Eötvös Loránd Tudományegyetem

Természettudományi Kar

Molnár Viktória

Matematika Bsc - Alkalmazott matematikus szakirány

Közönséges differenciálegyenletek numerikus megoldása

Egylépéses módszerek

Témavezető: Kurics Tamás

Alkalmazott Analízis és Számításmatematikai Tanszék

Budapest,

2010

Köszönetnyílvánítás

Szeretném kifejezni köszönetemet a szakdolgozat elkészítéséhez nyújtott segítségéért témavezetőmnek, Kurics Tamásnak.
Továbbá hálával tartozom tankörtársaimnak a LaTeX szövegszerkesztő alkalmazásánál nyújtott segítségükért.

Tartalomjegyzék

1. fejezet

Bevezetés

A gyakorlati életben sűrűn előfordulnak olyan feladatok, melyekhez differenciálegyenletek megoldása szükséges. Ezek között is előfordul olyan probléma, melynél analitikus módszerekkel nem tudunk célt érni. A számítástechnika fejlődésének köszönhetően ezeket a feladatokat is meg tudjuk oldani, vagy ha pontosan nem ismerjük a megoldást, de azért közelítő becslést kaphatunk róla.

Vegyünk például egy ejtőernyőst, aki $t = 0$ időpontban kiugrik egy repülőgépből és a kezdeti magassága legyen $y = y_0$. Az egyszerűség kedvéért feltesszük, hogy csak függőlegesen mozog. Az ugróra két erő hat: a gravitáció és a légellenállás. Newton második törvénye szerint a következő egyenletet kapjuk az ugró függőleges gyorsulására:

$$m\ddot{y} = -mg + mc\dot{y}^2,$$

ahol m az ejtőernyős tömege, g a gravitációs erő, c alaki tényező. A newtoni modellben az ejtőernyős mozgását ez a másodrendű differenciálegyenlet írja le. Ezt visszavezethetjük két elsőrendű differenciálegyenletre a következő módon: a sebességet jelöljük v-vel. Ekkor $v = \dot{y}$ miatt $\dot{v} = \ddot{y}$. Ha leosztunk m-mel, akkor a következő egyenletrendszert kapjuk:

$$\begin{aligned} \dot{y} &= \dot{v} \\ \dot{v} &= -g + cv^2. \end{aligned}$$

Ha hozzávesszük még a kezdeti feltételt is, akkor már egyértelmű megoldást kapunk. Az egyenletrendszert nem tudjuk megoldani analitikusan, így szükség van a numerikus számításokra. De mielőtt erre részletesebben kitérnénk, nézzük meg, hogy általános eset-

ben hogyan vezethető vissza egy p-edrendű közönséges differenciálegyenlet elsőfokú rendszerre. Legyen

$$F(t, y, \dot{y}, \ldots, y^{(p)}) = 0$$

adott kezdetiértékek mellett

$$y^{(s)}(t_0) = y_0^{(s)},$$

ahol $s = 0, 1, \ldots, p-1$. Az $y^{(p)}$ kifejezhető (az implicit függvény tétel segítségével) a p-edrendű egyenletből:

$$y^{(p)}(t) = G(t, y, \dot{y}, \ldots, y^{(p-1)}).$$

Vezessünk be új változókat: $z_1, z_2, \ldots, z_p$ a következő értékekkel:

$$\begin{aligned} z_1 &= y, \\ z_2 &= \dot{y}, \\ &\vdots \\ z_p &= y^{(p-1)}. \end{aligned}$$

Innen:

$$\begin{aligned} \dot{z}_1 &= \dot{y} = z_2, \\ \dot{z}_2 &= \ddot{y} = z_3, \\ &\vdots \\ \dot{z}_{p-1} &= y^{(p-1)} = z_p. \end{aligned}$$

Tehát

$$\begin{cases} \dot{z}_1(t) = z_2(t) \\ \dot{z}_2(t) = z_3(t) \\ \quad\vdots \\ \dot{z}_{p-1}(t) = z_p(t) \\ \dot{z}_p(t) = G(t, z_1(t), \ldots, z_p(t)) \end{cases}$$

a z-re egy elsőrendű közönséges egyenletrendszert kaptunk.

A továbbiakban a következő közönséges differenciálegyenletekből álló rendszert szeretnénk megoldani

$$y'(x) = f(x, y(x)), \tag{1.1}$$

ahol $f : \mathbf{R}^{n+1} \to \mathbf{R}^n$, és $x \in \mathbf{R}$, $y \in \mathbf{R}^n$. Feltesszük továbbá, hogy f folytonos az x-ben, és Lipschitz-folytonos az y-ban. Így a kezdeti feltétel $y(x_0) = y_0$ hozzávételével, a Picard–Lindelöf tétel alapján egyértelmű megoldást kapunk. Ha az (1.1) egyenlet jobb oldala nem függ az y-tól és például $n = 1$, akkor az

$$y' = f(x)$$

egyenlet ekvivalens az

$$y(x) = y_0 + \int_0^x f(t)\mathrm{d}t$$

integrálegyenlettel.

2. fejezet

Euler-módszerek

2.1. Explicit Euler-módszer

Nézzük (1.1)-et $n = 1$-re, és a $[0, 1]$ intervallumon keressük a megoldását. Osszuk fel ekvidisztánsan az intervallumot úgy, hogy jelölje $x_0 = 0$ a kezdőpontot, és $x_k = x_0 + kh$ a belső pontokat, $k = 0, \ldots, N$, ahol $h = 1/N$. Itt $N \geq 1$ egész szám, és h a lépésköz. Ha minden kis intervallumon közelítjük $y'(x_k)$-t a differenciálhányadosával, akkor a következőt kapjuk:

$$\frac{y(x_k + h) - y(x_k)}{h} \approx y'(x_k) = f(x_k, y(x_k)),$$

amiből úgy kapjuk a numerikus értékeket, ha egyenlőséget teszünk. Szokás a numerikus értékeket y_k-val jelölni, a pontos értékeket $y(x_k)$-val. Ezeket a jelöléseket alkalmazva kapjuk, hogy

$$\frac{y_{k+1} - y_k}{h} = f(x_k, y_k).$$

Ha megszorozzuk h-val, és átrendezzük az egyenletünket, akkor megkapjuk az úgynevezett haladó vagy explicit Euler-módszert:

$$y_{k+1} = y_k + hf(x_k, y_k).$$

Vezessünk be néhány definíciót:

1. Definíció. *Legyen $x^* \in [0, 1]$ egy rögzített pont. Osszuk fel a $[0, 1]$ intervallumot n részre úgy, hogy x^* az osztópontok között szerepeljen. Egy numerikus módszert az (1.1) megoldására akkor hívunk konvergensnek, ha egy adott függvényosztályból minden f-re,*

minden $y_0 = y(0)$ kezdeti értékre és minden rögzített x^* pontjára igaz, hogy ha $n \to \infty$, vagyis finomítjuk a felosztást, akkor $y_n \to y(x^*)$.

A konvergencia-vizsgálat érdekében definiáljuk a módszer lokális hibáját.

2. Definíció. *Az Euler-módszer képlethibájának vagy lokális hibájának hívjuk a következő mennyiséget:*

$$g_k := \frac{y(x_{k+1}) - y(x_k)}{h} - f(x_k, y(x_k)), \tag{2.1}$$

ahol $y(x)$ (1.1) pontos megoldását jelenti.

A lokális hibát becsülhetjük anélkül, hogy ismernénk a pontos megoldást. Alkalmazzuk kétszer a Lagrange-féle középérték tételt:

$$|g_k| = |y'(\xi_k) - y'(x_k)| = |y''(\tau_k)(\xi_k - x_k)| \leq h \max |y''(x)|,$$

ahol $\xi_k \in (x_k, x_{k+1})$ és $\tau_k \in (x_k, \xi_k)$.

3. Definíció. *A numerikus módszert konzisztensnek hívjuk, ha minden $0 \leq k \leq N$-re igaz $g_k = O(h^p)$, ahol $p > 0$. A legnagyobb egész p-t a módszer konzisztencia rendjének nevezzük.*

A lokális hibabecslésből azonnal látszik, hogy az Euler-módszer elsőrendben konzisztens.

4. Definíció. *A módszer globális hibájának nevezzük a pontos és a számított érték különbségét:*

$$e_k := y(x_k) - y_k \qquad\qquad k = 0, 1, \ldots, N.$$

A globális hiba becslése: Ha (2.1)-t beszorozzuk h-val és átrendezzük, a következő pontos értéket kifejezve kapjuk a következő egyenletet:

$$y(x_{k+1}) = y(x_k) + hf(x_k, y(x_k)) + hg_k.$$

Ebből kell kivonni a közelítő értékekkel kapott megoldást:

$$y_{k+1} = y_k + hf(x_k, y_k).$$

Vagyis e_{k+1}-re azt kapjuk, hogy

$$e_{k+1} = e_k + h(f(x_k, y(x_k)) - f(x_k, y_k)) + hg_k.$$

Kihasználva, hogy f Lipschitz-folytonos a második változójában, kapjuk a következő becslést:

$$|e_{k+1}| \leq |e_k| + hL_f|e_k| + h|g_k| = (1 + hL_f)|e_k| + h|g_k|.$$

Ezt tovább becsülve:

$$|e_{k+1}| \leq (1 + hL_f)^2|e_{k-1}| + h|g_k| + (1 + hL_f)h|g_{k-1}| \leq$$

$$\leq \ldots \leq (1 + hL_f)^{k+1}|e_0| + \sum_{i=0}^{k}(1 + hL_f)^{k-i}h|g_i| \leq$$

$$\leq (1 + hL_f)^{k+1}\big(|e_0| + \sum_{i=0}^{k} h|g_i|\big).$$

Az $(1 + hL_f)$-et felülről tudjuk becsülni $(e^h L_f)$-el az e^h Taylor sorfejtéséből adódóan, így

$$(1 + hL_f)^{k+1} \leq e^{(k+1)hL_f} \leq e^{NhL_f}$$

Nh pedig pont az intervallumunk hossza, ami jelen esetben 1. Ezáltal e^{L_f}-et kapjuk, ami egy rögzített C konstans. Ebből adódik a következő definíció:

5. Definíció. *Egy numerikus módszert stabilnak nevezünk, ha f Lipschitz-folytonos, és ha igaz a következő becslés:*

$$|e_{k+1}| \leq C\big(|e_0| + \sum_{i=0}^{k} h|g_i|\big),$$

ahol C konstans.

1. Tétel. *Egy numerikus módszer akkor konvergens, ha minden k-ra $e_k = O(h^p)$ teljesül. Ekkor a konvergencia rendje p.*

1. Állítás. *Ha a módszer p-edrendben konzisztens és stabil, akkor konvergens is, és a konvergencia rendje megegyezik a konzisztencia rendjével.*

2.2. Implicit Euler-módszer

Az Euler-módszer másik változatát úgy kaphatjuk, hogy $y'(x_{k+1})$-et helyettesítjük az Euler-módszernél látott differenciálhányadossal. Ezt az úgynevezett hátralépő vagy implicit Euler-módszernek nevezzük.

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

h-val való szorzás, és átrendezés után így néz ki az implicit Euler-módszer:

$$y_{k+1} = y_k + hf(x_{k+1}, y_{k+1}).$$

Hátránya az Euler-módszerrel szemben, hogy minden lépésben egy nemlineáris egyenletet kell megoldanunk. Állapítsuk meg a módszer rendjét! Ehhez tekintsük a g_k értékét:

$$g_k := \frac{y(x_{k+1}) - y(x_k)}{h} - f(x_{k+1}, y(x_{k+1})).$$

Szinte szó szerint elismételhetjük az Euler-módszernél látottakat, vagyis alkalmazzuk kétszer a Lagrange-féle középérték tételt:

$$|g_k| = |y'(\xi_k) - y'(x_{k+1})| = |y''(\tau_k)(\xi_k - x_{k+1})| \leq h \max |y''(x)|,$$

ahol $\xi_k \in (x_k, x_{k+1})$ és $\tau_k \in (\xi_k, x_{k+1})$ tetszőleges. A módszer így szintén elsőrendű.

3. fejezet

Javított Euler-módszerek

Felmerül a kérdés, hogy hogyan kaphatnánk magasabb rendben konzisztens megoldást. Tudnánk-e javítani az Euler-módszert?

Az (1.1) kezdeti érték feladatunk pontos megoldását így is megkaphatjuk:

$$y(x_{k+1}) - y(x_k) = \int_{x_k}^{x_{k+1}} f(t, y(t))\mathrm{d}t. \tag{3.1}$$

Viszont, ahogy már a korábbiakban említettük, nem mindig tudjuk ezt megoldani, így hát az integrált is numerikusan fogjuk közelíteni kvadratúraformulával.

3.1. Trapézformulával javított Euler-módszer

Maga a trapézformula:

$$\int_{x_k}^{x_{k+1}} F(t)\mathrm{d}t \approx \frac{h}{2}\big(F(x_k) + F(x_{k+1})\big)$$

melynek hibája $O(h^3)$. Így kapjuk, hogy

$$\int_{x_k}^{x_{k+1}} f(t, y(t))\mathrm{d}t = \frac{h}{2}\big(f(x_k, y(x_k)) + f(x_{k+1}, y(x_{k+1}))\big) + O(h^3).$$

Ha ezt írnánk be (3.1)-be, akkor nem explicit formulát kapnánk $y(x_{k+1})$ miatt. Ennek kiküszöbölésére az ötlet a következő: fejtsük Taylor-sorba $y(x_{k+1})$-t x_k körül. Tehát

$$y(x_{k+1}) = y(x_k) + hy'(x_k) + O(h^2).$$

Vegyük észre, hogy az itt szereplő $y'(x_k)$ egyenlő $f(x_k, y(x_k))$-val. És ha itt végig $y(x_k)$ közelítő értékével, vagyis y_k-val számolunk, és beírjuk az eddigi eredményeket a (3.1)-es

képletbe, akkor megkapjuk a trapézformulával javított Euler-módszert:

$$y_{k+1} = y_k + \frac{h}{2}(f(x_k, y_k) + f(x_{k+1}, y_k + hf(x_k, y_k)). \tag{3.2}$$

Lássuk, hogy ennek a módszernek mekkora a rendje: a rend meghatározásához ismét a lokális hibatagot kell vizsgálnunk, ami

$$g_k = \frac{y(x_{k+1}) - y(x_k)}{h} - \underbrace{\frac{1}{2}[f(x_k, y(x_k)) + f(x_{k+1}, y(x_k) + hf(x_k, y(x_k)))]}.$$

Kezdjük el először átalakítani a kapcsos részt. Tudjuk, hogy $y'(x_k) = f(x_k, y(x_k))$, és észrevesszük, hogy ha ezt a helyettesítést végrehajtanánk $y(x_k) + hf(x_k, y(x_k))$-ban, akkor f második argumentumában pont az $y(x_{k+1})$ Taylor-sorának elejét kapnánk. Ezért még hozzá kell vennünk $O(h^2)$-et, és végül bővítsük h-val ezt a részt. A kapott érték a következő:

$$\frac{h}{2}[f(x_k, y(x_k)) + f(x_{k+1}, y(x_{k+1}))] + O(h^3) = \int_{x_k}^{x_{k+1}} f(t, y(t))\mathrm{d}t + O(h^3).$$

Így a kapott hibatag:

$$g_k = \frac{y(x_{k+1}) - y(x_k)}{h} - \frac{1}{h}\int_{x_k}^{x_{k+1}} \underbrace{f(t, y(t))}_{y'(t)}\,\mathrm{d}t + O(h^2) =$$

$$= \frac{y(x_{k+1}) - y(x_k)}{h} - \frac{1}{h}(y(x_{k+1}) - y(x_k)) + O(h^2) = O(h^2).$$

Ezek alapján a trapézformulával javított Euler-módszer másodrendben konzisztens.

3.2. Érintőformulával javított Euler-módszer

Másik módszer a javításra az érintő formula alkalmazása, ami pedig a következő:

$$\int_{x_k}^{x_{k+1}} F(t)\mathrm{d}t \approx hF(x_{k+\frac{1}{2}}).$$

Ezt alkalmazva a függvényünkre kapjuk, hogy

$$\int_{x_k}^{x_{k+1}} f(t, y(t))\mathrm{d}t = hf\big(x_{k+\frac{1}{2}}, y(x_{k+\frac{1}{2}})\big) + O(h^3).$$

Ugyanúgy mint az előbb, fejtsük Taylor sorba $y(x_{k+\frac{1}{2}})$-t x_k körül:

$$\int_{x_k}^{x_{k+1}} f(t, y(t))\mathrm{d}t = hf(x_{k+\frac{1}{2}}, y(x_k) + \frac{h}{2}f(x_k, y(x_k))) + O(h^3).$$

Az ötlet megint ugyanaz, vagyis cseréljük ki a pontos értékeket a közelítőkre. Tehát az érintőformulával javított Euler-módszer a következő lesz:

$$y_{k+1} = y_k + hf(x_{k+\frac{1}{2}}, y_j + \frac{h}{2}f(x_k, y_k)).$$

Mivel egy ugyanolyan $O(h^3)$ hibájú módszerrel javítottunk, ezért az a sejtésünk, hogy a javítás mértéke is ugyanolyan lesz. Lássuk precízen g_k értékét:

$$g_k = \frac{y(x_{k+1}) - y(x_k)}{h} - \underbrace{f(x_{k+\frac{1}{2}}, y(x_k)) + \frac{h}{2}f(x_k, y(x_k))}.$$

Vessük be ugyanazokat a trükköket, mint az trapézformulával javított módszernél. A kapcsos rész így alakul:

$$hf\big(x_{k+\frac{1}{2}}, y(x_k + \frac{1}{2})\big) + O(h^2).$$

Az egészet tekintve ismét azt kapjuk, hogy másodrendű a javítás, mivel

$$g_k = \frac{y(x_{k+1}) - y(x_k)}{h} - \frac{1}{h}\int_{x_k}^{x_{k+1}} \underbrace{f(t, y(t))}_{y'(t)}\,\mathrm{d}t + O(h^2) =$$

$$= \frac{y(x_{k+1}) - y(x_k)}{h} - \frac{1}{h}(y(x_{k+1}) - y(x_k)) + O(h^2) = O(h^2).$$

3.3. θ-módszer

Az Euler- és a trapéz-módszer speciális esetei az úgynevezett θ-módszernek:

$$y_{k+1} = y_k + h[\theta f(x_k, y_k) + (1-\theta)f(x_{k+1}, y_{k+1})], \qquad k = 0, 1, \ldots.$$

A $\theta = 1$ és $\theta = \frac{1}{2}$ esetben az eredeti képletek adódnak. Különböző $\theta \in [0, 1]$-re különböző módszereket kapunk, ami egyedül $\theta = 1$-ra lesz explicit, minden más esetben pedig implicit. Például $\theta = 0$ esetben az implicit Euler-módszert kapjuk vissza. Tekintsük a lokális hibatagot, a rend meghatározásához:

$$g_k = \frac{y(x_{k+1}) - y(x_k)}{h} - \big(\theta f(x_k, y_k) + (1-\theta)f(x_{k+1}, y_{k+1})\big) =$$

$$= \frac{y(x_{k+1}) - y(x_k)}{h} - \big(\theta y'(x_k) + (1-\theta)y''(x_{k+1})\big).$$

Használjuk fel $y(x_{k+1})$ sorfejtését:

$$y(x_k) + hy'(x_k) + \frac{1}{2}h^2y''(x_k) + \frac{1}{6}h^3y^{(3)}(x_k) + O(h^4).$$

Emiatt a hiba:

$$g_k = \frac{1}{h}[y(x_k) + hy'(x_k) + \frac{1}{2}h^2y''(x_k) + \frac{1}{6}h^3y^{(3)}(x_k) - y(x_k)] -$$

$$-\big(\theta y'(x_k) + (1-\theta)[y'(x_k) + hy''(x_k) + \frac{1}{2}h^2y^{(3)}(x_k)]\big) + O(h^3) =$$

$$= \big(\theta - \frac{1}{2}\big)hy''(x_k) + \big(\frac{1}{2}\theta - \frac{1}{3}\big)h^2y^{(3)}(x_k) + O(h^3).$$

Tehát a módszer másodrendű, ha $\theta = \frac{1}{2}$, különben meg elsőrendű.

4. fejezet

Runge–Kutta-módszerek

Az eddigiekben láthattuk, hogy a kvadratúra formulák segítségével javítani tudtunk a módszer rendjén. Ezen az ötleten alapszanak a Runge–Kutta-módszerek is. Tekintsük újra (3.1)-et. Ezt tovább alakítva általánosan kapjuk a következőket:

$$y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} f(t, y(t))\mathrm{d}t = h \int_0^1 f(x_n + ht, y(x_n + ht))\mathrm{d}t.$$

Ha helyettesítjük az így kapott integrált valamilyen kvadratúrával, és a közelítő értékeket beírjuk, akkor újabb módszereket nyerhetünk:

$$y_{n+1} = y_n + h \sum_{j=1}^{m} b_j f(x_n + c_j h, y(x_n + c_j h)), \qquad n = 0, 1, \ldots$$

ahol b_j-k és c_j-k megfelelő súlyok. Ezek a számok határozzák majd meg, hogy melyik módszerről is beszélünk. Viszont nem ismerjük y értékeit az $x_n + c_1 h, x_n + c_2 h, \ldots$ pontokban. Az explicit Runge–Kutta-módszerek ötlete, hogy az $y(x_n + c_j h)$ értékeket approximáljuk a következő módon. Kezdetnek legyen $c_1 = 0$ és vezessük be rekurzívan a következő számokat:

$$\begin{aligned}
\xi_1 &= y_n, \\
\xi_2 &= y_n + h a_{2,1} f(x_n, \xi_1), \\
\xi_3 &= y_n + h a_{3,1} f(x_n, \xi_1) + h a_{3,2} f(x_n + c_2 h, \xi_2), \\
&\vdots \\
\xi_m &= y_n + h \sum_{j=1}^{m-1} a_{m,j} f(x_n + c_j h, \xi_j),
\end{aligned}$$

íly módon az

$$y_{n+1} = y_n + h \sum_{j=1}^{m} b_j f(x_n + c_j h, \xi_j) \tag{4.1}$$

képlet adódik. Az $A = (a_{i,j})_{i,j=1,2,\ldots,m}$ mátrixot Runge–Kutta-mátrixnak nevezzük, és a hiányzó elemeit nullának definiáljuk. A további együtthatókat vektorokba rendezhetjük:

$$b = [b_1 b_2 \ldots b_m], \qquad c = [c_1 c_2 \ldots c_m].$$

Azt mondjuk, hogy ez a módszer úgynevezett m-lépcsős explicit RK-módszer, mert ennyi lépcsőben értékeljük ki f-et, hogy megkaphassuk y_{n+1} értékét. Ennek a módszernek egy természetesen adódó általánosítása, ha az A együttható mátrix nem csupán alsó háromszög. Ilyenkor minden lépésben egy nemlineáris rendszert kell megoldanunk. Ezt hívjuk implicit Runge–Kutta-módszernek. Mindkét típusú módszernél az együtthatók tárolására úgynevezett Butcher-táblát használunk:

$$\begin{array}{c|c} c^T & A \\ \hline & b \end{array}.$$

Továbbiakban csak az explicit RK-módszerekről lesz szó. Ebben az esetben így néz ki kifejtve a Butcher-tábla:

$$\begin{array}{c|ccccc} 0 & & & & & \\ c_2 & a_{2,1} & & & & \\ c_3 & a_{3,1} & a_{3,2} & & & \\ \vdots & \vdots & & \ddots & & \\ c_m & a_{m,1} & a_{m,2} & \ldots & a_{m,m-1} & \\ \hline & b_1 & b_2 & \ldots & b_{m-1} & b_m \end{array}$$

A módszer stabilitásáról egy tételt fogalmazunk meg [6] könyv alapján, ami hasonló módszerrel látható be, mint a javított Euler-módszereknél.

2. Tétel. *(általános explicit Runge–Kutta-módszer stabilitása)*
Legyen f a második változójában lokálisan Lipschitz, L_f Lipschitz-állandóval. Ekkor a (4.1)*-ben definiált módszer stabil, mert érvényes a következő becslés:*

$$|e_k| \leq e^{L_\Phi} \Big(|e_0| + \sum_{i=0}^{k-1} |g_i| h \Big).$$

Itt

$$L_\Phi = L_f\Big(\sum_{j=1}^{m}|c_j| + P_{m-1}(hL_f)\Big),$$

és a $P_{m-1}(hL_f)$ *tag a* hL_f *mennyiség* $(m-1)$*-edfokú polinomja, amire igaz, hogy* $P_{m-1}(hL_f) = O(hL_f)$.

Már csak az maradt hátra, hogy hogyan is válasszuk meg a c_j-ket, b_j-ket, $a_{i,j}$-ket, és hogy mekkora lesz az egyes módszerek rendje m függvényében.

4.1. Kétlépcsős Runge–Kutta-módszer

Nézzük az egyik legegyszerűbb nemtriviális esetet, amikor $m = 2$, vagyis egy kétlépcsős RK-módszert:

$$y_{n+1} = y_n + hb_1 f(x_n + c_1 h, y_n) + hb_2 f(x_n + c_2 h, y_n + ha_{2,1} f(x_n, y_n)). \tag{4.2}$$

Az összeg utolsó tagját fejtsük Taylor sorba (x_n, y_n)-körül:

$$f(x_n+c_2h, y_n+ha_{2,1}f(x_n,y_n)) = f(x_n,y_n)+h[c_2f_x(x_n,y_n)+a_{2,1}f_y(x_n,y_n)f(x_n,y_n)]+O(h^2),$$

majd ezt visszaírva (4.2)-be kapjuk a következő összefüggést:

$$y_{n+1} = y_n + h(b_1+b_2)f(x_n,y_n) + h^2b_2[c_2f_x(x_n,y_n) + a_{2,1}f_y(x_n,y_n)f(x_n,y_n)] + O(h^3). \tag{4.3}$$

Össze kell vetnünk ezt az egyenletet, a pontos értékkel ugyanabban az (x_n, y_n) pontban. Az y első deriváltját ki tudjuk fejezni f-el, de lássuk, hogy néz ki a második derivált:

$$y'' = f_x + f_y f.$$

Így, ha az x_{n+1}-ben kapott pontos megoldás sorfejtését $\tilde{y}(x_{n+1})$-el jelöljük, akkor azt kapjuk, hogy

$$\tilde{y}(x_{n+1}) = y_n + hf(x_n,y_n) + \frac{1}{2}h^2[f_x(x_n,y_n) + f_y(x_n,y_n)f(x_n,y_n)] + O(h^3).$$

Azt szeretnénk, ha ez megegyezne (4.3)-al. Innen kapjuk az egyenletrendszert az együtthatókra:

$$\begin{cases} b_1 + b_2 = 1 \\ b_2c_2 = \frac{1}{2} \\ a_{2,1} = c_2. \end{cases}$$

Ebben a konkrét esetben az egyenletrendszernek nincs egyértelmű megoldása, mint ahogy általában sincs, mert több az ismeretlen, mint az egyenlet. De azért példának mutatunk két megoldást az alábbi két Butcher-táblában:

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

Ezekben a táblázatokban rendre felismerhetjük a trapéz- és érintőformulával javított Euler-módszert, aminek már ismerjük a rendjét. Így a kétlépcsős RK-módszer stabil és a rendje megegyik a lépcsők számával.

4.2. Háromlépcsős Runge–Kutta-módszer

Azt reméljük, hogy a háromlépcsős módszer nagyobb rendben lesz konzisztens. Az előző technika segítségével kiszámolható a háromlépcsős módszer Butcher-táblája. Másodrendig kell Taylor-sorba fejteni az összeg második és harmadik tagját, és kifejezni $y^{(3)}$-ot, vagyis y harmadik deriváltját, ugyanis szükség lesz rá az $\tilde{y}(x_{n+1})$ sorfejtésénél. Majd ezeket kell összevetni. Már itt $m = 3$ esetén is elég hosszadalmas kiszámítani az egyenletrendszert, de későbbiekben mutatunk rá egy egyszerűbb sémát. A kapott egyenletrendszer a következő:

$$\begin{cases} b_1 + b_2 + b_3 = 1 \\ b_2 c_2 + b_3 c_3 = \frac{1}{2} \\ b_2 c_2^2 + b_3 c_3^2 = \frac{1}{3} \\ b_3 a_{3,2} c_2 = \frac{1}{6}. \end{cases}$$

Több megoldás lehet itt is, példának nézzük a következő két Butcher-táblát:

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array} \qquad \begin{array}{c|ccc} 0 & & & \\ \frac{2}{3} & \frac{2}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \end{array}$$

Ezek alapján általánosan azt vesszük észre, hogy $\sum_{j=1}^{m-1} b_j = 1$, ugyanis emiatt lesz egyáltalán konzisztens a módszer. A következő egyenlőségek pedig megkönnyítik a szá-

molást: $c_2 = a_{2,1}$, $c_3 = a_{3,1} + a_{3,2}$, ..., vagyis tegyük fel ($c_i = \sum_{j=1}^{i-1} a_{i,j}$)-t minden explicit Runge–Kutta-módszerre.

Ezek után lássuk a háromlépcsős módszer rendjét. A $y(x_{n+1}) - y(x_n) - h\sum_{j=1}^{3} b_j f(x_n + c_j h, \xi_j)$ sorfejtésének hibatagja $O(h^4)$, tehát a g_n lokális hiba harmadrendű és a módszer globális hibája $O(h^3)$.

A p-edrendű módszer lokális hibája minden lépcsőben ugyanannyi (persze h-tól függően), ezért nézhetjük az első intervallum végpontjaiban:

$$e(h) = y(x_0 + h) - y_1.$$

Fejtsük Taylor-sorba $e(h)$-t 0-körül

$$e(h) = e(0) + he'(0) + \cdots + \frac{h^p}{p!} e^{(p)}(\theta h),$$

ahol $0 < \theta < 1$, és $e(0) = e'(0) = \cdots = e^{(p)}(0) = 0$. Viszont $e^{(p)}(h)$ pontos értékét a következő alakban kapjuk:

$$e^{(p)}(h) = E_1(h) + hE_2(h),$$

ahol $E_1(h)$ és $E_2(h)$ tartalmazzák f parciális deriváltjait $(p-1)$- és p-edrendig. Mivel $e^{(p)}(0) = 0$, ezért $E_1(0) = 0$ és mivel f parciális deriváltjai adottak, ezért $E_1(h)$ csak $O(h)$ és E_2 csak $O(1)$ lehet. Tehát létezik olyan C konstans, hogy $|e(h)| \leqq Ch$ és

$$|e(h)| \leqq C\frac{h^{p+1}}{p!}.$$

Azt hihetnénk, hogy minél nagyobb a lépcsőszám, annál nagyobb lesz a rend, de azért ez nem egészen van így. A következő összefüggések ismertek a lépcsőszám m és az elérhető $p(m)$ rend között:

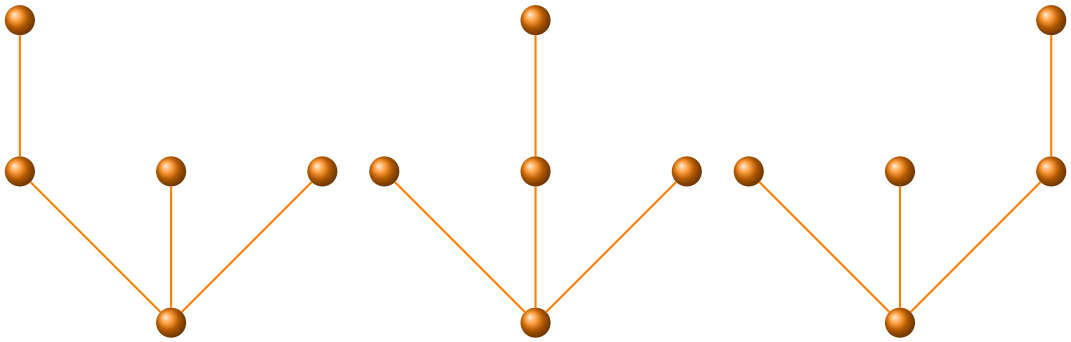
$$\begin{aligned} 1 \leq m \leq 4: &\quad p(m) = m \\ m \geq 5: &\quad p(m) \leq m - 1 \\ m \geq 8: &\quad p(m) \leq m - 2 \\ m \geq 10: &\quad p(m) \leq m - 3. \end{aligned}$$

Az m és $p(m)$ közötti hézag éppen azzal kapcsolatos, hogy f függvény y-tól is függ.

4.3. Általános Runge–Kutta-módszer együtthatói

Már ilyen kis lépcsőszámnál is nehezen számolhatók ki a Runge–Kutta-módszer együtthatói, de szerencsére van rá minta, hogy ez általánosságban hogyan is működik. Meglepő lehet, de nem más, mint a gráfelmélet adott választ a kérdéseinkre.

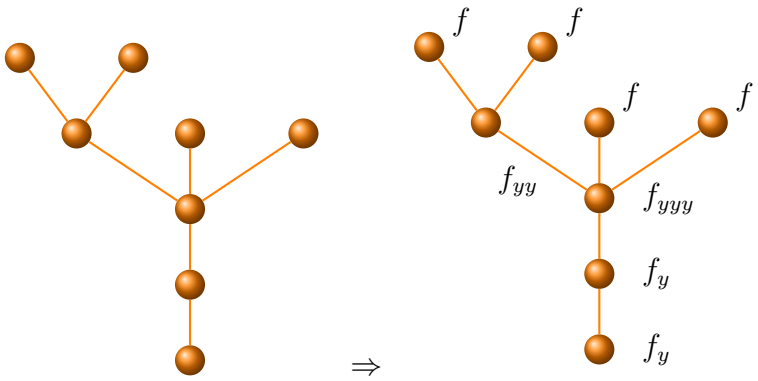
Legyen T gyökeres fák egy halmaza, ami alatt olyan irányítatlan körmentes gráfokat értünk, aminek van egy kitüntetett csúcsa: a gyökér. Továbbá egy fa (t) rendjén a csúcsainak a számát fogjuk érteni. Ha nagyobb, mint elsőrendű fáról van szó, akkor levelek alatt azokat a csúcsokat értjük, ami nem gyökér, és ami pontosan egy éllel csatlakozik a fához. A fán adott egy rendezés, miszerint a gyökércsúcs legyen a legalsó, a belőle 1 lépésben elérhető csúcsok legyenek a gyerekei. Szülő csúcsok pedig értelemszerűen az egy lépéssel elérhető leljebb lévő csúcsok. Két azonos rendű fát ekvivalensnek mondunk, ha az utak ugyanazt a mintát követik a fa tetejétől az aljáig. A következő ábra ekvivalens fákat ábrázol.



Szükségünk van y egyre magasabb deriváltjaira - elég arra az esetre szorítkozni, mikor f nem függ x-től -, amik 1-től 4-ig a következőképp néznek ki:

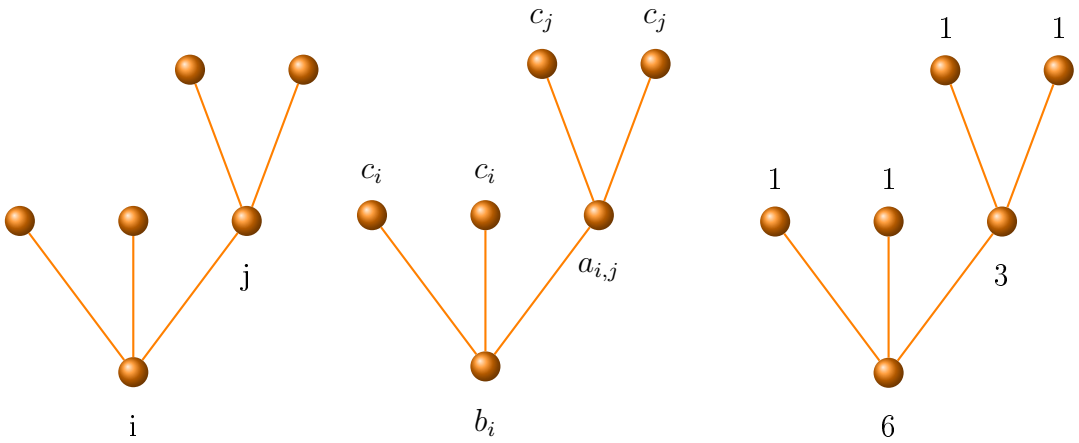
$$y' = f(y)$$
$$y'' = f_y(y)f(y)$$
$$y^{(3)} = f_{yy}(y)[f(y)]^2 + [f_y(y)]^2 f(y)$$
$$y^{(4)} = f_{yyy}(y)[f(y)]^3 + 4f_{yy}(y)f_y(y)[f(y)]^2 + [f_y(y)]^3 f(y)$$

Ha mondjuk a k-adik deriváltra van szükségünk, akkor vegyünk egy k csúcsból álló fát az összes lehetséges nem ekvivalens módon. Egy fa csúcsainak feleltessük meg f annyiadik deriváltját, ahány gyereke van, majd ezeket szorozzuk össze minden egyes fán, és adjuk össze őket. Az ábrán egy példa látható:



Ha már a deriváltakat előállítottuk, akkor már csak az együtthatókat kell nekik megfeleltetni. Egyrészről minden fa meghatároz egy polinomot - jelöljük $\Phi(t)$-vel -, amiből az együtthatókat nyerjük majd ki, másrészről pedig minden t-hez hozzárendelünk egy természetes számot, amit jelöljünk $\gamma(t)$-vel. Lássuk a konstrukciót!

Indexeljük meg a csúcsokat - kivéve a leveleket - úgy, hogy a gyökér legyen az i, a többi pedig rendre $j, k, \ldots$. Most rendeljük hozzá a csúcsokhoz az együtthatókat, úgy, hogy a gyökér legyen b_i, a többi csúcs, ami nem levél, pedig legyen $a_{j,k}$ ahol j a szülőjének az indexe, k pedig a csúcs saját indexe. A levelekben lévő együtthatók pedig legyenek c_k-k, ha a k indexű csúcsból induló él tartja őket. $\Phi(t)$-t úgy kapjuk, hogy összeszorozzuk a fában lévő együtthatókat, majd összeadjuk öket az összes lehetséges módon. $\gamma(t)$-hez az kell, hogy megnézzük mennyi a gyökér rendje, majd elhagyjuk azt, és az így kapott gráfban vegyük a legalsó csúcsok rendjét, majd ezeket is elhagyva folytatjuk az algoritmust. Ezeknek a számoknak a szorozata lesz $\gamma(t)$.



$$\Phi(t) = \sum_{ij} b_i c_i^2 a_{i,j} c_j^2 \qquad \gamma(t) = 1 * 1 * 3 * 1 * 1 * 6 = 18$$

Az egyenletek, amiket ezekből kapunk a következőképp néznek ki:

$$\Phi(t) = \frac{1}{\gamma(t)}$$

Legjobban megérteni talán egy példán keresztül lehet. Ez alapján írjuk fel a négylépé-

ses RK-módszerben kapott egyenlet rendszert.

4.4. Négylépcsős Runge–Kutta-módszer

A (4.1) képletben m helyére 4-et írva a módszer explicit, ezért minden $i > j$-re $a_{i,j} = 0$. Az eddigiek alapján a következő egyenletrendszert kapjuk:

$$\begin{cases} b_1 + b_2 + b_3 + b_4 = 1 \\ b_2c_2 + b_3c_3 + b_4c_4 = \frac{1}{2} \\ b_2c_2^2 + b_3c_3^2 + b_4c_4^2 = \frac{1}{3} \\ b_3a_{3,2}c_2 + b_4a_{4,2}c_2 + b_4a_{4,3}c_3 = \frac{1}{6} \\ b_2c_2^3 + b_3c_3^3 + b_4c_4^3 = \frac{1}{4} \\ b_3c_3a_{3,2}c_2 + b_4c_4a_{4,2}c_2 + b_4c_4a_{4,3}c_3 = \frac{1}{8} \\ b_3a_{3,2}c_2^2 + b_4a_{4,2}c_2^2 + b_4a_{4,3}c_3^2 = \frac{1}{12} \\ b_4a_{4,3}a_{3,2}c_2 = \frac{1}{24}. \end{cases}$$

A második, harmadik és ötödik egyenletből kifejezhető b_2, b_3, b_4, a negyedik, hatodik, hetedik egyenletből pedig $a_{3,2}, a_{4,2}, a_{4,3}$. Az utolsó egyenletbe való behelyettesítés után kapjuk, hogy $c_4 = 1$. A sok megoldásból mutatunk két példát:

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

0				
$\frac{1}{4}$	$\frac{1}{4}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	1	−2	2	
	$\frac{1}{6}$	0	$\frac{2}{3}$	$\frac{1}{6}$

Ezek közül az elsőt nevezik a klasszikus (negyedrendű) Runge–Kutta-módszernek, mert maga Kutta határozta meg őket és valóban negyedrendben konzisztens a módszer.

4.5. Folytonos Runge–Kutta képletek

Ha ki szeretnénk rajzoltatni a megoldást, vagy a megoldás szélsőértékeit keressük, akkor nem érdemes a h lépéstávolságot nagyon kicsire választani, hiszen ez túl sok függvénykiértékeléssel jár. Ilyenkor előnyös folytonos Runge–Kutta képleteket használni. Ilyenkor

a ξ_j számokat ugyanúgy választjuk, mint eddig, viszont a b_j súlyok már nem számok lesznek, hanem a t változó folytonos függvényei, amivel a képlet a következő:

$$y_{n+t} = y_n + h\sum_{j=1}^{m} b_j(t) f(x_n + c_j h, \xi_j) \qquad 0 \le t \le 1. \tag{4.4}$$

Itt $y_{n+t} = y_n(t)$ az $y(x_n + th)$ közelítésére szolgál. Azt szeretnénk, hogy a $t = 0$ és $t = 1$ helyen ne változzon meg a hagyományos Runge–Kutta képlet, és a képlet hibája se változzon meg menet közben, vagyis minden $0 < t < 1$-re ugyanakkora maradjon.

Példaként vizsgáljuk meg a másodrendű, két lépcsőjű, folytonos Runge–Kutta képletek esetét:

$$y_{n+t} = y_n + hb_1(t) f(x_n + c_1 h, y_n) + hb_2(t) f(x_n + c_2 h, y_n + ha_{2,1} f(x_n, y_n)) \qquad 0 \le t \le 1. \tag{4.5}$$

Először biztosítsuk, hogy $t = 0$-ra és $t = 1$-re a hagyományos képletet kapjuk meg. Az első követelmény azt jelenti, hogy

$$b_1(0) = b_2(0) = 0.$$

A második feltételhez ugyanazt csináljuk, mint amikor a kétlépcsős képletet határoztuk meg. Taylor sorfejtések összevetéséből kapjuk a következőket:

$$b_1(1) + b_2(1) = 1, \qquad b_2(1)a_{2,1} = \frac{1}{2}.$$

Ebből következik, ahogy a nem folytonos képletben, úgy itt is a lokális hiba másodrendű. Az általános explicit Runge–Kutta-módszer stabilitásáról szóló tételből következik, hogy $y_n, y_{n+1}, \ldots$ értékek hibája is másodrendű. Ezután (4.4) már folytonos interpolációt ad $y_n = y(x_n) + O(h^2)$ és $y_{n+1} = y(x_{n+1}) + O(h^2)$ között. Már csak azt kell elérnünk, hogy $0 < t < 1$-re y_{n+t} eltérése $y(x_n + th)$-tól másodrendű legyen. A (4.5) képletből és az $y_n = y(x_n) + O(h^2)$-ből kiindulva kapjuk, hogy

$$\begin{aligned}
y_{n+t} &= y(x_n) + O(h^2) + hb_1(t) f(x_n + c_1 h, y(x_n) + O(h^2)) + \\
&+ hb_2(t) f(x_n + c_2 h, y(x_n) + O(h^2) + ha_{2,1} f(x_n, y(x_n) + O(h^2))) = \\
&= y(x_n) + h[b_1(t) + b_2(t)] f(x_n, y(x_n)) + O(h^2) = \\
&= y(x_n) + hty'(x_n) + O(h^2) = y(x_n + th) + O(h^2)
\end{aligned}$$

kell, hogy teljesüljön, vagyis

$$b_1(t) + b_2(t) = t, \qquad 0 \le t \le 1.$$

A legegyszerűbb lehetőség ahhoz, hogy az eddigi eredmények mind teljesüljenek:

$$b_1(t) \equiv 0, \qquad b_2(t) = t, \qquad a_{2,1} = \frac{1}{2}.$$

Harmadrendű módszer szerkesztéséhez tekintsük (4.4)-ben $m = 3$ esetet, és $t = 0$-ra írjuk elő, hogy

$$b_1(0) = b_2(0) = b_3(0) = 0$$

legyen, majd $t = 1$-re követeljük meg a háromlépcsős Runge–Kutta módszernél kapott egyenletrendszert. Ezzel biztosítjuk a az y_n értékek kiszámításának stabilitását, valamint a harmadrendűségüket. Az előző módszer analógiájára $0 < t < 1$-re y_n helyére $y(x_n) + O(h^3)$-at írva kapjuk, hogy

$$y_{n+t} = y(x_n) + O(h^3) + h\sum_{j=1}^{3} b_j(t) f(x_n + c_j h, \xi_j).$$

Ezt a képlépcsős módszerhez hasonlóan sorbafejtve kapjuk, hogy

$$y_{n+t} = y(x_n) + h[b_1(t) + b_2(t) + b_3(t)] f(x_n, y(x_n)) + [b_2(t)a_{2,1} + b_3(t)(a_{3,1} + a_{3,2})] f_x(x_n, y(x_n)) + O(h^3)$$

kell, hogy megegyezzen a következővel:

$$y(x_n) + hty'(x_n) + \frac{h^2 t^2}{2} y''(x_n) + O(h^3) = y(x_n + th) + O(h^3).$$

Összességében a következő feltételek adódnak:

$$\begin{cases} b_1(t) + b_2(t) + b_3(t) = t \\ b_2(t)a_{2,1} + b_3(t)(a_{3,1} + a_{3,2}) = \frac{t^2}{2} \\ b_2(1)a_{2,1}^2 + b_3(1)(a_{3,1} + a_{3,2})^2 = \frac{1}{3} \\ b_3(1)a_{3,2}a_{2,1} = \frac{1}{6}. \end{cases}$$

Ennek egy megoldása a következő lehet:

$$a_{2,1} = \frac{1}{2}, \qquad a_{3,1} = -1, \qquad a_{3,2} = 2$$

jön a sima háromlépcsős módszerből, és

$$b_1(t) = t(1 - \frac{5}{6}t), \qquad b_2(t) = \frac{2}{3}t^2, \qquad b_3(t) = \frac{1}{6}t^2.$$

5. fejezet

Beágyazott módszerek

Ebben a fejezetben szeretnénk minél jobb lokális hiba becslést kapni az egylépéses módszerekre. Ahogy az Euler-módszernél is láthattuk, a globális hibát viszonylag könnyen tudjuk becsülni. Nem ez a helyzet a lokális hibánál, ugyanis a közelítő eredmények hibáját sajnos csak ritkán tudjuk becsülni. Ezért fontos a számítás során, hogy a numerikus eredmények birtokában azonnal hibabecslést készítsünk. De minden esetben célszerű, hogy összehasonlítási érték is álljon a rendelkezésünkre.

5.1. Richardson-extrapoláció

Richardson ötletének lényege, hogy h függvényeként kell kezelni a hibát. Tegyük fel, hogy egy adott kezdeti érték (x_0, y_0) és egy adott h lépésköz mellett használunk egy p-edrendű RK-módszert. Kiszámolunk két lépést, aminek az eredménye legyen y_1 és y_2. Ezután ugyanazzal a kezdeti értékből indulva kiszámoljuk $(2h)$ lépésközzel tett nagy lépés eredményét, amit jelöljünk ω-val. y_1 globális hibája ismert:

$$e_1 = y(x_0 + h) - y_1 = Ch^{p+1} + O(h^{p+2}),$$

ahol C tartalmazza az úgynevezett hiba együtthatókat, és y_0 deriváltjait $p+1$-edrendig bezárólag. Az y_2 hibája két részből tevődik össze. Az első lépésből származó hibából $(I + h\frac{\delta f}{\delta y} + O(h^2))e_1$ illetve a második lépés lokális hibájából, ami pont akkora, mint e_1. Tehát a következőt kapjuk:

$$e_2 = y(x_0+2h)-y_2 = (I+O(h))Ch^{p+1}+(C+O(h))h^{p+1}+O(h^{p+2}) = 2C(h^p+1)+O(h^{p+2}).$$

A nagy lépés hibája:

$$y(x_0 + 2h) - \omega = C(2h)^{p+1} + O(h^{p+2})$$

Az előző egyenletből kifejezzük ki az ismeretlen C-t, és így extrapolálunk egy jobb értéket $y(x_0 + 2h)$-hoz, amit jelöljünk $\hat{y}$-al. A következő tétel az [5] könyvben található.

3. Tétel. *Tegyük fel, hogy egy p-edrendű RK módszerrel tett két h méretű lépés közelítő értéke y_2, és ω az egy nagy, $2h$ méretű lépéssel tett lépés eredménye, akkor az y_2 hibáját a következőképp határozzuk meg:*

$$y(x_0 + 2h) - y_2 = \frac{y_2 - \omega}{2^p - 1} + O(h^{p+2})$$

és

$$\hat{y}_2 = y_2 + \frac{y_2 - \omega}{2^p - 1}$$

egy $p + 1$-edrendű közelítés $y(x_0 + 2h)$-re.

5.2. Runge–Kutta–Fehlberg-módszerek

Az egyik nevezetes módszer a hibabecslésre a beágyazott RK-módszer. A lényege, hogy két különböző rendű módszert futtatunk párhuzamosan, így kapjuk a hibabecsléseket. Ha két azonos rendű módszert futtatnánk, akkor előfordulhatna, hogy egyszer az egyiknek lesz kisebb hibája, másszor pedig a másiknak. A két különböző rendű módszernél az egyik hibája elhanyagolható lesz.

Fehlberg ötlete a következő volt: a p-edrendű módszernek legyenek ugyanazok a c_j, $a_{i,j}$ együtthatói, mint a $p + 1$-edrendű módszernek. Ekkor viszont a b_j-k különbözők lesznek, ezért ezekre vezessük be $\hat{b}_j$ jelölést. Mostantól ezt úgy hívjuk, hogy $p + 1(p)$ módszer. Butcher-táblában pedig a következőképp jelöljük:

c^T	A
	b
	$\hat{b}$

Ez a módszer sokkal előnyösebb Richardson ötletéhez képest, ugyanis nincs szükség újabb függvénykiértékelésekre az azonos együtthetők miatt. Nézzük meg, hogyan is néz

ki a 3(2) módszer. Vegyünk egy harmadrendű módszert, például az előző fejezetben tárgyaltat. Ebbe ágyazzunk be egy másodrendű módszert. Szükségünk lesz $\hat{b}_j$-kre. A következő összefüggést kapjuk a másodrendű módszer együtthatóira:

$$\begin{cases} \hat{b}_1 + \hat{b}_2 + \hat{b}_3 = 1 \\ \frac{1}{2}\hat{b}_2 + \hat{b}_3 = \frac{1}{2}. \end{cases}$$

Következésképp a módszer:

$$y_{n+1} = y_n + h(\hat{b}_1 f(\xi_1) + \hat{b}_2 f(\xi_2) + \hat{b}_3 f(\xi_3)).$$

Rögzítsük le $\hat{b}_2$-t $\frac{1}{2}$-nek, így $\hat{b}_1 = \hat{b}_2 = \frac{1}{4}$ lesz, a módszer Butcher-táblázata pedig:

0			
$\frac{1}{2}$	$\frac{1}{2}$		
1	-1	2	
b	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
$\hat{b}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

A két módszer segítségével becsüljük a hibát:

$$g_{n+1} = \hat{y}_{(n+1)} - y(x_{n+1}) = h((\hat{b}_1 - b_1)f(\xi_1) + (\hat{b}_2 - b_2)f(\xi_2) + (\hat{b}_3 - b_3)f(\xi_3)) =$$

$$= h(d_1 f(\xi_1) + d_2 f(\xi_2) + d_3 f(\xi_3)).$$

Ebben az esetben $d_1 = \frac{1}{12}$, $d_2 = -\frac{1}{6}$, $d_3 = \frac{1}{12}$.

Nézzük most meg az előző fejezetben tárgyalt klasszikus negyededrendű módszert, hogy van-e hozzá beágyazott harmadrendű módszer. A harmadrendű feltételeknek kell teljesülniük $\hat{b}_j$-kre, a negyedrendű $a_{i,j}$-kkel, és c_j-kkel.

$$\begin{cases} \hat{b}_1 + \hat{b}_2 + \hat{b}_3 + \hat{b}_4 = 1 \\ \hat{b}_2 c_2 + \hat{b}_3 c_3 + \hat{b}_4 c_4 = \frac{1}{2} \quad \rightarrow \quad \frac{1}{2}\hat{b}_2 + \frac{1}{2}\hat{b}_3 + \hat{b}_4 = \frac{1}{2} \\ \hat{b}_2 c_2^2 + \hat{b}_3 c_3^2 + \hat{b}_4 c_4^2 = \frac{1}{3} \quad \rightarrow \quad \frac{1}{4}\hat{b}_2 + \frac{1}{4}\hat{b}_3 + \hat{b}_4 = \frac{1}{3} \\ \hat{b}_3 a_{3,2} c_2 + \hat{b}_4 a_{4,2} c_2 + \hat{b}_4 a_{4,3} c_3 = \frac{1}{6} \quad \rightarrow \quad \frac{1}{4}\hat{b}_3 + \frac{1}{2}\hat{b}_4 = \frac{1}{6} \end{cases}$$

Ha felírjuk ennek az egyenletrendszernek a mátrixát, és megnézzük a determinánsát, akkor azt kapjuk, hogy nem egyenlő 0-val. Így viszont egyértelmű a megoldása. Vagyis nincs 3-adrendű beágyazott módszer ebben az esetben. Ellenben van egy módszer, ami

ilyenkor is megoldást jelenthet. Vegyük hozzá az $a_{i,j}$ mátrixhoz ötödik sornak a b vektort, vagyis $a_{5,i} = b_i$ ahol $i = 1, \ldots, 4$. c_5 pedig értelem szerűen legyen $\sum_{i=1}^{4} a_{5,i}$. Így a $\hat{b}_i$-kre $i = 1, \ldots, 4$ ismeretlenekre felírva a harmadrendű feltételeket, kapunk 4 lineáris egyenletet 5 ismeretlennel. Célszerű feltenni, hogy $\hat{b}_5 \neq 0$, ezért legyen $\hat{b}_5 = 1$. Megoldva a rendszert kapjuk:

$$\hat{b}_1 = \frac{1}{6}, \qquad \hat{b}_2 = -1, \qquad \hat{b}_3 = \frac{5}{3}, \qquad \hat{b}_4 = \frac{5}{6}, \qquad \hat{b}_5 = 1.$$

A módszer neve FSAL(First Same As Last)

A különböző matematikai programokban (például Matlab, Mathematica) gyakran használják azt a beágyazott 5(4) módszert, ami Runge, Kutta és Felhlberg nevéhez fűződik. A módszer együtthatóit Butcher-táblázatban adjuk meg, a felső sarok nulláit elhagyva.

0						
$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{216}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$			
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$		
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	
$b^{(5)}$	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0
$\hat{b}^{(4)}$	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$

Itt $\hat{b}^{(4)}$-vel illetve $b^{(5)}$-vel jelöltük a két együttható halmazt, amely mutatja a hozzátatozó képlet rendjét.

5.3. Lépésválasztás

Azzal, hogy most már láttunk módszereket a lokális hiba becslésre, ezeket ki is használhatnánk, hogy ne fix h lépéssel tegyünk meg egy lépést, hanem a lokális hiba nagyságától függően. Jelölje h_j az x_{j-1} és x_j közötti távolságot minden $j = 1, 2, \ldots, N$-re.

Olyan $\hat{h}$ kellene választani, hogy a globális hiba kisebb maradjon egy adott tolerancia határnál (τ). Vagyis egy p-edrendű módszernél $g_n = \Psi(y_n)h^p$, tehát

$$\tau = \Psi(y_n)\hat{h}^p$$

Osszuk el egymással az egyenleteket!

$$\frac{\tau}{g_n} = \frac{\Psi(y_n)\hat{h}^p}{\Psi(y_n)h^p} = \frac{\hat{h}^p}{h^p}$$

$$\Rightarrow \hat{h} = h\sqrt[p]{\frac{\tau}{g_n}}$$

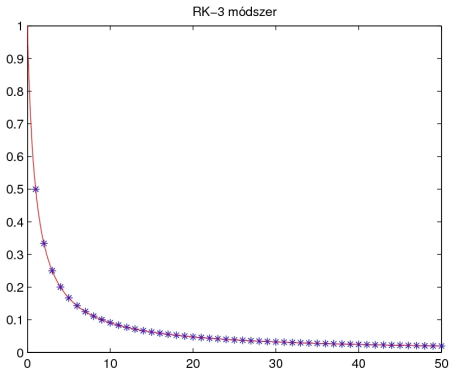
Ez a becsült optimális lépésköz.

5.4. Módszerek a gyakorlatban

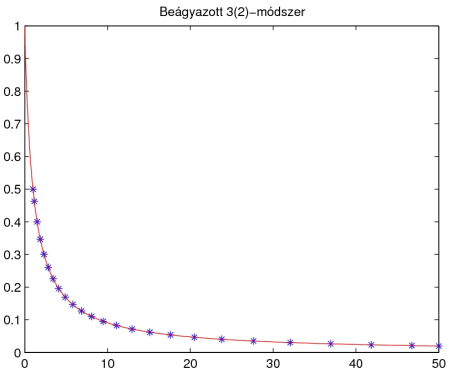
Lássuk, hogy az eddig említett módszerek hogyan viselkednek a gyakorlatban. Példának tekintsük a következő kezdetiérték feladatot:

$$\begin{aligned} y' &= \frac{y^2}{x} - \frac{y}{x} \\ y(1) &= \frac{1}{2} \end{aligned}$$

Ennek analitikusan ki tudjuk számolni a pontos megoldását: $y(x) = \frac{1}{1+x}$. Matlab program segítségével megoldottuk 3-lépcsős Runge–Kutta-módszerrel, és beágyazott 3(2) módszerrel is. Az eredményeket ábrázoltuk a $[0, 50]$ intervallumon.



(a) RK-3 módszer



(b) Beágyazott 3(2) módszer

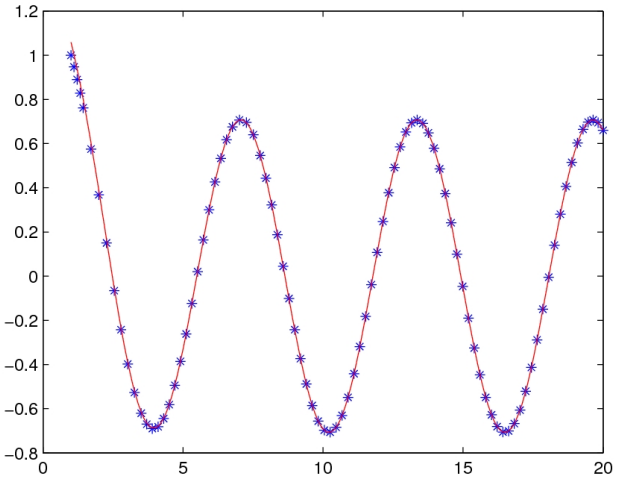
Az ábrákon piros vonal jelzi a pontos megoldást, és a kék pontok a közelítő megoldás pontjait mutatják. Az sima RK-3 módszernél jól látszik, hogy azonosan 1 lépéstávolsággal helyezkednek el a pontok, míg a beágyazott módszernél ez változik a beépített hibabecsléstől függően. Egy módszert akkor mondunk előnyösebbnek a másiknál, ha kevesebb függvénykiértékelésre van szükség (hiszen egy bonyolult függvénynél ez sokáig tarthat), és a megoldástól sem tér el nagyon. Az első ábrán 50 kiértékelésre volt szükség és a 0

közelében, a függvény nagy változásához képest ritkán helyezkednek el a pontok. Ellenben a beágyazott módszernél 25 kiértékelés is elég volt, ráadásul láthatóan jobban közelíti a valódi megoldást.

A következő feladat azt hivatott szemléltetni, hogy még az explicit beágyazott módszerek sem tudnak minden kezdetiérték feladatot ügyesen közelíteni. A feladat:

$$y' = -\lambda(y - \cos(x))y(1) = 1$$

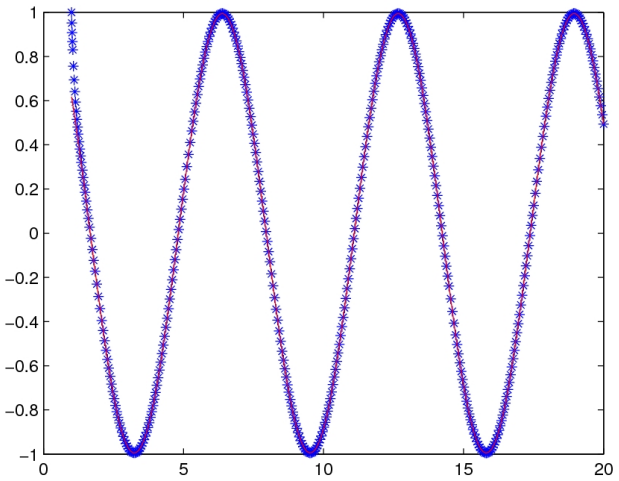
Beágyazott 5(4) módszerrel megoldott feladatot $[0, 20]$ intervallumon ábrázoltuk $\lambda = 1$ paraméter mellett, lásd az 5.1-es ábrán.



5.1. ábra.

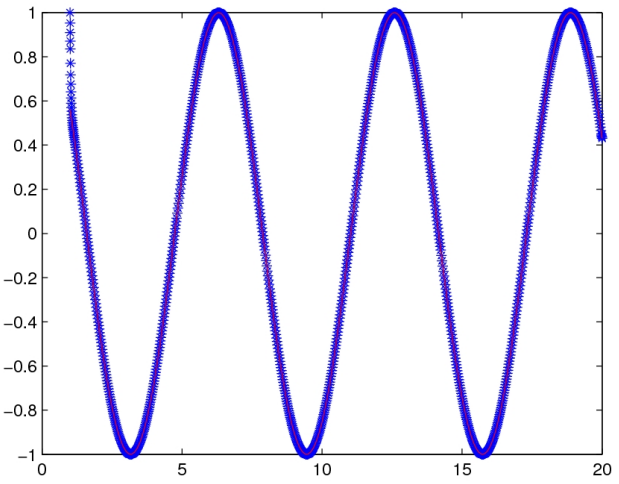
Piros vonal ismét a pontos megoldást jelenti, a kék pontok meg a közelítő megoldást ábrázolják. A megoldó statisztikája: 23 sikeres lépés, 4 hibás kísérlet, és 163 függvénykiértékelés. Újra megoldottuk a feladatot ugyanezzel a módszerrel $\lambda = 10$ paraméter mellett. A változás szemmel látható az 5.2-es ábrán.

Statisztikája: 103 sikeres lépés, 1 hibás kísérlet, 625 függvénykiértékelés. Mivel egy módszer hatékonysága egyenesen arányos a függvénykiértékelések számával, így ez a módszer már nem tűnik annyira jónak. Tehát azt a következtetést tudjuk levonni, hogy a λ paraméter növelésével a feladat úgynevezett merev egyenletté válik. A merevségnek nincs általános definíciója, hatékonysági kérdésen alapszik, hiszen az explicit módszerek is megoldják, csak rendkívül lassan. Az explicit módszereknél a lépéshossz indokolatlanul kicsi lesz, mert a lépéstávolság meghatározásához a stabilitási követelmények jobban hozzájárulnak, mint a pontossági követelmények. Ezért kedvezőbbek az olyan módszerek,



5.2. ábra.

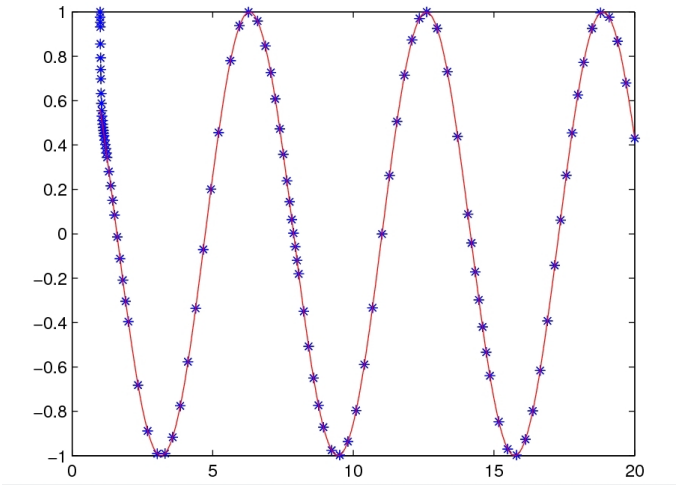
amelyek abszolút stabilak. Ha tovább növeljük λ értékét mondjuk 40-re, akkor az 5.3 ábra adódik.



5.3. ábra.

Ehhez már 277 sikeres lépés, 4 hibás kísérlet, és 1687 függvénykiértékelés kellett. Ugyanezt a feladatot egy merev egyenletekre kifejlesztett programmal oldjuk meg (Matlabban ez az ode15s). A végeredmény az 5.4-es ábrán látható.

Statisztikái: 105 sikeres lépés, 14 hibás kísérlet, 241 függvénykiértékelés. Ez a módszer szemmel láthatóan hatékonyabb. Viszont, ahogy az ábrán is látszik, nem mindegy a kezdeti érték sem, ahonnan a feladat kiindul. Ha közelebbi adatot adtunk volna, akkor még hatékonyabb lett volna a módszer.



5.4. ábra.

6. fejezet

Összefoglalás

Összefoglalásként elmondhatjuk, hogy az előbbiekben ismertetett egylépéses módszerek egyben jópár problémára megoldást jelentenek, viszont merev rendszerek megoldására nem alkalmasak. Megismerkedtünk az explicit és implicit Euler-módszerrel, ennek javításaival, a trapéz- és érintőformulával, valamint az ezeket összefoglaló θ-módszerrel. Ismertettük a Runge–Kutta-módszercsaládot, betekintést nyertünk az együtthatók általános kiszámításába, amihez gráfelméleti eszközöket is alkalmaztunk. Láthattunk folytonos Runge–Kutta képletet, valamint megtudhattuk, hogy használhatók a beágyazott módszerek a lokális hiba becslésében. Végül példákon keresztül illusztráltuk a különböző módszerek ugyanazon feladat megoldásával összefüggő előnyeit és hátrányait.

Irodalomjegyzék

[1] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations.* Wiley, 2003.

[2] A. Iserles. *A First Course in the Numerical Analysis of Differencial Equations.* Cambridge University Press, 1996.

[3] E. Hairer és G. Wanner. *Solving Ordinary Differencial Equations 2.* Springer, 1980.

[4] J. Stoer és R. Bulirsch. *Introduction to Numerical Analysis.* Springer-Verlag, 1993.

[5] E. Hairer és S. P. Nørsett and G. Wanner. *Solving Ordinary Differential Equations 1.* Springer, 1993.

[6] Stoyan G. és Takó G. *Numerikus módszerek 2.* Elte-Typotex, Budapest, 1995.

[7] P. G. Thomsen. *Numerical ODEs -analysis and applications, manuscript*, 2007.