tg(\alpha)$$

szög alatt kell lőnünk, hogy a célba találjunk. Éppen ezt a szöget nem ismertük, különben megoldhattuk volna azonnal az

$$y'' = -\frac{g}{u^2}$$
$$y(0) = 0, \;\; y'(0) = \frac{gL}{2u^2}$$

kezdetiérték feladatot. Az $u \neq 0$ feltételről kiderül, hogy feltétele a megoldhatóságnak. $u = 0$ esetén nincs megoldás. Fizikai tapasztalatunkkal szemben viszont minden $u > 0$ esetén van megoldás. Az ellentmondás feloldásához a légellenállást is figyelembe vevő nemlineáris egyenletrendszerből kellene kiindulnunk. Az ebből eredő peremérték feladat megoldásához már numerikus módszerre lenne szükség.

8-2. Példa. Vegyi reaktort modellezünk a következő egyenlettel:

$$Dc'' - vc' + f(x, c) = 0, \;\; 0 \leq x \leq L,$$

ahol $D > 0$ a diffúziós állandó, $v \leq 0$ az áramló közeg sebessége, f a reakció kinetikáját írja le, c a reakcióban keletkező anyag koncentrációja. Legtöbbször f nemlineáris c-ben és nem függ közvetlenül x-től. A reaktor elején, ahol az anyagok beadagolása történik, még 0 a c koncentráció ($c(0) = 0$), a kimenetnél viszont feltesszük, hogy a koncentráció már nem változik ($c'(L) = 0$). A következő elsőrendű rendszerre térhetünk át, bevezetve a c függvény deriváltját egy új ismeretlennel ($y_1 := c, y_2 := c'$-t):

$$y_1' = y_2$$
$$y_2' = \frac{1}{D}(vy_2 - f(x, y_1))$$
$$y_1(0) = 0, \;\; y_2(L) = 0.$$

Ha f nemlineáris c-ben, akkor a feladat analitikus megoldását nem tudjuk megadni. Ha f lineáris c-ben, pl. $f(x, c) = rc$, ahol r konstans, akkor a megoldás az exponenciális függvények segítségével írható le. Írjuk fel az alapmátrixot.

1. eset: Vizsgáljuk az egyszeres sajátértékek esetét, tegyük fel, hogy $v^2 \neq 4rD$, ekkor a differenciálegyenlet-rendszer mátrixa

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\frac{r}{D} & \frac{v}{D} \end{bmatrix}.$$

A karakterisztikus polinomja:

$$p(\lambda) = \det\left(\begin{bmatrix} -\lambda & 1 \\ -\frac{r}{D} & \frac{v}{D} - \lambda \end{bmatrix}\right) = \lambda^2 - \frac{v}{D}\lambda + \frac{r}{D} = 0,$$

a sajátértékei (különbözők), sajátvektorai:

$$\lambda_{1,2} = \frac{1}{2D}(v \pm \sqrt{v^2 - 4rD}), \quad \mathbf{z_i} = \begin{bmatrix} 1 \\ \lambda_i \end{bmatrix}, \ (i = 1, 2).$$

$$\mathbf{A} \cdot \mathbf{z_i} = \begin{bmatrix} 0 & 1 \\ -\frac{r}{D} & \frac{v}{D} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \lambda_i \end{bmatrix} = \begin{bmatrix} \lambda_i \\ -\frac{r}{D} + \lambda_i \frac{v}{D} \end{bmatrix} = \begin{bmatrix} \lambda_i \\ \lambda_i^2 \end{bmatrix} = \lambda_i \cdot \begin{bmatrix} 1 \\ \lambda_i \end{bmatrix},$$

mivel a karakterisztikus polinomból $-\frac{r}{D} + \lambda_i \frac{v}{D} = \lambda_i^2$.
Legyen

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{bmatrix} \quad \to \quad \mathbf{Z^{-1}} = \frac{1}{\lambda_2 - \lambda_1} \begin{bmatrix} \lambda_2 & -1 \\ -\lambda_1 & 1 \end{bmatrix},$$

ekkor az alapmátrix

$$\begin{aligned} \mathbf{Y}(x) = \mathbf{Z}e^{\mathbf{\Lambda} x}\mathbf{Z^{-1}} &= \frac{1}{\lambda_2 - \lambda_1} \begin{bmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{bmatrix} \begin{bmatrix} e^{\lambda_1 x} & 0 \\ 0 & e^{\lambda_2 x} \end{bmatrix} \begin{bmatrix} \lambda_2 & -1 \\ -\lambda_1 & 1 \end{bmatrix} \\ &= \frac{1}{\lambda_2 - \lambda_1} \begin{bmatrix} \lambda_2 e^{\lambda_1 x} - \lambda_1 e^{\lambda_2 x} & e^{\lambda_2 x} - e^{\lambda_1 x} \\ \lambda_1 \lambda_2 (e^{\lambda_1 x} - e^{\lambda_2 x}) & \lambda_2 e^{\lambda_2 x} - \lambda_1 e^{\lambda_1 x} \end{bmatrix}. \end{aligned}$$

2. eset: Vizsgáljuk a többszörös sajátérték esetét, legyen $v^2 = 4rD$, ekkor a differeciálegyenletrendszer mátrixa

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\frac{r}{D} & 2\sqrt{\frac{r}{D}} \end{bmatrix}.$$

A karakterisztikus polinomja:

$$p(\lambda) = \det\left(\begin{bmatrix} -\lambda & 1 \\ -\frac{r}{D} & 2\sqrt{\frac{r}{D}} - \lambda \end{bmatrix} \right) = \lambda^2 - 2\sqrt{\frac{r}{D}}\lambda + \frac{r}{D} = 0,$$

a sajátértékei:

$$\lambda_{1,2} = \sqrt{\frac{r}{D}} = \frac{v}{2D} =: \lambda.$$

A differenciálegyenletünk

$$Dc'' - 2\sqrt{rD}c' + rc = 0.$$

Ennek $c_1(x) = e^{\lambda x}$, $c_2(x) = xe^{\lambda x}$ megoldásaiból felírható az általános megoldás:

$$c(x) = \alpha_1 c_1(x) + \alpha_2 c_2(x) = \alpha_1 e^{\lambda x} + \alpha_2 x e^{\lambda x} = e^{\lambda x} \begin{bmatrix} 1 & x \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}.$$

Mivel $c = y_1$ és $c' = y_2$, így

$$\mathbf{y}(x) = \begin{bmatrix} c(x) \\ c'(x) \end{bmatrix} = \begin{bmatrix} \alpha_1 e^{\lambda x} + \alpha_2 x e^{\lambda x} \\ \lambda \alpha_1 e^{\lambda x} + \alpha_2 e^{\lambda x}(1 + \lambda x) \end{bmatrix} = e^{\lambda x} \begin{bmatrix} 1 & x \\ \lambda & 1 + \lambda x \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}.$$

Innen az alapmátrix

$$\mathbf{Y}(x) = \begin{bmatrix} \mathbf{y^{(1)}} & \mathbf{y^{(2)}} \end{bmatrix} = e^{\lambda x} \begin{bmatrix} 1 & x \\ \lambda & 1 + \lambda x \end{bmatrix} \begin{bmatrix} \alpha_1^{(1)} & \alpha_1^{(2)} \\ \alpha_2^{(1)} & \alpha_2^{(2)} \end{bmatrix} = e^{\lambda x} \begin{bmatrix} 1 & x \\ \lambda & 1 + \lambda x \end{bmatrix} \mathbf{C}.$$

Ha $x = 0$, akkor

$$\mathbf{Y}(0) = \mathbf{I} = \begin{bmatrix} 1 & 0 \\ \lambda & 1 \end{bmatrix} \mathbf{C} \quad \rightarrow \quad \mathbf{C} = \begin{bmatrix} 1 & 0 \\ \lambda & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ -\lambda & 1 \end{bmatrix}.$$

Tehát

$$\mathbf{Y}(x) = e^{\lambda x} \begin{bmatrix} 1 & x \\ \lambda & 1+\lambda x \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\lambda & 1 \end{bmatrix} = e^{\lambda x} \begin{bmatrix} 1-\lambda x & x \\ -\lambda^2 x & 1+\lambda x \end{bmatrix}.$$

Ha az 1. esetben felírt képletben a $\lambda_{1,2} \rightarrow \lambda$ határátmenetet végrehajtjuk, akkor

$$\frac{e^{\lambda_2 x} - e^{\lambda_1 x}}{\lambda_2 - \lambda_1} \rightarrow x e^{\lambda x},$$

vagyis ugyanezt a képletet kapjuk.

8.2. Fredholm alternatíva tétel

Az előző példában a megoldás előállításához szükségünk volt a differenciálegyenlet-rendszer mátrixának sajátértékeire, sajátvektoraira, vagyis a diagonalizálására. Az alapmátrixot egyszerűbb előállítani, hogy megoldjuk a következő n db kezdetiérték feladatot, ekkor az $\mathbf{A}$ mátrix x-től is függhet. Ebben az esetben nincs szükségünk a diagonalizálhatóságra, elég az a simasági feltétel, hogy $\mathbf{A}$ elemei pl. folytonosak legyenek.

$$\begin{aligned} \frac{d\mathbf{y^{(i)}}}{dx} &= \mathbf{A}(x)\mathbf{y^{(i)}}, \;\; 0 \le x \le 1 \\ \mathbf{y^{(i)}}(0) &= \mathbf{e_i} \end{aligned}$$

Vizsgáljuk most az inhomogén differenciálegyenlet-rendszert

$$\mathbf{y}'(x) = \mathbf{A}(x)\mathbf{y}(x) + \mathbf{f}(x), \;\; 0 \le x \le 1.$$

A vegyi reaktor modellje felírható ilyen alakban, ha $f(x,c) = rc + g(x)$, ekkor

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \;\; \mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\frac{r}{D} & \frac{v}{D} \end{bmatrix}, \;\; \mathbf{f} = \begin{bmatrix} 0 \\ -\frac{g}{D} \end{bmatrix}.$$

Az inhomogén rendszer egy partikuláris megoldása egy további kezdetiérték feladat megoldásával kapható meg:

$$\begin{aligned} \mathbf{z}' &= \mathbf{A}\mathbf{z} + \mathbf{f}, \;\; 0 \le x \le 1 \\ \mathbf{z}(0) &= \mathbf{0}. \end{aligned}$$

Ezután az inhomogén rendszer általános megoldása

$$\mathbf{y}(x) = \mathbf{z}(x) + \mathbf{Y}(x)\mathbf{q}$$

alakban írható fel, ahol a konstans $\mathbf{q}$ vektor adja az $\mathbf{y}$ még ismeretlen kezdetiértékeit. Mivel $\mathbf{q}$-ban n db szabad paraméterünk van, összesen n db feltételt adhatunk meg. Vizsgáljuk a lineáris peremfeltételek esetét a $[0;1]$ intervallumon. A feltételek alakja:

$$\mathbf{B_0}\mathbf{y}(0) + \mathbf{B_1}\mathbf{y}(1) = \mathbf{p},$$

ahol $\mathbf{p} \in \mathbb{R}^n$ adott vektor és $\mathbf{B_0}, \mathbf{B_1} \in \mathbb{R}^{n \times n}$ adott mátrixok. A peremfeltételt szeparáltnak nevezzük, ha

$$\mathbf{B_0} = \begin{bmatrix} \mathbf{B} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{B_1} = \begin{bmatrix} \mathbf{0} \\ \mathbf{C} \end{bmatrix},$$

ahol $1 \leq m < n$-re

$$\mathbf{B} \in \mathbb{R}^{m\times n},\ \ \mathbf{C} \in \mathbb{R}^{(n-m)\times n},\ \ \mathbf{p} = \begin{bmatrix} \mathbf{p_B} \\ \mathbf{p_C} \end{bmatrix},\ \ \mathbf{p_B} \in \mathbb{R}^m,\ \ \mathbf{p_C} \in \mathbb{R}^{n-m}.$$

Ekkor a peremfeltételek elkülöníthetők:

$$\mathbf{B}\mathbf{y}(0) = \mathbf{p_B},\ \ \mathbf{C}\mathbf{y}(1) = \mathbf{p_C}.$$

Célszerű még feltennünk, hogy a $\mathbf{B}$ és $\mathbf{C}$ mátrixok maximális rangúak. A vegyi reaktor peremfeltételei ilyenek

$$\mathbf{B_0} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},\ \ \mathbf{B_1} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},\ \ \mathbf{p} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

8-1. T *(Fredholm-alternatíva tétel)*
Legyenek a differenciálegyenlet-rendszer $\mathbf{A}$ *mátrixának elemei az* x *folytonos függvényei. Ekkor a*

$$\mathbf{y}'(x) = \mathbf{A}(x)\mathbf{y}(x) + \mathbf{f}(x),\ \ 0 \leq x \leq 1$$
$$\mathbf{B_0}\mathbf{y}(0) + \mathbf{B_1}\mathbf{y}(1) = \mathbf{p}$$

peremérték feladatnak pontosan akkor van egyértelmű megoldása (tetszőleges $\mathbf{p}$ *vektor és tetszőleges folytonos komponensű* $\mathbf{f}$ *esetén), ha a*

$$\mathbf{G} = \mathbf{B_0}\mathbf{Y}(0) + \mathbf{B_1}\mathbf{Y}(1)$$

mátrix reguláris.
Jelöljük $\mathbf{z}$*-vel az inhomogén egyenlet egy partikuláris megoldását. Ha* $\mathbf{G}$ *szinguláris, akkor*

$$\mathbf{p} - \mathbf{B_1}\mathbf{z}(1) \in \operatorname{Im}_G$$

esetén végtelen sok megoldás van, máskülönben nincs megoldás.

Bizonyítás. Az inhomogén rendszer lineáris peremfeltétellel történő megoldásához a $\mathbf{q}$ vektor komponenseit kell meghatározni. Helyettesítsük be az általános megoldást a peremfeltételbe

$$\mathbf{p} = \mathbf{B_0}\mathbf{y}(0) + \mathbf{B_1}\mathbf{y}(1) = \mathbf{B_0}[\underbrace{\mathbf{z}(0)}_{=\mathbf{0}} + \mathbf{Y}(0)\mathbf{q}] + \mathbf{B_1}\left[\mathbf{z}(1) + \mathbf{Y}(1)\mathbf{q}\right] =$$
$$= \mathbf{B_0}\mathbf{Y}(0)\mathbf{q} + \mathbf{B_1}\left[\mathbf{z}(1) + \mathbf{Y}(1)\mathbf{q}\right] = \mathbf{B_1}\mathbf{z}(1) + \left[\mathbf{B_0}\mathbf{Y}(0) + \mathbf{B_1}\mathbf{Y}(1)\right]\mathbf{q}$$

és rendezzük át

$$\mathbf{p} - \mathbf{B_1}\mathbf{z}(1) = \left[\mathbf{B_0}\mathbf{Y}(0) + \mathbf{B_1}\mathbf{Y}(1)\right]\mathbf{q} = \mathbf{G}\mathbf{q}.$$

Tehát a következő lineáris egyenletrendszert kell megoldanunk:

$$\mathbf{G}\mathbf{q} = \mathbf{p} - \mathbf{B_1}\mathbf{z}(1), \quad \mathbf{G} := \mathbf{B_0}\mathbf{Y}(0) + \mathbf{B_1}\mathbf{Y}(1),$$

ennek megoldhatóságán múlik a peremérték feladat megoldhatósága.

Ha $\mathbf{G}$ reguláris, akkor $\mathbf{q} = \mathbf{G^{-1}}(\mathbf{p} - \mathbf{B_1}\mathbf{z}(1))$ és $\mathbf{y} = \mathbf{z} + \mathbf{Y}\mathbf{q}$.

Ha $\mathbf{G}$ szinguláris, akkor $\mathbf{p} - \mathbf{B_1}\mathbf{z}(1) \in \operatorname{Im}_G$ esetén $\mathbf{y} = \mathbf{z} + \mathbf{Y}(\mathbf{q} + \mathbf{s})$, ahol $\mathbf{q}$ a lineáris egyenletrendszer egy megoldása és $\mathbf{s} \in Ker_G$, vagyis $\mathbf{G}$ magteréből való. ■

Megjegyzések.

1. Az egyértelműség csak az $\mathbf{A}, \mathbf{B_0}, \mathbf{B_1}$ mátrixoktól függ, $\mathbf{f}$-től és $\mathbf{p}$-től nem. (Az $\mathbf{A}, \mathbf{B_0}, \mathbf{B_1}$ mátrixok meghatározzák a rendszer belső fizikai természetét, $\mathbf{f}$ és $\mathbf{p}$ input adatok.)

2. Ha a szeparált peremfeltétel esetén $\mathbf{B}$ és $\mathbf{C}$ rangja nem maximális, akkor $\mathbf{BY}(0)$ és $\mathbf{CY}(1)$ rangja sem az, így $\mathbf{G}$ szinguláris.

3. A tétel tartalmazza az $n = 1$ esetet is. Elsőrendű egyenletre csak kivételes esetben lehet olyan (egyértelműen megoldható) peremérték feladatot megfogalmazni, amely két peremfeltétel és tetszőleges peremérték is szerepel, pl. $y' = \sqrt{y}$, $y(0) = 0$, $y(1) = p$ és $|p| \leq \frac{\pi}{4}$. Ilyen feladatokkal nem foglalkozunk.

4. Figyeljünk a peremérték feladat és kezdetiérték feladat lényeges különbségeire. Az előző tétel szerint algebrai tulajdonságon múlik a peremérték feladat megoldhatósága, az adatok simasága itt nem segít. A feltételeink mellett a differenciálegyenlet-rendszer jobb oldala y-ban Lipschitz-folytonos és ez biztosítja a kezdetiérték feladat megoldásának létezését. Így lehetséges, hogy a peremérték feladatnak nincs megoldása, amikor a kezdetiérték feladatnak van. Ezenkívül a kezdetiérték feladathoz képest elég műveletigényes a számítása, mert $n + 1$ db kezdetiérték feladatot és egy lineáris egyenletrendszert kell megoldanunk.

8-3. Példa. Nézzük a vegyi reaktor példáján a tétel által leírt eseteket, a peremfeltétel legyen a korábbi

$$\mathbf{B_0 y}(0) + \mathbf{B_1 y}(1) = \mathbf{p} = \mathbf{0}, \text{ ahol } \mathbf{Y}(0) = \mathbf{I}.$$

1. eset: $v^2 = 4rD$, láttuk, hogy

$$\mathbf{Y}(x) = e^{\lambda x} \begin{bmatrix} 1 - \lambda x & x \\ -\lambda^2 x & 1 + \lambda x \end{bmatrix}, \quad \lambda = \frac{v}{2D}$$

és $\mathbf{G} = \mathbf{B_0 Y}(0) + \mathbf{B_1 Y}(1)$, innen

$$\mathbf{G} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + e^{\lambda} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 - \lambda & 1 \\ -\lambda^2 & 1 + \lambda \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\lambda^2 e^{\lambda} & (1 + \lambda) e^{\lambda} \end{bmatrix}.$$

Ekkor $\mathbf{G}$ reguláris.

2. eset: $v^2 \neq 4rD$, ekkor

$$\mathbf{Y}(x) = \frac{1}{\lambda_2 - \lambda_1} \begin{bmatrix} \lambda_2 e^{\lambda_1 x} - \lambda_1 e^{\lambda_2 x} & e^{\lambda_2 x} - e^{\lambda_1 x} \\ \lambda_1 \lambda_2 (e^{\lambda_1 x} - e^{\lambda_2 x}) & \lambda_2 e^{\lambda_2 x} - \lambda_1 e^{\lambda_1 x} \end{bmatrix},$$

ahol

$$\lambda_{1,2} = \frac{1}{2D}(v \pm \sqrt{v^2 - 4rD})$$

és $\mathbf{G} = \mathbf{B_0 Y}(0) + \mathbf{B_1 Y}(1)$. Innen

$$\mathbf{G} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{1}{\lambda_2 - \lambda_1} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \lambda_2 e^{\lambda_1} - \lambda_1 e^{\lambda_2} & e^{\lambda_2} - e^{\lambda_1} \\ \lambda_1 \lambda_2 (e^{\lambda_1} - e^{\lambda_2}) & \lambda_2 e^{\lambda_2} - \lambda_1 e^{\lambda_1} \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 \\ \frac{\lambda_1 \lambda_2 (e^{\lambda_1} - e^{\lambda_2})}{\lambda_2 - \lambda_1} & \frac{\lambda_2 e^{\lambda_2} - \lambda_1 e^{\lambda_1}}{\lambda_2 - \lambda_1} \end{bmatrix}.$$

a) Ha $r \leq 0$, akkor a $\lambda_{1,2}$ értékek nem csak különböznek, hanem valósak is (a másodfokú egyenlet diszkriminánsa nem negatív), így $\mathbf{G}$ reguláris.

b) Ha $r > 0$ és $v = 0$, $D = 1$, ekkor a peremérték-problémánk

$$\mathbf{y}' = \begin{bmatrix} 0 & 1 \\ -r & 0 \end{bmatrix} \mathbf{y} + \begin{bmatrix} 0 \\ -g \end{bmatrix} = \mathbf{A}\mathbf{y} + \mathbf{f}$$
$$y_1(0) = 0, \;\; y_2(1) = 0.$$

Ekkor $\mathbf{A}$ sajátértékei: $\lambda_{1,2} = \pm i\sqrt{r}$.

$$\begin{aligned}
\lambda_1\lambda_2 &= r \\
\lambda_2 - \lambda_1 &= -i\sqrt{r} - i\sqrt{r} = -2i\sqrt{r}, \\
e^{\lambda_1} - e^{\lambda_2} &= (\cos(\sqrt{r}) + i\sin(\sqrt{r})) - (\cos(\sqrt{r}) - i\sin(\sqrt{r})) = 2i\sin(\sqrt{r}) \\
\Rightarrow &\frac{\lambda_1\lambda_2(e^{\lambda_1} - e^{\lambda_2})}{\lambda_2 - \lambda_1} = \frac{2ir\sin(\sqrt{r})}{-2i\sqrt{r}} = -\sqrt{r}\sin(\sqrt{r}) \\
\lambda_2 e^{\lambda_2} - \lambda_1 e^{\lambda_1} &= -i\sqrt{r}\left(\cos(\sqrt{r}) - i\sin(\sqrt{r})\right) - i\sqrt{r}\left(\cos(\sqrt{r}) + i\sin(\sqrt{r})\right) = -2i\sqrt{r}\cos(\sqrt{r}) \\
\Rightarrow &\frac{\lambda_2 e^{\lambda_2} - \lambda_1 e^{\lambda_1}}{\lambda_2 - \lambda_1} = \frac{-2i\sqrt{r}\cos(\sqrt{r})}{-2i\sqrt{r}} = \cos(\sqrt{r})
\end{aligned}$$

$$\mathbf{G} = \begin{bmatrix} 1 & 0 \\ -\sqrt{r}\sin(\sqrt{r}) & \cos(\sqrt{r}) \end{bmatrix}, \quad \det(\mathbf{G}) = \cos(\sqrt{r}) = 0 \quad \Leftrightarrow \quad r = \left(\frac{\pi}{2} + k\pi\right)^2, \;\; k = 0, 1, \ldots.$$

Tekintsük $k = 0$, $g \equiv 0$ esetben a következő peremérték-problémát:

$$c'' + \frac{\pi^2}{4}c = 0$$
$$c(0) = 0, \;\; c'(1) = 0,$$

Behelyettesítéssel ellenőrizhető, hogy ennek megoldása $c(x) = c_0 \sin\left(\frac{\pi}{2}x\right)$ illetve vektoros alakban

$$\mathbf{y}(x) = c_0 \begin{bmatrix} \sin\left(\frac{\pi}{2}x\right) \\ \frac{\pi}{2}\cos\left(\frac{\pi}{2}x\right) \end{bmatrix},$$

ahol c_0 tetszőleges konstans. Tehát itt végtelen sok megoldással találkozunk.
A tétel feltételeit vizsgálva $\mathbf{p} = \mathbf{0}, \mathbf{z} \equiv \mathbf{0}$ miatt

$$\mathbf{p} - \mathbf{B_1}\mathbf{z}(1) = \mathbf{0} \in \operatorname{Im} G.$$

A $k = 1$ esetben $r = \left(\frac{3}{2}\pi\right)^2$.

$$\mathbf{G} = \begin{bmatrix} 1 & 0 \\ -\sqrt{r}\sin(\sqrt{r}) & \cos(\sqrt{r}) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{3}{2}\pi & 0 \end{bmatrix}, \quad \Rightarrow \quad \operatorname{Im} G = \operatorname{Span}\left\{ \begin{bmatrix} 1 \\ \frac{3}{2}\pi \end{bmatrix} \right\}.$$

$$\mathbf{p} - \mathbf{B_1}\mathbf{z}(1) = \begin{bmatrix} 0 \\ -z_2(1) \end{bmatrix} \notin \operatorname{Im} G, \;\; \text{ha} \;\; z_2(1) \neq 0.$$

Látjuk, hogy ekkor megoldhatatlan feladatot kapunk. Egy konkrét ilyen feladat, mely teljesíti a $z_2(1) \neq 0$ feltételt:

$$\mathbf{y}' = \begin{bmatrix} 0 & 1 \\ -\frac{9}{4}\pi^2 & 0 \end{bmatrix} \mathbf{y} + \begin{bmatrix} 0 \\ 2 + \frac{9}{4}\pi^2 x^2 \end{bmatrix} = \mathbf{A}\mathbf{y} + \mathbf{f}$$
$$y_1(0) = 0, \;\; y_2(1) = 0.$$

Ekkor a $\mathbf{z}(0) = \mathbf{0}$ kezdetiértékhez tartozó inhomogén egyenlet partikuláris megoldása $\mathbf{z}(x) = \begin{bmatrix} x^2 \\ 2x \end{bmatrix}$ és teljesíti a $z_2(1) \neq 0$ feltételt. Ellenőrizzük:

$$\mathbf{z}' = \begin{bmatrix} 2x \\ 2 \end{bmatrix}$$
$$\mathbf{A}\mathbf{y} + \mathbf{f} = \begin{bmatrix} 0 & 1 \\ -\frac{9}{4}\pi^2 & 0 \end{bmatrix} \begin{bmatrix} x^2 \\ 2x \end{bmatrix} + \begin{bmatrix} 0 \\ 2 + \frac{9}{4}\pi^2\, x^2 \end{bmatrix} = \begin{bmatrix} 2x \\ 2 \end{bmatrix}$$
$$z_1(0) = 0, \;\; z_2(0) = 0 \;\text{ és }\; z_2(1) = 2.$$

8.3. A másodrendű egyenlet és klasszikus peremfeltételei

Vizsgáljuk a vegyi reaktor azon esetét, amikor a közeg nyugszik (sebessége: $v = 0$), a forrástag (f) csak x függvénye, a koncentrációt c helyett u-val jelöljük

$$Du'' + f(x) = 0, \quad 0 \leq x \leq 1,$$

$D > 0$ konstans továbbra is a diffúziós együttható. A kapott egyenlet úgy is interpretálható, hogy $u(x)$ egy rúd hőmérséklete az x pontban (a rúd hosszát 1-re normalizáljuk), $f(x)$ a hőforrások sűrűsége x-ben és D a hővezetési együttható.

Ehhez a **hővezetési egyenlet**hez a következő (szeparált) peremfeltételeket csatoljuk:

$$\alpha_1 u(0) - \beta_1 u'(0) = c_1, \quad \alpha_2 u(1) + \beta_2 u'(1) = c_2.$$

Ezek után a peremérték feladatot a következő lineáris differenciálegyenlet-rendszer alakjában írjuk fel:

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g}$$
$$\mathbf{y} := \begin{bmatrix} u(x) \\ u'(x) \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \mathbf{g} := \begin{bmatrix} 0 \\ -f(x)/D \end{bmatrix}$$
$$\mathbf{B_0 y}(0) + \mathbf{B_1 y}(1) = \mathbf{p}$$
$$\mathbf{p} := \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \quad \mathbf{B_0} = \begin{bmatrix} \alpha_1 & -\beta_1 \\ 0 & 0 \end{bmatrix} \quad \mathbf{B_1} := \begin{bmatrix} 0 & 0 \\ \alpha_2 & \beta_2 \end{bmatrix}$$

Az $\mathbf{A}$ mátrix sajátértékei $\lambda_{1,2} = 0$, így a 8-2. Példa 2. esetét felhasználva kapjuk az $\mathbf{Y}(x)$ alapmátrixot. A peremfeltételek $\mathbf{G}$ mátrixa

$$\mathbf{Y}(x) = \begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix}, \quad \mathbf{G} = \mathbf{B_0} + \mathbf{B_1}\mathbf{Y}(1) = \begin{bmatrix} \alpha_1 & -\beta_1 \\ \alpha_2 & \alpha_2 + \beta_2 \end{bmatrix}.$$

A peremfeltételek $\mathbf{q}$ vektora egyértelműen meghatározott, ha

$$\det(\mathbf{G}) = \alpha_1(\alpha_2 + \beta_2) + \alpha_2\beta_1 \neq 0.$$

Ez például a következő esetekben teljesül:

1. eset: ha $\alpha_1, \alpha_2 > 0$, $\beta_1 = \beta_2 = 0$ vagyis a peremfeltétel

$$u(0) = a, \;\; u(1) = b, \quad \text{ahol} \quad a := \frac{c_1}{\alpha_1}, \;\; b := \frac{c_2}{\alpha_2}.$$

Ezt **elsőfajú vagy Dirichlet-féle peremfeltételnek** nevezzük. Ha hővezetési egyenletnek interpretáljuk az egyenletet, akkor ez azt jelenti, hogy a peremen megadjuk az a és b hőmérsékletet. Diffúz feladat esetén azt jelenti, hogy megadjuk a koncentrációt a peremen.

2. eset: ha $\beta_1, \beta_2 > 0$, $\alpha_1, \alpha_2 > 0$, vagyis a peremfeltétel

$$\sigma_0 u(0) - \frac{du}{dx}(0) = a, \quad \sigma_1 u(1) + \frac{du}{dx}(1) = b,$$

ahol

$$\sigma_0 := \frac{\alpha_1}{\beta_1}, \;\; \sigma_1 := \frac{\alpha_2}{\beta_2}, \;\; a := \frac{c_1}{\beta_1}, \;\; b := \frac{c_2}{\beta_2},$$

ezt **vegyes illetve harmadfajú vagy Robin-féle peremfeltételnek** nevezzük. Harmadfajú peremfeltételekkel modellezhető a turbulens hőcsere a külvilággal. Diffúziós egyenlet esetén azt jelenti, hogy a külső fal részlegesen áteresztő.

3. eset: ha $\beta_1, \beta_2 > 0$, $\alpha_1, \alpha_2 = 0$, tehát a peremfeltétel

$$-\frac{du}{dx}(0) = a := \frac{c_1}{\beta_1}, \quad \frac{du}{dx}(1) = b := \frac{c_2}{\beta_2},$$

ezt **másodfajú vagy Neumann-féle peremfeltételnek** nevezzük. Ekkor $\det(\mathbf{G}) = 0$, így $\mathbf{q}$ nincs egyértelműen meghatározva. Ez a peremfeltétel azt adja vissza a hővezetési feladatban, hogy a külső fal hőszigetelt, diffúz feladatnál a koncentrációt nem engedi át.

Megjegyzések.

1. A harmadfajú peremfeltétel az első és másodfajút, mint határesetet tartalmazza. Fizikai szemléltetését lásd [10] 163. oldalán. Amennyiben (akár parciális) differenciálegyenlettel kapcsolatban nem tudjuk, hogy milyen peremfeltételeket kellene kitűzni, akkor írjuk fel a harmadfajúakat és igyekezzünk megszerezni annak σ és c együtthatóit.

2. Másodrendű egyenleteknél az itt vizsgált peremfeltételektől eltérő feltételek közül a periodikus peremfeltétel fontos gyakorlatban és elméletben is. Ezzel most nem foglalkozunk.

8.4. A peremérték feladatok kondicionáltsága

A peremérték feladat megoldása az alapmegoldáson keresztül elavultnak számít. Azzal, hogy a különféle $\mathbf{y}^{(\mathbf{i})}$ megoldásokat alkalmasan kombináljuk felléphetnek rosszul kondicionált műveletek, például amikor két igen közeli számot vonunk ki egymásból. Ennek szemléltetésére nézzük a következő feladatot.

8-4. Példa. Vizsgáljuk most a vegyi reaktor differenciálegyenlet-rendszerét a $[0;1]$ intervallumon

$$\mathbf{y}'(x) = \mathbf{A}(x)\mathbf{y}(x) + \mathbf{f}(x), \;\; 0 \le x \le 1,$$

ahol $g \equiv 0$, $r = -400 =: -\sigma^2$, $D = 1$, $v = 0$. Ez annak felel meg, hogy a vegyi reaktorban a c koncentrációjú anyag gyorsan szétesik. Tehát

$$\mathbf{y}' = \begin{bmatrix} 0 & 1 \\ \sigma^2 & 0 \end{bmatrix} \mathbf{y} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Leftrightarrow \quad \begin{bmatrix} c' \\ c'' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \sigma^2 & 0 \end{bmatrix} \begin{bmatrix} c \\ c' \end{bmatrix} \quad \Leftrightarrow \quad c''(x) = \sigma^2\, c(x),$$

a peremfeltételek legyenek

$$\begin{aligned} x = 0: \quad & -\frac{\sigma}{2}\, c(0) - c'(0) &&= \frac{\sigma}{2} \\ x = 1: \quad & \sigma\, c(1) + c'(1) &&= 0. \end{aligned}$$

Ezt mátrixos alakban felírva

$$\mathbf{B_0 y}(0) + \mathbf{B_1 y}(1) = \mathbf{p},$$

ahol

$$\mathbf{B_0} = \begin{bmatrix} -\frac{\sigma}{2} & -1 \\ 0 & 0 \end{bmatrix}, \ \mathbf{B_1} = \begin{bmatrix} 0 & 0 \\ \sigma & 1 \end{bmatrix}, \ \mathbf{p} = \begin{bmatrix} \frac{\sigma}{2} \\ 0 \end{bmatrix}.$$

Könnyen ellenőrizhető, hogy a peremfeltételeket is kielégítő pontos megoldás

$$c(x) = e^{-\sigma x}, \quad c'(x) = -\sigma e^{-\sigma x}.$$

Írjuk fel az $\mathbf{Y}$ alapmátrixot, majd ebből a peremfeltételnek eleget tevő megoldást a fejezet elején megadott módon. Mivel az inhomogén egyenlet $\mathbf{z}(0) = \mathbf{0}$ peremfeltételhez tartozó partikuláris megoldása $\mathbf{z}(x) \equiv 0$, ezért a lineáris egyenletrendszer jobboldala: $\mathbf{p} - \mathbf{B_1 z}(1) = \mathbf{p}$. A $\mathbf{Gq} = \mathbf{p}$ lineáris egyenletrendszerből számítjuk $\mathbf{q}$-t, mellyel a peremfeltételnek eleget tevő megoldás előállítható. A 8-3. Példa 2. esetbeli alapmátrixot felhasználva $\lambda_{1,2} = \pm\sigma$ sajátértékekkel a következő alapmátrixot kapjuk:

$$\begin{aligned} \lambda_2 - \lambda_1 &= -2\sigma \\ \mathbf{Y}(x)_{11} &= \frac{-\sigma\, e^{\sigma x} - \sigma\, e^{-\sigma x}}{-2\sigma} = \frac{e^{\sigma x} + e^{-\sigma x}}{2} = \operatorname{ch}(\sigma x) \\ \mathbf{Y}(x)_{12} &= \frac{e^{-\sigma x} - \sigma\, e^{\sigma x}}{-2\sigma} = \frac{1}{\sigma}\frac{e^{\sigma x} - e^{-\sigma x}}{2} = \frac{1}{\sigma}\operatorname{sh}(\sigma x) \\ \mathbf{Y}(x)_{21} &= \frac{-\sigma^2(e^{\sigma x} - \sigma\, e^{-\sigma x})}{-2\sigma} = \sigma\frac{e^{\sigma x} - e^{-\sigma x}}{2} = \sigma \operatorname{sh}(\sigma x) \\ \mathbf{Y}(x)_{22} &= \frac{-\sigma\, e^{-\sigma x} - \sigma\, e^{\sigma x}}{-2\sigma} = \frac{e^{\sigma x} + e^{-\sigma x}}{2} = \operatorname{ch}(\sigma x) \\ \mathbf{Y}(x) &= \begin{bmatrix} \operatorname{ch}(\sigma x) & \frac{1}{\sigma}\operatorname{sh}(\sigma x) \\ \sigma \operatorname{sh}(\sigma x) & \operatorname{ch}(\sigma x) \end{bmatrix}, \quad \det(\mathbf{Y}(x)) = 1. \end{aligned}$$

A $\mathbf{G}$ mátrix és a megoldandó lineáris egyenletrendszer

$$\begin{aligned} \mathbf{G} = \mathbf{B_0 I} + \mathbf{B_1 Y}(1) &= \begin{bmatrix} -\frac{\sigma}{2} & -1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \sigma & 1 \end{bmatrix} \cdot \begin{bmatrix} \operatorname{ch}(\sigma) & \frac{1}{\sigma}\operatorname{sh}(\sigma) \\ \sigma \operatorname{sh}(\sigma) & \operatorname{ch}(\sigma) \end{bmatrix} = \\ &= \begin{bmatrix} -\frac{\sigma}{2} & -1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \sigma \operatorname{ch}(\sigma) + \sigma \operatorname{sh}(\sigma) & \operatorname{ch}(\sigma) + \operatorname{sh}(\sigma) \end{bmatrix} = \begin{bmatrix} -\frac{\sigma}{2} & -1 \\ \sigma e^{\sigma} & e^{\sigma} \end{bmatrix} \\ \mathbf{Gq} = \mathbf{p} \quad \Leftrightarrow \quad \begin{bmatrix} -\frac{\sigma}{2} & -1 \\ \sigma e^{\sigma} & e^{\sigma} \end{bmatrix} \cdot \mathbf{q} &= \begin{bmatrix} \frac{\sigma}{2} \\ 0 \end{bmatrix} \quad \Rightarrow \quad \mathbf{q} = \begin{bmatrix} 1 \\ -\sigma \end{bmatrix}. \end{aligned}$$

Az c megoldásunk tehát előáll, mint

$$c(x) = (\mathbf{Y}(x)\mathbf{q})_1 = \operatorname{ch}(\sigma x) - \operatorname{sh}(\sigma x) = \frac{e^{\sigma x} + e^{-\sigma x}}{2} - \frac{e^{\sigma x} - e^{-\sigma x}}{2} = e^{-\sigma x}.$$

Ha a lineáris egynletrendszert pontosan oldjuk meg, és $c(1)$-et a fenti mátrix-vektor szorzásból számoljuk, akkor két közeli nagy számot vonunk ki egymásból, a kivonási jegyveszteség nagy, így pontatlan eredményt kapunk. Példaként nézzük, mit számol a Matlab 16 jegy pontossággal:

$$\begin{aligned} \mathrm{ch}\,(20) &= 2.425825977048951\,e+08,\ \ \mathrm{sh}\,(20) = 2.425825977048951\,e+08 \\ \text{így}\ \ \mathrm{ch}\,(20) - \mathrm{sh}\,(20) &= 0,\ \ \text{ugyanakkor}\ \ e^{-20} \approx 2.061153622438558\,e-09 \\ \mathrm{cond}_\infty(\mathbf{G}) &= 2.037693822821119\,e+10. \end{aligned}$$

Másrészt $\mathbf{q}$-t nem tudjuk pontosan kiszámolni, mert egy rosszul kondicionált lineáris egyenletrendszerből kapjuk, ahol $\mathrm{cond}_\infty(\mathbf{G}) = (\sigma+1)(2e^\sigma+1) \approx 2\cdot 10^{10}$ (lásd még a 8-5. Példát). Továbbá $\mathbf{z}(1)$ numerikus kiszámítása az egyenletrendszer jobboldalán szintén hibát eredményez. Ha a példában szereplő $\sigma = 20$ helyett ennél nagyobbat vennénk, akkor a probléma még erősebben jelentkezik.

Ezzel elérkeztünk ahhoz a kérdéshez, hogy a peremérték feladat mikor jól meghatározott? A peremérték feladat kondicionáltságát az alapmátrix és a $\mathbf{G}$ mátrix határozza meg. Tekintsük a $\mathbf{w}(x) := \mathbf{y}(x) - \mathbf{z}(x)$ függvényt, ahol $\mathbf{z}(x)$ egy partikuláris megoldás. Nyilván $\mathbf{w}$ a homogén egyenlet megoldása:

$$\mathbf{w} = \mathbf{Yq} = \mathbf{YG^{-1}r}, \quad \mathbf{r} := \mathbf{p} - \mathbf{B_1 z}(1).$$

Vizsgáljuk meg a változást, ha a

$$\mathbf{w} = \mathbf{YG^{-1}r} \quad \text{helyett a} \quad \widetilde{\mathbf{w}} = \mathbf{YG^{-1}}\widetilde{\mathbf{r}}$$

megváltozott lineáris egyenletrendszert oldjuk meg.

8-2. T *(Lineáris egyenletrendszer perturbációs tétele)*
Rögzített x esetén a megváltoztatott lineáris egyenletrendszer megoldásának abszolút és relatív hibájára a következő becslés adható

$$\|\mathbf{w}(x) - \widetilde{\mathbf{w}}(x)\| \le \|\mathbf{Y}(x)\mathbf{G^{-1}}\| \cdot \|\mathbf{r}(x) - \widetilde{\mathbf{r}}(x)\|,$$

$$\frac{\|\mathbf{w}(x) - \widetilde{\mathbf{w}}(x)\|}{\|\mathbf{w}(x)\|} \le \mathrm{cond}(\mathbf{Y}(x)\mathbf{G^{-1}}) \frac{\|\mathbf{r}(x) - \widetilde{\mathbf{r}}(x)\|}{\|\mathbf{r}(x)\|}.$$

Bizonyítás. A rövidebb írásmód kedvéért elhagyjuk az x-től való függés felírását, majd a kész becslésben visszatérünk rá. Tehát az egyenleteink

$$\mathbf{w} = \mathbf{YG^{-1}r} \quad \text{és} \quad \widetilde{\mathbf{w}} = \mathbf{YG^{-1}}\widetilde{\mathbf{r}}.$$

Ezeket egymásból kivonva

$$\mathbf{w} - \widetilde{\mathbf{w}} = \mathbf{YG^{-1}}(\mathbf{r} - \widetilde{\mathbf{r}}).$$

Innen a normabecslés

$$\|\mathbf{w} - \widetilde{\mathbf{w}}\| \le \|\mathbf{YG^{-1}}\| \cdot \|\mathbf{r} - \widetilde{\mathbf{r}}\|,$$

ez volt a tétel egyik állítása. Másrészt

$$\mathbf{w} = \mathbf{YG^{-1}r} \quad \Rightarrow \quad \mathbf{r} = (\mathbf{YG^{-1}})^{-1} \cdot \mathbf{w}$$

$$\|\mathbf{r}\| \le \|(\mathbf{YG^{-1}})^{-1}\| \cdot \|\mathbf{w}\| \quad \Rightarrow \quad \|\mathbf{w}\| \ge \frac{\|\mathbf{r}\|}{\|(\mathbf{YG^{-1}})^{-1}\|}.$$

A kétféle becslést felhasználva

$$\frac{\|\mathbf{w} - \widetilde{\mathbf{w}}\|}{\|\mathbf{w}\|} \le \|\mathbf{YG^{-1}}\| \cdot \frac{\|(\mathbf{YG^{-1}})^{-1}\|}{\|\mathbf{r}\|} \cdot \|\mathbf{r} - \widetilde{\mathbf{r}}\| = \mathrm{cond}(\mathbf{YG^{-1}}) \frac{\|\mathbf{r} - \widetilde{\mathbf{r}}\|}{\|\mathbf{r}\|}.$$

Tehát rögzített x-re

$$\frac{\|\mathbf{w}(x)-\widetilde{\mathbf{w}}(x)\|}{\|\mathbf{w}(x)\|} \le \operatorname{cond}(\mathbf{Y}(x)\mathbf{G}^{-1})\frac{\|\mathbf{r}(x)-\widetilde{\mathbf{r}}(x)\|}{\|\mathbf{r}(x)\|}.$$

■

8-3. T *(A perturbált peremérték-feladat abszolút- és relatív hibája)*

$$\max_{x\in[0;1]}\|\mathbf{w}(x)-\widetilde{\mathbf{w}}(x)\| \le \max_{x\in[0;1]}\|\mathbf{Y}(x)\mathbf{G}^{-1}\|\cdot\max_{x\in[0;1]}\|\mathbf{r}(x)-\widetilde{\mathbf{r}}(x)\|$$

$$\max_{x\in[0;1]}\frac{\|\mathbf{w}(x)-\widetilde{\mathbf{w}}(x)\|}{\|\mathbf{w}(x)\|} \le \max_{x\in[0;1]}\operatorname{cond}(\mathbf{Y}(x)\mathbf{G}^{-1})\max_{x\in[0;1]}\frac{\|\mathbf{r}(x)-\widetilde{\mathbf{r}}(x)\|}{\|\mathbf{r}(x)\|}.$$

Bizonyítás. Ha a lineáris egyenletrendszer perturbációs tételét alkalmazzuk minden $x\in[0;1]$ esetén, akkor megkapjuk a tétel becsléseit. ■

Ezek után definiálhatjuk a peremérték feladat abszolút- és relatív kondíciószámát.

8-1. Definíció. Legyen a peremérték feladathoz tartozó $\mathbf{G}=\mathbf{B_0}\mathbf{Y}(0)+\mathbf{B_1}\mathbf{Y}(1)$ mátrix reguláris, ekkor a

$$\kappa_{abs} := \max_x \|\mathbf{Y}(x)\mathbf{G}^{-1}\|$$

számot a peremérték feladat abszolút kondíciószámának nevezzük, a

$$\kappa_{rel} := \max_x \operatorname{cond}(\mathbf{Y}(x)\mathbf{G}^{-1})$$

számot pedig a peremérték feladat relatív kondíciószámának.

8-5. Példa. Számítsuk ki az előző példára a peremérték feladat abszolút és relatív kondíciószámait!

$$\mathbf{Y}(0)=\mathbf{I},\quad \mathbf{G}=\begin{bmatrix}-\frac{\sigma}{2} & -1\\ \sigma e^{\sigma} & e^{\sigma}\end{bmatrix},\quad \det(\mathbf{G})=-\frac{\sigma}{2}e^{\sigma}+\sigma e^{\sigma}=\frac{\sigma}{2}e^{\sigma},$$

$$\mathbf{G}^{-1}=\frac{2}{\sigma}e^{-\sigma}\begin{bmatrix}e^{\sigma} & 1\\ -\sigma e^{\sigma} & -\frac{\sigma}{2}\end{bmatrix}=\begin{bmatrix}\frac{2}{\sigma} & \frac{2}{\sigma}e^{-\sigma}\\ -2 & -e^{-\sigma}\end{bmatrix}$$

Számítsuk ki a mátrixok végtelen normáját $\sigma\ge 1.14$ esetén:

$$\|\mathbf{G}\|_\infty=\max\left\{1+\frac{\sigma}{2};e^{\sigma}(\sigma+1)\right\}=e^{\sigma}(\sigma+1)$$

$$\|\mathbf{G}^{-1}\|_\infty=\max\left\{\frac{2}{\sigma}(1+e^{-\sigma});2+e^{-\sigma}\right\}=2+e^{-\sigma}$$

A fenti σ-ra tett feltétel az inverz normájából jön ki. Maple segítségével felrajzolható a $2+e^{-\sigma}$ és a $\frac{2}{\sigma}(1+e^{-\sigma})$ függvény, így ellenőrizhető a kapott σ érték.

$$\kappa_{abs}\ge\|\mathbf{Y}(0)\mathbf{G}^{-1}\|_\infty=\|\mathbf{G}^{-1}\|_\infty=2+e^{-\sigma}\quad\text{és}$$
$$\kappa_{rel}\ge\operatorname{cond}_\infty(\mathbf{G}^{-1})=\operatorname{cond}_\infty(\mathbf{G})=(\sigma+1)(2e^{\sigma}+1),$$

tehát a peremérték feladat ugyanolyan rosszul kondicionált, mint a lineáris egyenletrendszer.
Az előző feladatban megadott $\sigma=20$ esetén

$$\kappa_{rel}\ge\operatorname{cond}_\infty(\mathbf{G})=21(2e^{20}+1)\approx 2\cdot 10^{10}.$$

Megjegyzések.

1. A peremérték feladat kondíciószámai függetlenek attól, hogy milyen $\mathbf{M}$ reguláris mátrixot választunk $\mathbf{Y}(0)$-nak. Ugyanis, ha $\mathbf{Y_I}(x)$ illetve $\mathbf{Y_M}(x)$ az $\mathbf{Y_I}(0) = \mathbf{I}$-nek és $\mathbf{Y_M}(0) = \mathbf{M}$-nek eleget tevő alapmátrixot, akkor

$$\mathbf{Y_M}(x) = \mathbf{Y_I}(x)\mathbf{M} \quad \text{és} \quad \mathbf{G_M} = (\mathbf{B_0}\mathbf{Y}(0) + \mathbf{B_1}\mathbf{Y}(1))\mathbf{M} = \mathbf{G_I} \cdot \mathbf{M}.$$

Az $\mathbf{M}$ szorzó a norma számításakor kiesik.

2. Vegyük észre (lásd a példát), hogy a definíció szerint jól kondicionált peremérték feladathoz jól kondicionált lineáris egyenletrendszer is tartozik, hiszen

$$\kappa_{rel} \geq \operatorname{cond}(\mathbf{Y}(0)\mathbf{G}^{-\mathbf{1}}) = \operatorname{cond}(\mathbf{G}).$$

8.5. Egy modellfeladat

Vizsgáljuk a következő speciális ($D = 1, \alpha_1 = \alpha_2 = 1, \beta_1 = \beta_2 = 0$) elsőfajú peremérték feladatot:

$$\begin{aligned} u''(x) + f(x) = 0, \;\; 0 \leq x \leq 1 \\ u(0) = a, \;\; u(1) = b. \end{aligned}$$

Nézzük meg a feladat kondicionáltságát, mielőtt a megoldásnak egy más fajta előállítását tárgyaljuk.

$$\mathbf{Y}(x) = \begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix} \;\text{ és }\; \mathbf{G} = \begin{bmatrix} \alpha_1 & -\beta 1 \\ \alpha_2 & \alpha_2 + \beta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix},$$

innen végtelen normában a peremérték feladat abszolút és relatív kondíciószáma:

$$\kappa_{abs} = \max_{x\in[0;1]} \|\mathbf{Y}(x)\mathbf{G}^{-\mathbf{1}}\|_\infty = \max_{x\in[0;1]} \left\| \begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \right\|_\infty = \max_{x\in[0;1]} \left\| \begin{bmatrix} 1-x & x \\ -1 & 1 \end{bmatrix} \right\|_\infty = 2$$

$$\kappa_{rel} = \max_{x\in[0;1]} \operatorname{cond}_\infty(\mathbf{Y}(x)\mathbf{G}^{-\mathbf{1}}) = \max_{x\in[0;1]} \operatorname{cond}_\infty \left(\begin{bmatrix} 1-x & x \\ -1 & 1 \end{bmatrix} \right) =$$

$$= \max_{x\in[0;1]} \left\| \begin{bmatrix} 1-x & x \\ -1 & 1 \end{bmatrix} \right\|_\infty \cdot \left\| \begin{bmatrix} 1 & -x \\ 1 & 1-x \end{bmatrix} \right\|_\infty = 2 \cdot 2 = 4.$$

Modellfeladatunk tehát jól kondicionált. Megoldását nem csak az alapmátrix segítségével állíthatjuk elő, hanem közvetlenül kétszeres integrálással.

$$\begin{aligned} u'(x) &= -\int_0^x f(r)\,dr + u'(0), \\ u(x) &= \int_0^x u'(s)\,ds + u(0) = -\int_0^x \int_0^s f(r)\,dr ds + u'(0)x + u(0). \end{aligned}$$

Figyelembe véve a peremfeltételeket

$$b = u(1) = -\int_0^1 \int_0^s f(r)\,dr ds + u'(0) + a,$$

innen $u'(0)$-t kifejezve

$$u'(0) = \int_0^1 \int_0^s f(r)\,dr ds + (b - a).$$

Írjuk vissza $u(x)$ képletébe:

$$u(x) = a + (b-a)x + x\int_0^1 \int_0^s f(r)\,drds - \int_0^x \int_0^s f(r)\,drds.$$

Fubini tételét felhasználva alakítsuk át az integrálokat. Cseréljük fel az integrálások sorrendjét úgy, hogy az integrálási tartomány ne változzon. Az első integrálnál a tartomány a $(0;0),(1;0),(1;1)$ pontok által meghatározott derékszögű háromszög, a másodiknál a $(0;0),(x;0),(x;x)$ pontok által meghatározott derékszögű háromszög.

$$\int_0^1 \int_0^s f(r)\,drds = \int_0^1 \int_r^1 f(r)\,dsdr = \int_0^1 (1-r)f(r)\,dr,$$
$$\int_0^x \int_0^s f(r)\,drds = \int_0^x \int_r^x f(r)\,dsdr = \int_0^x (x-r)f(r)\,dr$$

Összesítve kapjuk, hogy

$$\begin{aligned}
u(x) &= a + (b-a)x + x\int_0^1 (1-r)f(r)\,dr - \int_0^x (x-r)f(r)\,dr = \\
&= a + (b-a)x + \left(\int_0^x x(1-r)f(r)\,dr + \int_x^1 x(1-r)f(r)\,dr\right) - \int_0^x (x-r)f(r)\,dr = \\
&= a + (b-a)x + \left(\int_0^x x(1-r)f(r)\,dr - \int_0^x (x-r)f(r)\,dr\right) + \int_x^1 x(1-r)f(r)\,dr = \\
&= a + (b-a)x + \int_0^x \left(x(1-r) - (x-r)\right) f(r)\,dr + \int_x^1 x(1-r)f(r)\,dr = \\
&= a + (b-a)x + \int_0^x r(1-x)f(r)\,dr + \int_x^1 x(1-r)f(r)\,dr.
\end{aligned}$$

Vezessük be a peremérték feladat **Green-függvény**ét:

$$G(x,r) := \begin{cases} x(1-r) & 0 \le x \le r \le 1, \\ r(1-x) & 0 \le r \le x \le 1, \end{cases}$$

mellyel a megoldást a következő egyszerű alakban tudjuk felírni:

$$u(x) = a + (b-a)x + \int_0^1 G(x,r)f(r)\,dr.$$

A Green-függvény tulajdonságai:

1) Szimmetrikus: $G(x,r) = G(r,x)$.

2) Pozitív: $G(x,r) > 0$ bármely $r, x \in (0;1)$-re.

3) Eleget tesz

   (a) a homogén peremfeltételeknek: $G(0,r) = G(x,0) = 0$,

   (b) a homogén differenciálegyenletnek is, az $x = r$ egyenest nem számítva:

$$G_{rr}(x,r) = 0, \;\; G_{xx}(x,r) = 0, \;\; \text{bármely } \; x \neq r, \;\; 0 \le x, r \le 1.$$

   Itt G_{rr}, G_{xx} a parciális deriváltakat jelöli.

4) Az $x = r$ egyenesen érvényes

$$\lim_{\varepsilon\to 0+} G_x(x+\varepsilon, x) = -x, \quad \lim_{\varepsilon\to 0+} G_x(x-\varepsilon, x) = 1-x,$$

tehát

$$G_x(x+0, x) - G_x(x-0, x) = -1.$$

$$G_{xx}(x, r) = 0, \text{ ha } x \neq r,$$
$$G_{xx}(x, x) = -\infty,$$

vagyis

$$G_{xx}(x, r) = -\delta(x, r)$$

a Dirac-féle delta függvénnyel (az elektronikában az egységnyi impulzus szimbóluma, disztribúció). Ezt úgy kell érteni, hogy igaz

$$\frac{d}{dx}\int_0^1 G_x(x,r)f(r)\,dr = \frac{d}{dx}\left(\int_0^x -rf(r)\,dr - \int_x^1 (1-r)f(r)\,dr\right) =$$
$$= \frac{d}{dx}\left(\int_x^1 f(r)\,dr - \int_0^1 rf(r)\,dr\right) = -f(x).$$

5) Az

$$u(x) = a + (b-a)x + \int_0^1 G(x,r)f(r)\,dr$$

képletből leolvashatjuk, hogy $a = b = 0$ esetén $u(x) = \int_0^1 G(x,r)f(r)\,dr$ a peremérték feladat megoldása, továbbá folytonos f esetén az u megoldásunk klasszikus lesz, az u, u', u'' megoldás és deriváltjai folytonosak. A képlet akkor is érvényes, ha f csak integrálható, hiszen kétszer integráltunk a képlet előállításához. A Green-függvény nemnegativitása miatt igaz az is, hogy ha $a, b \geq 0$ és $f(x) \geq 0$, $x \in [0;1]$, akkor u nemnegatív.

9. fejezet

Véges differencia eljárások

9.1. Bevezetés, alapvető fogalmak

Ebben a fejezetben a peremértékfeladatok numerikus megoldásával foglakozunk. Tekintsük a következő feladatot:

$$y''(x) + f(x) = 0, \;\; 0 \le x \le 1$$
$$u(0) = a, \;\; u(1) = b.$$

A $[0;1]$ intervallumot osszuk N részre. Vezessük be a következő jelöléseket. A belső pontokat tartalmazó rács

$$\omega_h := \left\{ x_i := ih \,|\, i = 1, \ldots, N-1; h = \frac{1}{N} \right\},$$

a perempontokat is tartalmazó rács

$$\varpi_h := \left\{ x_i := ih \,|\, i = 0, \ldots, N; h = \frac{1}{N} \right\}$$

és a perempontok halmaza $\gamma_h := \{0, 1\}$. Legyen $v : [0;1] \to \mathbb{R}$ folytonos függvény és vezessük be a következő jelöléseket:

$$v_i := v(x_i), \quad v_{\overline{x}x,i} := \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2}.$$

9-1. L *Ha $v \in C^4[0;1]$, akkor a fenti másodrendű differenciahányados a második derivált másodrendű közelítését adja, azaz bármely $x_i \in \varpi_h$ esetén létezik olyan $|\eta_i| < 1$, melyre*

$$\frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} = v''(x_i) + \frac{h^2}{12} v^{(4)}(x_i + \eta_i h).$$

Bizonyítás. A Taylor-formula alapján létezik olyan $\xi_i^+ \in (x_i, x_{i+1})$ és $\xi_i^- \in (x_{i-1}, x_i)$, hogy

$$\begin{aligned} v_{i+1} + v_{i-1} &= \left(v_i + hv_i' + \frac{h^2}{2} v_i'' + \frac{h^3}{6} v_i''' + \frac{h^4}{24} v^{(4)}(\xi_i^+) \right) + \\ &+ \left(v_i - hv_i' + \frac{h^2}{2} v_i'' - \frac{h^3}{6} v_i''' + \frac{h^4}{24} v^{(4)}(\xi_i^-) \right) = \\ &= 2v_i + h^2 v_i'' + \frac{h^4}{12} \cdot \frac{1}{2} \left(v^{(4)}(\xi_i^+) + v^{(4)}(\xi_i^-) \right) = \\ &= 2v_i + h^2 v_i'' + \frac{h^4}{12} v^{(4)}(x_i + \eta_i h), \;\; |\eta_i| < 1. \end{aligned}$$

Innen

$$\frac{v_{i+1}-2v_i+v_{i-1}}{h^2} = v''(x_i) + \frac{h^2}{12}\,v^{(4)}(x_i+\eta_i h),\ \ |\eta_i|<1,$$

a másodrendű differencia hányadost kaptuk, ami v második deriváltját közelíti. ■

Ezzel az eredménnyel a peremérték feladat a következő lineáris egyenletrendszerrel helyettesíthető:

$$\begin{aligned} y_0 &= a,\\ -y_{\overline{x}x,i} &= f_i,\ \ i=1,\dots,N-1,\\ y_N &= b. \end{aligned}$$

9-1. Definíció. A peremértékfeladatnak a kapott lineáris egyenletrendszerrel való helyettesítését véges differencia diszkretizációnak nevezzük. A $h=1/N$ paraméterű rendszersereget differenciasémának nevezzük. A h paraméter a differenciaséma lépéstávolsága.

A differenciaséma indexmentes alakban

$$\begin{aligned} y_{\overline{x}x} + f(x) &= 0,\ \ x\in\omega_h\\ y(x) &= c(x),\ \ x\in\gamma_h, \end{aligned}$$

ahol $c(0)=a$ és $c(1)=b$. Írjuk fel az $(N+1)\times(N+1)$-es lineáris egyenletrendszer mátrixát, valamint a jobboldalát és megoldását leíró vektort:

$$\widetilde{\mathbf{A}}_{\mathbf{h}} := \frac{1}{h^2}\begin{bmatrix} h^2 & 0 & 0 & . & . & . & 0\\ -1 & 2 & -1 & 0 & . & . & 0\\ 0 & -1 & 2 & -1 & 0 & . & 0\\ . & . & . & . & . & . & .\\ 0 & . & . & -1 & 2 & -1 & 0\\ 0 & . & . & . & -1 & 2 & -1\\ 0 & . & . & . & . & 0 & h^2 \end{bmatrix},\quad \widetilde{\mathbf{f}} := \begin{bmatrix} a\\ f_1\\ f_2\\ .\\ .\\ f_{N-1}\\ b \end{bmatrix},\quad \widetilde{\mathbf{y}} := \begin{bmatrix} y_0\\ y_1\\ y_2\\ .\\ .\\ y_{N-1}\\ y_N \end{bmatrix}.$$

9-2. L *Az $\widetilde{\mathbf{A}}_{\mathbf{h}}$ mátrix M-mátrix.*

Bizonyítás. Az előjelfeltételek teljesülnek $\widetilde{\mathbf{A}}_{\mathbf{h}}$-ra.

Készítsünk olyan $\mathbf{g}>\mathbf{0}$ vektort, melyre $\mathbf{Ag}>\mathbf{0}$. Vezessük be a $w(x) := 1+x(1-x)$ konkáv majoráns függvényt és a $\widetilde{\mathbf{w}}$ vektort, mely a w függvény rácspontokon felvett értékeit tartalmazza:

$$\widetilde{\mathbf{w}} = (w(x_0), w(x_1), \dots, w(x_{N-1}), w(x_N)).$$

Ekkor $\widetilde{\mathbf{w}}>\mathbf{0}$ és

$$(\widetilde{\mathbf{A}}\widetilde{\mathbf{w}})_i = \begin{cases} w(x_i)=1 & i=0,N\\ \frac{1}{h^2}(-w(x_{i-1})+2w(x_i)-w(x_{i-1})) = -w''(x_i) = 2 & i=1,\dots,N-1. \end{cases}$$

Tehát a $\mathbf{g} := \widetilde{\mathbf{w}}$ választás jó, ezzel beláttuk, hogy $\widetilde{\mathbf{A}}_{\mathbf{h}}$ M-mátrix. ■

9-3. L *A fenti jelölésekkel a következő a-priori hibabecslés adható (a megoldás ismerete nem szükséges hozzá)*

$$\max_{i=0}^{N}|\widetilde{y}_i| =: \|\widetilde{\mathbf{y}}\|_{C(\widetilde{\omega}_h)} \le \frac{5}{4}\|\widetilde{\mathbf{f}}\|_{C(\widetilde{\omega}_h)} := \frac{5}{4}\max_{i=0}^{N}|\widetilde{f}_i|.$$

Bizonyítás. A rövidebb írásmód kedvéért $\|\widetilde{\mathbf{y}}\|_{C(\widetilde{\omega}_h)}$ helyett $\|\widetilde{\mathbf{y}}\|_\infty$-t írunk a bizonyításban.

$$\max_{i=0}^{N}|\widetilde{y}_i| =: \|\widetilde{\mathbf{y}}\|_\infty = \max_{i=0}^{N}|((\widetilde{\mathbf{A}}_\mathbf{h})^{-\mathbf{1}}\widetilde{\mathbf{f}})_i| = \|(\widetilde{\mathbf{A}}_\mathbf{h})^{-\mathbf{1}}\widetilde{\mathbf{f}}\|_\infty \leq$$
$$\leq \|(\widetilde{\mathbf{A}}_\mathbf{h})^{-\mathbf{1}}\|_\infty \|\widetilde{\mathbf{f}}\|_\infty$$

Az M-mátrixokról tanult 14. tételt alkalmazzuk az $\|(\widetilde{\mathbf{A}}_\mathbf{h})^{-\mathbf{1}}\|_\infty$ becslésére a $\mathbf{g} := \widetilde{\mathbf{w}}$ és $\mathbf{A} := \widetilde{\mathbf{A}}_\mathbf{h}$ jelöléssel ($\widetilde{\mathbf{w}}$ az előző lemmából)

$$\|\mathbf{A}^{-\mathbf{1}}\|_\infty \leq \frac{\|\mathbf{g}\|_\infty}{\min_{i=1}^n(\mathbf{A}\mathbf{g})_i} \leq \frac{\|\widetilde{\mathbf{w}}\|_\infty}{\min_{i=1}^n(\widetilde{\mathbf{A}}_\mathbf{h}\widetilde{\mathbf{w}})_i} = \frac{\frac{5}{4}}{1} = \frac{5}{4},$$

ahol

$$\|\widetilde{\mathbf{w}}\|_\infty \leq \max_{x\in[0;1]}|w(x)| = \left|w\left(\frac{1}{2}\right)\right| = 1 + \frac{1}{2}\cdot\frac{1}{2} = \frac{5}{4}.$$

A kapott becslést beírva kapjuk a tétel állítását.

$$\|\widetilde{\mathbf{y}}\|_\infty \leq \|(\widetilde{\mathbf{A}}_\mathbf{h})^{-\mathbf{1}}\|_\infty\|\widetilde{\mathbf{f}}\|_\infty \leq \frac{5}{4}\|\widetilde{\mathbf{f}}\|_\infty.$$

■

A fenti lineáris egyenletrendszert rövidített Gauss-eliminációval oldjuk meg, melynek műveletigénye $8N + O(1)$. Az módszer alkalmazásához szükséges feltételek $\widetilde{\mathbf{A}}_\mathbf{h} = \text{tridiag}(-a_i, b_i, -c_i)$:

$$a_i, b_i, c_i > 0, \quad b_i \geq a_i + c_i \text{ és } \exists\, j : b_j > a_j + c_j.$$

Behelyettesítve a konkrét elemeket

$$b_0 = 1 > c_0 = 0 \quad b_i = a_i + c_i \left(= \frac{2}{h^2}\right), i = 1, \ldots, N-1, \quad b_N = 1 > a_N = 0,$$

a feltételek teljesülnek, vagyis az algoritmus minden esetben végrehajtható.

A rövidített Gauss-elimináció algoritmusa

$\mathbf{A} = \text{tridiag}(a_i, b_i, c_i)_{i=0}^N$ és $(f_i)_{i=0}^N$ jobboldal esetén:

$$\alpha_0 := 0, \;\; \beta_0 := 0,$$
$$i = 0, \ldots, N-1:$$
$$\delta := b_i - a_i\alpha_i$$
$$\alpha_{i+1} := \frac{c_i}{\delta}$$
$$\beta_{i+1} := \frac{f_i + a_i\beta_i}{\delta}$$
$$\beta_{N+1} := \frac{f_N + a_N\beta_N}{b_N - a_N\alpha_N}$$
$$y_N := \beta_{N+1}$$
$$i = N-1, \ldots, 0:$$
$$y_i = \alpha_{i+1}y_{i+1} + \beta_{i+1}$$

Ezután eliminálják a peremfeltételeket.

a) $f_i = 0$, $i = 1, \ldots, N-1$ esetén

$$y_i = a + (b-a)x_i, \;\; i = 0, \ldots, N$$

megoldása a lineáris egyenletrendszernek. $y_0 = a$ és $y_N = b$ teljesül, ellenőrizzük, hogy az többi egyenlet is rendben van-e. Az i. egyenlet $1/h^2$ nélküli alakja $i = 1, \ldots, N-1$-re

$$\begin{aligned} -y_{i-1} + 2y_i - y_{i+1} &= -[a + (b-a)x_{i-1}] + 2[a + (b-a)x_i] - [a + (b-a)x_{i+1}] = \\ &= -a + 2a - a + (b-a)[-x_{i-1} + 2x_i - x_{i+1}] = \\ &= (b-a)[-x_i + h + 2x_i - x_i - h] = 0. \end{aligned}$$

b) Az általános megoldást a következő alakban írjuk fel

$$y(x) = a + (b-a)x + v(x), \quad x \in \varpi_h.$$

Ekkor a v-re felírt egyenlet:

$$\begin{aligned} v_{\overline{x}x,i} + f(x) = 0, \quad x \in \omega_h \\ v(x) = 0, \quad x \in \gamma_h. \end{aligned}$$

Mátrixos alakja

$$\mathbf{A_h v} = \overrightarrow{\mathbf{f}}, \text{ ahol}$$

$$\mathbf{A_h} := \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & . & . \\ -1 & 2 & -1 & 0 & . \\ . & . & . & . & . \\ . & . & -1 & 2 & -1 \\ . & . & . & -1 & 2 \end{bmatrix}, \quad \overrightarrow{\mathbf{f}} := \begin{bmatrix} f_1 \\ f_2 \\ . \\ . \\ f_{N-1} \end{bmatrix}, \quad \mathbf{v} := \begin{bmatrix} v_1 \\ v_2 \\ . \\ . \\ v_{N-1} \end{bmatrix},$$

ahol $\overrightarrow{\mathbf{f}}$ a perempontokat nem tartalmazza.

9-4. L *A perempontok eliminációja utáni* $\mathbf{A_h}$ *mátrix szimmetrikus és az M-mátrix tulajdonsága is megmarad.*

Bizonyítás. Ezt a korábbiakhoz hasonlóan a $w(x) := x(1-x)$ függvénnyel és a $\mathbf{g} := \overrightarrow{\mathbf{w}} > \mathbf{0}$ vektorral érhetjük el és

$$(\mathbf{A_h g})_i = -w''(x_i) = 2 \quad \forall\, i = 1, \ldots, N-1.$$

■

Megjegyzések.

1. Az $\mathbf{A_h}$ mátrix felírható a következő egyszerű módon $\mathbf{A_h} = \frac{1}{h^2}(\mathbf{K} + \mathbf{K^T})$, ahol

$$\mathbf{K} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & -1 & 1 \end{bmatrix} \quad \text{és} \quad \mathbf{S} := \mathbf{K^{-1}} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 1 & \ldots & \ldots & 1 \end{bmatrix}.$$

A [3] jegyzet 26. oldalán található ötletet felhasználva, vegyük észre, hogy az $\frac{1}{h}\mathbf{K}$ mátrix az előre mutató differencia operátor mátrixa ($\mathbf{K}$: különbségképzés mátrix), míg az $h\,\mathbf{S}$ mátrix a differencia operátor inverzének mátrixa ($\mathbf{S}$: összegzésmátrix). Ebből a gondolatmenetből következik, hogy $h\,\mathbf{S} = h\,\mathbf{K^{-1}}$ az integráloperátor diszkretizációja. Az előző fejezetbeli Green-függvény diszkrét megfelelője pedig az $\mathbf{A_h^{-1}}$ mátrix.

2. Írjuk fel az $\mathbf{A_h}$ mátrix inverzét:

$$\mathbf{A_h^{-1}} = h^2\,(\mathbf{K}+\mathbf{K^T})^{-1} = h^2\,\left[\mathbf{K}\overbrace{\underbrace{(\mathbf{K^{-1}})^{\mathbf{T}}}_{\mathbf{S^T}}\mathbf{K^T}}^{I} + \overbrace{\mathbf{K}\underbrace{\mathbf{K^{-1}}}_{\mathbf{S}}}^{I}\mathbf{K^T}\right]^{-1} = h^2\left[\mathbf{K}\left(\mathbf{S^T}+\mathbf{S}\right)\mathbf{K^T}\right]^{-1} =$$
$$= h^2\,(\mathbf{K^{-1}})^{\mathbf{T}}(\mathbf{I}+\mathbf{ee^T})^{-1}\mathbf{K^{-1}} = h^2\,\mathbf{S^T}\left(\mathbf{I}-\frac{1}{N}\mathbf{ee^T}\right)\mathbf{S},$$

Az $\mathbf{S^T}+\mathbf{S} = \mathbf{I}+\mathbf{ee^T}$ inverzének ellenőrzése:

$$(\mathbf{I}+\mathbf{ee^T})\cdot\left(\mathbf{I}-\frac{1}{N}\mathbf{ee^T}\right) = \mathbf{I}+\mathbf{ee^T}-\frac{1}{N}\mathbf{ee^T}-\frac{1}{N}\mathbf{e}\underbrace{(\mathbf{e^Te})}_{=N-1}\mathbf{e^T} = \mathbf{I}.$$

A fenti képletből az $\mathbf{v} = \mathbf{A_h^{-1}}\overrightarrow{\mathbf{f}}$ megoldásvektor a rövidített Gauss-elimináció nélkül is könnyen előállítható. A fenti alak segítségével az inverz elemei zárt alakban is előállíthatók.

9-5. L *A korábbi a-priori hibabecslésnél pontosabb becslés adható*

$$\|\widetilde{\mathbf{y}}\|_{C(\widetilde{\omega}_h)} \le \max\{|a|,|b|\} + \frac{1}{8}\|\overrightarrow{\mathbf{f}}\|_{C(\widetilde{\omega}_h)}.$$

Bizonyítás. A rövidebb írásmód kedvéért $\|\mathbf{v}\|_{C(\widetilde{\omega}_h)}$ helyett $\|\mathbf{v}\|_\infty$-t írunk a bizonyításban.

$$\|\mathbf{v}\|_\infty = \|\mathbf{A_h^{-1}}\overrightarrow{\mathbf{f}}\|_\infty \le$$
$$\le \|\mathbf{A_h^{-1}}\|_\infty\|\overrightarrow{\mathbf{f}}\|_\infty$$

Az M-mátrixokról tanult 14. tételt alkalmazzuk az $\|\mathbf{A_h^{-1}}\|_\infty$ becslésére a $\mathbf{g} := \overrightarrow{\mathbf{w}}$ és $\mathbf{A} := \mathbf{A_h}$ jelöléssel

$$\|\mathbf{A^{-1}}\|_\infty \le \frac{\|\mathbf{g}\|_\infty}{\min_{i=1}^n(\mathbf{Ag})_i} \le \frac{\|\overrightarrow{\mathbf{w}}\|_\infty}{\min_{i=1}^n(\mathbf{A_h}\overrightarrow{\mathbf{w}})_i} = \frac{\frac{1}{4}}{2} = \frac{1}{8},$$

ahol $w(x) := x(1-x)$ az előző lemmában szereplő függvény

$$\|\overrightarrow{\mathbf{w}}\|_\infty \le \max_{x\in[0;1]}|w(x)| = \left|w\left(\frac{1}{2}\right)\right| = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}.$$

A kapott becslést beírva és a peremfeltételeket felhasználva kapjuk a tétel állítását.

$$\|\widetilde{\mathbf{y}}\|_\infty \le \max\{|a|,|b|\} + \|\mathbf{A_h^{-1}}\|_\infty\|\overrightarrow{\mathbf{f}}\|_\infty \le \max\{|a|,|b|\} + \frac{1}{8}\|\overrightarrow{\mathbf{f}}\|_\infty.$$

■

9-2. Definíció. A differenciaséma stabil (összhangban a stabilitás általános definíciójával), ha adott $\|.\|$ norma esetén, melyre $\|\mathbf{e_h}\| = 1$, létezik olyan M konstans, mely nem függ h-tól és

$$\|\widetilde{\mathbf{y}}\| \le M\|\widetilde{\mathbf{f}}\|.$$

Megjegyzések.

1. Az előző lemmákban éppen két ilyen alakú becslést kaptunk.

2. Ha a differenciaséma megoldására sikerül ilyen becslést levezetni, akkor a hozzátartozó lineáris egyenletrendszer mátrixa reguláris, ugyanis a homogén peremfeltételeket is jelentő $\widetilde{\mathbf{f}} = \mathbf{0}$ jobboldalhoz csak $\widetilde{\mathbf{y}} = \mathbf{0}$ megoldás tartozhat. Ekkor tehát pontosan egy megoldás létezik.

3. Rögzített peremértékek és változó jobboldal mellett a differenciaséma $\widetilde{\mathbf{y}}$ megoldása f-ben Lipschitz-folytonos és $M = \|\widetilde{\mathbf{A}}_\mathbf{h}^{-\mathbf{1}}\|$ a Lipschitz-konstans:

$$\|\widetilde{\mathbf{y}}(\widetilde{\mathbf{f}}) - \widetilde{\mathbf{y}}(\widetilde{\mathbf{g}})\| = \|\widetilde{\mathbf{A}}_\mathbf{h}^{-\mathbf{1}}\widetilde{\mathbf{f}} - \widetilde{\mathbf{A}}_\mathbf{h}^{-\mathbf{1}}\widetilde{\mathbf{g}}\| = \|\widetilde{\mathbf{A}}_\mathbf{h}^{-\mathbf{1}}(\widetilde{\mathbf{f}} - \widetilde{\mathbf{g}})\| \le \|\widetilde{\mathbf{A}}_\mathbf{h}^{-\mathbf{1}}\| \cdot \|\widetilde{\mathbf{f}} - \widetilde{\mathbf{g}}\| = M\|\widetilde{\mathbf{f}} - \widetilde{\mathbf{g}}\|.$$

4. Mivel $\widetilde{\mathbf{y}} = \widetilde{\mathbf{A}}_\mathbf{h}^{-\mathbf{1}}\widetilde{\mathbf{f}}$, ezért a stabilitás ekvivalens azzal, hogy $\|\widetilde{\mathbf{A}}_\mathbf{h}^{-\mathbf{1}}\|$ egyenletesen korlátos h-ban és azzal is, hogy $\|\mathbf{A}_\mathbf{h}^{-\mathbf{1}}\|$ egyenletesen korlátos h-ban.

9-3. Definíció. A $\widetilde{\Psi}$ maradékvektort (reziduum vektort) a pontos megoldásnak a differenciasémába való helyettesítésével a két oldal különbségéből kapjuk

$$\widetilde{\Psi}(x) = f(x) + u_{\overline{x}x}, \;\; x \in \omega_h; \quad \widetilde{\Psi}(x) = y(x) - u(x), \;\; x \in \gamma_h$$

A $\widetilde{\mathbf{z}} = \widetilde{\mathbf{y}} - \widetilde{\mathbf{u}}$ a hibavektor a közelítő ($\widetilde{\mathbf{y}}$) és a pontos megoldás rácspontokon vett eltérése, ahol $\widetilde{\mathbf{y}}$ a differenciasémából kapott megoldás és $\widetilde{\mathbf{u}}$ a pontos megoldás projekciója $C[0;1]$-ből $\mathbb{R}^{N+1}$-be.

$$\widetilde{\mathbf{u}} := (u(x_0), u(x_1), \ldots, u(x_{N-1}), u(x_N))^T$$

A két vektor kapcsolata

$$\widetilde{\mathbf{A}}_\mathbf{h}\widetilde{\mathbf{z}} = \widetilde{\Psi}.$$

9-4. Definíció. A séma approximálja a peremértékfeladatot, vagyis a séma konzisztens, ha a stabilitási becslésben használt $\|.\|$-val

$$\|\widetilde{\Psi}\| \to 0, \;\; h \to 0.$$

p az approximáció rendje illetve a konzisztencia rend, ha

$$\|\widetilde{\Psi}\| = O(h^p).$$

A séma konvergens, ha

$$y_i \to u(x_i), \;\; h \to 0 \quad (i = 0, \ldots, N).$$

9-6. T *Ha a peremértékfeladat megoldására $u \in C^4[0;1]$, akkor a $\widetilde{\mathbf{z}} = \widetilde{\mathbf{y}} - \widetilde{\mathbf{u}}$ hibavektorra*

$$\|\widetilde{\mathbf{z}}\|_{C(\varpi_h)} = \max_{i=0}^{N} |y_i - u(x_i)| \le \frac{h^2}{96}\|u^{(4)}\|_{C[0;1]} = \frac{h^2}{96}\|f''\|_{C[0;1]}.$$

Ez azt jelenti, hogy $y_i = u(x_i) + O(h^2)$, vagyis a közelítés másodrendű.

Bizonyítás. Tekintsük a $\widetilde{\mathbf{z}} = \widetilde{\mathbf{y}} - \widetilde{\mathbf{u}}$ hibavektort. Ez eleget tesz az

$$\widetilde{\mathbf{A}}_{\mathbf{h}}\widetilde{\mathbf{z}} = \widetilde{\Psi}$$

egyenletrendszernek, ahol

$$\widetilde{\Psi}(x) = \begin{cases} \widetilde{\mathbf{A}}_{\mathbf{h}}\widetilde{\mathbf{z}}(x) = -y_{\overline{x}x} + u_{\overline{x}x} = f(x) + u_{\overline{x}x} & x \in \omega_h \\ \widetilde{\mathbf{A}}_{\mathbf{h}}\widetilde{\mathbf{z}}(x) = \widetilde{z}(x) = 0 & x \in \gamma_h. \end{cases}$$

A $\widetilde{\Psi}$ vektort felírva

$$\widetilde{\Psi} = (0, f_1 + u_{\overline{x}x,1}, \ldots, f_{N-1} + u_{\overline{x}x,N-1}, 0)^T.$$

A korábbi hibabecslésből

$$\|\widetilde{\mathbf{z}}\|_{C(\varpi_h)} = \|\widetilde{\mathbf{y}} - \widetilde{\mathbf{u}}\|_{C(\varpi_h)} \leq \frac{1}{8}\|\widetilde{\mathbf{\Psi}}\|_{C(\varpi_h)}.$$

Ha $u \in C^4[0;1]$, akkor a sorfejtését

$$u_{\overline{x}x,i} = u''(x_i) + \frac{h^2}{12}u^{(4)}(x_i + \upsilon_i h), \;\; |\upsilon_i| < 1$$

felhasználva

$$\psi_i = f_i + u_{\overline{x}x,i} = -u''(x_i) + u''(x_i) + \frac{h^2}{12}u^{(4)}(x_i + \upsilon_i h) = \frac{h^2}{12}u^{(4)}(x_i + \upsilon_i h).$$

A $\widetilde{\Psi}$ vektor normájának becslése

$$\|\widetilde{\mathbf{\Psi}}\|_{C(\varpi_h)} \leq \frac{h^2}{12}\|u^{(4)}\|_{C[0;1]},$$

amit beírva a korábbi becslésbe

$$\|\widetilde{\mathbf{z}}\|_{C(\varpi_h)} =\leq \frac{1}{8}\|\widetilde{\mathbf{\Psi}}\|_{C(\varpi_h)} \leq \frac{1}{8}\cdot\frac{h^2}{12}\|u^{(4)}\|_{C[0;1]} = \frac{h^2}{96}\|f''\|_{C[0;1]}.$$

■

9-7. T *(A modellfeladat differenciasémájának tulajdonságai)*
Ha $f \in C^2[0;1]$, akkor a

$$y''(x) + f(x) = 0, \;\; 0 \leq x \leq 1$$
$$u(0) = a, \;\; u(1) = b.$$

feladat

$$y_0 = a,$$
$$-y_{\overline{x}x,i} = f_i, \;\; i = 1, \ldots, N-1,$$
$$y_N = b.$$

differenciasémája stabil és érvényes a következő becslés

$$\|\widetilde{\mathbf{y}}\|_{C(\widetilde{\omega}_h)} \leq \max\{|a|, |b|\} + \frac{1}{8}\|\overrightarrow{\mathbf{f}}\|_{C(\widetilde{\omega}_h)}.$$

Továbbá a séma konzisztencia rendje és konvergencia rendje 2.

Bizonyítás. A korábbi lemmákból és tételekből következik. ■

9-8. T *(Stabilitás+Approximáció=Konvergencia)*
A kezdetiérték-feladatokhoz hasonlóan a séma stabilitásából és az approximációból következik a konvergencia.

Bizonyítás. Ugyanis a stabilitásból $\|\tilde{\mathbf{z}}\| = \|\tilde{\mathbf{y}} - \tilde{\mathbf{u}}\| \leq M\|\widetilde{\mathbf{\Psi}}\|$ és az approximációból

$$\|\widetilde{\Psi}\| \to 0, \;\; h \to 0,$$

ezért $\|\tilde{\mathbf{y}} - \tilde{\mathbf{u}}\| \to 0, \;\; h \to 0$, ami ekvivalens azzal, hogy
minden i-re $(i = 0, \ldots, N) \quad y_i \to u(x_i), \;\; h \to 0$, vagyis a séma konvergens. ■

Megjegyzések.

1. A 6. tételbeli pontossági becslés előnyös tulajdonsága, hogy elvileg kiértékelhető az input adatokból. Ilyen becslést bonyolultabb feladatoknál nem tudunk megadni. Ahhoz, hogy az ε pontosságot biztosító h-t meghatározzuk f''-t és annak maximumát kellene meghatározni.

2. Gyakorlatban jól használható a Runge-fogás (a Richardson-extrapoláció legegyszerűbb formája). Számítsuk ki a megoldást h-t mindig felezve. Ha az $\mathbf{y^{(h)}}$ és $\mathbf{y^{(h/2)}}$ közelítések rendelkezésünkre állnak, akkor

$$e := \left\| \frac{4}{3} \left(\mathbf{y^{(h/2)}} - \mathbf{y^{(h)}} \right) \right\|_{C(\omega_h)}$$

adja a hiba becslését.

Irodalomjegyzék

[1] N. Sz. Bahvalov: *A gépi matematika numerikus módszerei*, Műszaki könyvkiadó, Budapest, 1977.

[2] Gergó Lajos: *Numerikus módszerek*, ELTE Eötvös Kiadó, Budapest, 2010

[3] Hegedűs Csaba: *Numerikus analízis jegyzet matematika tanároknak*, http://numanal.inf.elte.hu/ hegedus/na1.pdf

[4] Móricz Ferenc: *Differenciálegyenletek numerikus módszerei*, Polygon, Szeged, 1998

[5] G. Söderlind: *The Logarithmic Norm. History and Modern Theory*, BIT Numerical Mathematics (2006) 46

[6] H. R. Schwarz: *Numerical Analysis: A Comprehensive Introduction*, Wiley, New York, 1989

[7] M.N. Spijker: *Numerical Stability, Stability Estimates and Resolvent Conditions in the Numerical Solution of Initial Value Problems*, Lecture Notes, December 1998

[8] J. Stoer, R. Bulirsch: *Introduction to Numerical Analysis*, Springer, New York, 1980

[9] Stoyan Gisbert, Takó Galina: *Numerikus módszerek 1.*, ELTE-Typotex, Budapest, 1993

[10] Stoyan Gisbert, Takó Galina: *Numerikus módszerek 2.*, ELTE-Typotex, Budapest, 1995

[11] Stoyan Gisbert: *Numerikus matematika*, Egyetemi jegyzet, Typotex, Budapest, 2007

[12] T. Störm: *On Logarithmic Norms*, SIAM J. NUMER. ANAL. Vol. 12, No. 5, October 1975

[13] E. Süli, D. F. Mayers: *An Introduction to Numerical Analysis*, Univesity Press, Cambridge, 2003