16. Nemlineáris egyenletek megoldása I.

Eddig lényegében lineáris egyenletrendszerek megoldásával foglalkoztunk. De sokszor felvetődik az

$$f(x)=0 \tag{16.1}$$

egyenlet egy (vagy esetleg több) gyökének keresése, ahol az $f(x)\in C[a,b]$ egyváltozós függvény. Minden olyan x^* érték, amelyre $f(x^*)=0$, a (16.1) egyenlet *gyöke* vagy $f(x)$ *zérushelye*. A gyök az x^* helyen m*-edrendű*, ha $f(x)=(x-x^*)^m g(x),\ g(x^*)\neq 0$ alakban írható. Azzal az esettel foglalkozunk, amikor a megoldás közelítése valamilyen numerikus módszer segítségével végezhető.

16.1. A gyököt tartalmazó intervallum

Ha a függvény előjelet vált: $f(a)f(b)<0$, akkor a folytonosság miatt legalább 1 gyök található $[a,b]$-ben. Ha létezik $f(x)$ első deriváltja is, és előjeltartó $[a,b]$-ben, akkor csak 1 gyök van.

Ha a függvény nem monoton, akkor $[a,b]$-t célszerű olyan részintervallumokra bontani, ahol az intervallum két végpontja között előjelváltás van. Ílymódon $f(x)$ páratlan gyökeit el tudjuk különíteni. Deriválható függvény esetén a páros gyököket kereshetjük $f'(x)$ gyökeiként, mert ekkor a párosakat páratlanná tettük, de kereshetjük $f(x)/f'(x)$ gyökeit is, amelyek mind egyszeresek.

16.2. Fixpont iteráció

Egy lehetséges eljárás, hogy megpróbáljuk az $f(x)=0$ egyenletet fixpont-egyenletté alakítani:

$$x=F(x). \tag{16.2}$$

Példa: legyen a megoldandó egyenlet: $x^2-\sin(x)=0$. Ekkor próbálkozhatunk az $x_{k+1}=\sqrt{\sin(x_k)}$ iterációval. Fixpont-egyenletet mindig tudunk készíteni, hiszen $x=x+cf(x)$ is ilyen, ahol c nemzérus állandó, de választhatunk valamely $c(x)$ függvényt is olymódon, hogy az iteráció konvergencia-tulajdonságai javuljanak. A fixpont létezéséről szól a

16.2.1 Brouwer fixponttétel[1]

Ha $F(x)$ folytonos $[a,b]$-ben és $F:\ [a,b]\to[a,b]$, akkor létezik fixpontja.

Bizonyítás. Legyen $g(x)=x-F(x)$, ekkor $g(a)\le 0$ és $g(b)\ge 0$, amiből $g(a)g(b)\le 0$. Ha itt egyenlőségjel érvényes, akkor már van egy gyök, ha pedig a < jel érvényes, akkor a folytonosság miatt kell léteznie gyöknek $[a,b]$-ben. ■

16.2.2 Tétel, kontrakció

Ha $F:S\to\mathbb{R}$ folytonosan differenciálható az S zárt intervallumon és $|F'(x)|<1,\ \forall x\in S$, akkor F kontrakció.

Bizonyítás. A Lagrange középérték tétel alapján $x,y\in S$-re $\exists\zeta$: $F(x)-F(y)=F'(\zeta)(x-y)$. Térjünk át az abszolút értékre és használjuk fel, hogy $\exists|F'(x)|$ maximuma S-ben:

[1] A tétel többdimenziós megfogalmazása: ha a folytonos $F(x)$ függvény a gömböt önmagába képezi le, akkor van fixpontja.

$$|F(x)-F(y)| \le \max_{x\in S}|F'(x)||x-y|, \quad x,y \in S$$

tehát F kontrakció $q=\max_{x\in S}|F'(x)|<1$ kontrakciós állandóval. ■

16.2.3 Következmény

Ha $F(x)$ kontrakció, akkor a Banach fixponttétel szerint csak egy gyök van és a kontrakciós állandó ismeretében a közelítés pontosságát is becsülni tudjuk.

Visszatérve a fenti példához: a kapott iteráció biztosan konvergens abban a tartományban, ahol $\left(\sqrt{\sin(x)}\right)' = \dfrac{\cos(x)}{2\sqrt{\sin(x)}} < 1$. Látjuk, ha $x=0$ vagy $x=\pi$, ezzel a kifejezéssel baj van, mert a formula kiértékelésekor 0-val kéne osztani. De például $x=\pi/4$ esetén a kifejezés értéke $1/\sqrt{8}$, ami már jobbnak tűnik. Ha megrajzoljuk az x^2 parabolát és a $\sin(x)$ függvény képét, látjuk, hogy két nemnegatív gyök van, az egyik a zérus, a másik pedig közel $x=\pi/4$-hez, tehát remélhető, hogy az iteráció $\pi/4$-gyel indítva konvergens. De az is látszik, hogyha nagyon kicsi pozitív értékkel indítjuk az iterációt, akkor sem kapjuk meg a zérus gyököt, mert az iteráció mindig elvisz a nagyobbik gyök irányába.

Ha azonban az $x=\arcsin(x^2)$ iterációt készítjük, könnyen meggyőződhetünk arról, hogy kis pozitív x-re zérushoz tart. Ha azonban $x=1$-gyel indítunk, akkor először a $\pi/2$ értéket kapjuk, majd komplex számokat, mivel az argumentum nagyobb 1-nél.

A tanulság: ügyelnünk kell, a kapott függvény hova képez le, és a leképezés tartományában megmaradnak-e a konvergencia tulajdonságok, illetve azt a gyököt kapjuk-e, amit szeretnénk meghatározni.

Ha a megoldandó egyenletben több helyen is szerepel x, akkor több $x=F(x)$ kifejezés is készíthető. Például szerepeljen két helyen, ekkor meg lehet mutatni: a kapott két iteráció egy adott helyen egyszerre nem lehet konvergens. Legyen ugyanis $f(x_1,x_2)=0$ a megoldandó egyenlet, ahol a két előfordulást x_1 és x_2-vel azonosítjuk. Legyen $F_i(x)$ az az iterációs függvény, amelyet x_i kifejezésével kaptunk. Ez azt jelenti, hogy

$$f(F_1(x),x)=0 \quad \text{és} \quad f(x,F_2(x))=0. \tag{16.3}$$

Legyen α egyszeres gyök: $f(\alpha,\alpha)=0$. Ha a (16.3)-ben szereplő kifejezéseket deriváljuk x szerint és helyettesítjük $x=\alpha$ -t, az eredmény:

$$f_1'(\alpha,\alpha)F_1'(\alpha)+f_2'(\alpha,a)=0,$$
$$f_1'(\alpha,\alpha)+f_2'(\alpha,\alpha)F_2'(\alpha)=0,$$

ahol f alsó indexe azt jelöli, melyik hely szerint deriváltunk. Ahhoz, hogy egyszeres gyök mellett nemzérus megoldást kapjunk $f_i'(\alpha,\alpha)$-ra kell, hogy a kapott rendszer determinánsa 0 legyen:

$$\begin{vmatrix} F_1'(\alpha) & 1 \\ 1 & F_2'(\alpha) \end{vmatrix} = F_1'(\alpha)F_2'(\alpha)-1=0,$$

ahonnan

$$|F_1'(\alpha)| = 1/|F_2'(\alpha)|. \tag{16.4}$$

Hacsak nem 1 abszolút értékűek a deriváltak, a gyök közelében az egyik iterációs függvény konvergens, a másik meg divergens lesz. Hasonlóan lehet vizsgálni azt az esetet, amikor x 2-nél többször fordul elő. De ekkor a helyzet rosszabb, az is lehetséges, hogy egyik iterációs függvény sem konvergens. Emiatt azt célszerű tennünk, hogy az x-ek előfordulását két csoportba osztjuk és x-et az egyik csoportból teljesen kifejezük. Például $3x^2 - 2x + \exp(2.2x) + 1 = 0$ -nél az iterációs függvényre kereshetjük a másodfokú polinom gyökeit úgy, hogy a konstans tag helyére $\exp(2.2x) + 1$-et írunk.

16.3. A konvergencia-sebesség

Legyen az x_n sorozat konvergens, $\lim_{n\to\infty} x_n = x^*$. Jelölje $\varepsilon = x_n - x^*$ az n-edik hibát. Ekkor, ha létezik c állandó és $p \ge 1$ szám úgy, hogy

$$\left|\varepsilon_{n+1}\right| \le c\left|\varepsilon_n\right|^p, \quad n = 0,1,\ldots, \tag{16.5}$$

akkor az x_n sorozat konvergenciája p-*edrendű*. Ha

- $p = 1$, akkor a konvergencia *lineáris* vagy *elsőrendű*,
- $1 < p < 2$, akkor a konvergencia *szuperlineáris*,
- $p = 2$, akkor *kvadratikus* vagy *másodrendű*,
- $p = 3$, akkor *köbös*, vagy *harmadrendű*.

A p szám jellemzi az iterációs módszer konvergenciájának sebességét. Ha például $p = 2$, akkor ez nagyjából azt jelenti, hogy lépésenként az értékes jegyek száma megduplázódik.

A fixpont iteráció nem rendelkezik ezzel a sebességgel. Megmutatjuk, hogy $p = 1$, azaz a konvergenciája elsőrendű, amennyiben $\left|F'(x^*)\right| \ne 0$. Ugyanis

$$\left|\varepsilon_{n+1}\right| = \left|x_{n+1} - x^*\right| = \left|F(x_n) - F(x^*)\right| \le q\left|x_n - x^*\right| = q\left|\varepsilon_n\right|. \tag{16.6}$$

Ha $F'(x^*) = 0$, akkor a konvergencia magasabb rendű. Erre vonatkozik a következő

16.3.1 Tétel

Legyen F valós függvény: $F(S) \subset S \subset \mathbb{R}$, S zárt. Tegyük fel, $F \in C^m(S)$ és $F^{(k)}(x^*) = 0$, $k = 1,2,\cdots,m-1$. Ekkor az F által meghatározott iteráció konvergencia-sebessége $p = m$-edrendű.

Bizonyítás. Az x^* körüli Taylor-polinom m-edrendű maradéktaggal

$$F(x) = F(x^*) + F'(x^*)(x - x^*) + \ldots + \frac{F^{(m-1)}(x^*)}{(m-1)!}(x - x^*)^{m-1} + \frac{F^{(m)}(\xi_x)}{m!}(x - x^*)^m$$

ahol a feltevés szerint az első, második, … , $m-1$-edik derivált eltűnik. Helyettesítsük $x = x_n$-et, vegyük figyelembe, hogy $x^* = F(x^*)$ és $x_{n+1} = F(x_n)$, ezzel

$$x_{n+1} - x^* = \frac{F^{(m)}(\xi_x)}{m!}(x_n - x^*)^m,$$

ahonnan

$$\left|\varepsilon_{n+1}\right| = \frac{\left|F^{(m)}(\xi_x)\right|}{m!}\left|x_n - x^*\right|^m \le \frac{M_m}{m!}\left|\varepsilon_n\right|^m, \quad n = 0,1,\ldots$$

ahol $M_k = \max\limits_{x\in S}\left|F^{(k)}(x)\right|$. Innen látható, a konvergencia m-edrendű. ■

16.4. Newton-iteráció (Newton-Raphson módszer) és a szelőmódszer

Ha a függvény első deriváltja létezik a gyök környezetében, akkor a gyököt közelíthetjük úgy, hogy az x_n pontban a függvényhez húzott érintő metszéspontját vesszük az x tengellyel. Ez ugyanaz, mint amikor az x_n körüli elsőfokú Taylor-polinomot zérussá tesszük és x_{n+1}-re megoldjuk:

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n),$$

innen a Newton-Raphson módszer iterációs formulája:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = F(x_n). \tag{16.7}$$

A *szelőmódszert* ebből úgy nyerjük, hogy a derivált helyére az utolsó két pontra felírt osztott differenciát tesszük:

$$x_{n+1} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})} = \frac{f(x_n)x_{n-1} - f(x_{n-1})x_n}{f(x_n) - f(x_{n-1})} = F(x_{n-1}, x_n), \tag{16.8}$$

tehát ez az iterációs függvény két pontra támaszkodik. A módszer előnye a Newton-módszerrel szemben, hogy nem kell hozzá a derivált, amit néha eléggé körülményes kiszámítani. Hátránya pedig a kisebb konvergencia-sebesség.

16.4.1 Tétel, a szelőmódszer hibája

Legyen $f(x) \in C^2[x_{n-1}, x_n, x^*]$, ekkor a szelőmódszernél az $n+1$-edik iterált hibájára fennáll

$$\varepsilon_{n+1} = \varepsilon_n \varepsilon_{n-1} \frac{f''(\xi)}{2f'(\eta)}, \quad \xi, \eta \in [x^*, x_{n-1}, x_n], \tag{16.9}$$

ahol x^* a zérushely és $[x^*, x_{n-1}, x_n]$ az adott pontok által lefedett intervallum.

<u>Bizonyítás.</u> Az állítást (16.8)-ból osztott differenciák segítségével származtatjuk. Kihasználjuk, hogy $f(x^*) = 0$:

$$\varepsilon_{n+1} = x_{n+1} - x^* = x_n - x^* - (x_n - x^*) \frac{f(x_n) - f(x^*)}{x_n - x^*} \frac{1}{f[x_{n-1}, x_n]} = \varepsilon_n \left(1 - \frac{f[x^*, x_n]}{f[x_{n-1}, x_n]} \right) =$$

$$= \varepsilon_n \left(\frac{f[x_{n-1}, x_n] - f[x^*, x_n]}{f[x_{n-1}, x_n]} \frac{\varepsilon_{n-1}}{x_{n-1} - x^*} \right) = \varepsilon_n \varepsilon_{n-1} \frac{f[x^*, x_{n-1}, x_n]}{f[x_{n-1}, x_n]}$$

és innen az osztott differenciák és a deriváltak között érvényes összefüggés (14.1.1 Következmény) segítségével kapjuk az eredményt. ■

16.4.2 Következmény

A Newton-módszerre vonatkozó eredményt az $x_{n-1} \to x_n$ határátmenettel kapjuk:

$$\varepsilon_{n+1} = \varepsilon_n^2 \frac{f''(\xi)}{2f'(x_n)}, \quad \xi \in [x_n, x^*], \tag{16.10}$$

Látjuk, ha van konvergencia, akkor az másodrendű, feltéve, hogy $f'(x^*) \neq 0$.

16.4.3 Tétel, monoton konvergencia

Legyen $f \in C^2[a,b]$, $f(x^*) = 0$, $x^* \in [a,b]$, *az* $f'(x), f''(x)$ *deriváltak ne váltsanak előjelet* $[a,b]$-*ben, továbbá az* $x_0 \in [a,b]$ *kezdőpontra teljesüljön* $f(x_0)f''(x_0) > 0$. *Ekkor a Newton-módszer konvergens és az általa készített* x_n *sorozat monoton módon tart az* x^* *zérushelyhez.*

Bizonyítás. A Newton-módszer (16.10) formulája szerint az összes iterált a gyöktől vagy jobbra, vagy balra helyezkedik el, mert f''/f' előjele állandó. A (16.7) formulából x^*-ot levonva

$$\varepsilon_{n+1} = \varepsilon_n - f(x_n)/f'(x_n) \tag{16.11}$$

következik. Az $f(x_0)f''(x_0) > 0$ feltétel miatt $f''(x_0)/f'(x_0)$ és $f(x_0)/f'(x_0)$ előjele megegyezik. Emiatt, ha (16.10)-ben $\varepsilon_1 > 0$, akkor $f''(x_0)/f'(x_0)$ pozitív és (16.11)-ben ε_0 kissebítve van és az összes további lépésben $\varepsilon_n > 0$ kisebbedik. Hasonlóan kapjuk, hogy $\varepsilon_1 < 0$ esetén az összes további $\varepsilon_n < 0$ nagyobbodik, tehát az ε_n-ek vagy felülről vagy alulról monoton módon tartanak 0-hoz. ■

Következmény. A (16.9) formula mutatja, hogyha a szelőmódszert úgy indítjuk, hogy $x_0, x_1 \in [a,b]$, ε_0 ε_1 és $f''(x_0)/f'(x_0)$ előjele megegyezik, akkor a tétel feltételei mellett a szelőmódszer is monoton konvergens sorozatot állít elő, mert a formulájában szereplő osztott differencia mindig helyettesíthető egy intervallum-beli deriválttal, aminek az előjele $[a,b]$-ben állandó.

16.4.4 Tétel, lokális konvergencia

Legyen $f \in C^2[a,b]$, $f(x^*) = 0$, $f'(x) \neq 0$, $x, x^* \in [a,b]$, *és az* $x_0 \in [a,b]$ *kezdőpontra teljesüljön*

$$\left|x_0 - x^*\right| < \frac{2\min_{[a,b]}|f'(x)|}{\max_{[a,b]}|f''(x)|} = \frac{1}{M}. \tag{16.12}$$

Ilyen x_0*-ból indítva a Newton-Raphson módszer konvergál* x^**-hoz. A szelőmódszer konvergál, ha* x_0 *mellett* x_1 *is kielégíti a* (16.12) *feltételt.*

Bizonyítás. Az első lépéstől kezdve van kontrakció, ha (16.9) vagy (16.10) alapján

$$\left|\varepsilon_0 \frac{f''(\xi)}{2f'(\eta)}\right| \le \left|x_0 - x^*\right| \frac{\max_{[a,b]}|f''(x)|}{2\min_{[a,b]}|f'(x)|} < 1.$$

Az állítás innen átrendezéssel adódik. A szelőmódszernél a második lépéshez még ε_1-re is meg kell követelnünk ugyanezt a feltételt. ■

A fentiek alapján a Newton-Raphson módszernél megbecsüljük az $n+1$-edik hibát. Bevezetve a $d_k = M|\varepsilon_k|$ jelölést

$$d_{n+1} = M|\varepsilon_{n+1}| \le M^2\varepsilon_n^2 \quad \to \quad d_{n+1} \le d_0^{2^n} \quad \to \quad |\varepsilon_{n+1}| \le (M\varepsilon_0)^{2^n}. \tag{16.13}$$

16.4.5 Tétel, szelőmódszer konvergencia-sebessége

A 16.4.4 Tétel feltételei mellett az x_0, x_1 *kezdőpontokból indítva a szelőmódszer* $p = (1+\sqrt{5})/2 \approx 1{,}62$ *aszimptotikus sebességgel konvergál* x^**-hoz.*

Bizonyítás. Most

$$|\varepsilon_{n+1}| \le M|\varepsilon_n||\varepsilon_{n-1}|$$

érvényes (16.9) alapján, ahol M ugyanaz, mint (16.12)-ben. Ismét a $d_k = M|\varepsilon_k|$ jelöléssel

$$d_{n+1} \le d_n d_{n-1}, \quad n = 1,2,\ldots$$

Az indításkor $\left|x_0 - x^*\right| < 1/M$ és $\left|x_1 - x^*\right| < 1/M$, ezzel $d_0, d_1 < 1$. Igaz tehát, hogy $\exists d < 1$: $d_0, d_1 \le d$, amellyel $d_2 \le d^2$, $d_3 \le d^3$, $d_4 \le d^5$, általában

$$d_n \le d^{f_n}, \quad f_0 = f_1 = 1, \quad f_{n+1} = f_{n-1} + f_n, \quad n = 1,2,\ldots$$

Itt f_n-ek a jólismert Fibonacci-sorozat tagjai, melyeknek explicit előállítása ismert:

$$f_n = \frac{1}{\sqrt{5}}\left[b_1^{n+1} - b_2^{n+1}\right], \quad b_1 = \frac{1+\sqrt{5}}{2}, \quad b_2 = \frac{1-\sqrt{5}}{2}. \tag{16.14}$$

Mivel $\left|b_2\right| < 1$, a növekvő hatványai zérushoz fognak tartani. Emiatt létezik egy K szám, hogy minden n-re $d^{s_{n+1}} \le K$, $s_{n+1} = -b_2^{n+1}/\sqrt{5}$. Tehát írható

$$d_n \le K\left(d^{b_1/\sqrt{5}}\right)^{b_1^n} = K\left(\tilde{d}\right)^{b_1^n}, \quad \tilde{d} = d^{b_1/\sqrt{5}}.$$

Kaptuk, hogy a szelőmódszerhez tartozó hibák majorálhatók egy olyan sorozattal, amelynek konvergenciarendje $b_1 = \dfrac{1+\sqrt{5}}{2} \approx 1{,}62$, azaz a módszer *szuperlineáris*. ■

16.5. Példák

1. Legyen $f \in C^3[a,b]$. Parabola interpolációval készítsünk három pontra támaszkodó iterációs módszert $f(x)$ egy $[a,b]$-beli lokális minimumának meghatározására!

Megoldás. Legyen $[a,b]$-ben három pont $(x_{i-2}, f_{i-2}), (x_{i-1}, f_{i-1}), (x_i, f_i)$. Newton-interpolációval $p_2(x) = f_{i-2} + f[x_{i-2}, x_{i-1}](x - x_{i-2}) + f[x_{i-2}, x_{i-1}, x_i](x - x_{i-2})(x - x_{i-1})$. A deriváltjának zérushelye:

$$x_{i+1} = \frac{x_{i-2} + x_{i-1}}{2} - \frac{f[x_{i-2}, x_{i-1}]}{2f[x_{i-2}, x_{i-1}, x_i]}. \tag{16.15}$$

2. Egyenletes lépésközzel haladva hogyan derítenénk fel egy minimumhelyet?

Megoldás. Legyen a lépésköz h és $x_j = a + jh$, $j = 0,1,\ldots$. Az x_{j-1}, x_j, x_{j+1} alappont-hármas megfelelő, ha $f[x_{j-1}, x_j] < 0$ és $f[x_j, x_{j+1}] > 0$. Ekkor a lokális minimumot a következő egyszerűsített formulával becsülhetjük, ha (16.15)-ben x_j-t vesszük a középső pontnak:

$$x_{\min} \approx \frac{x_{j-1} + x_{j+1}}{2} - \frac{f[x_{j-1}, x_{j+1}]}{2f[x_{j-1}, x_j, x_{j+1}]} = x_j - \frac{h}{2}\frac{f_{j+1} - f_{j-1}}{f_{j+1} - 2f_j + f_{j-1}}. \tag{16.16}$$

3. A kapott iterációs módszerre fogalmazzunk meg ahhoz hasonló tételt, mint amit a Newton-módszernél láttunk a lokális konvergenciára!

Megoldás. Az egyszerűség kedvéért tekintsük (16.15)-ben azt az esetet, amikor $i = 2$. Az osztott differenciák tulajdonságait kihasználva a hibák terjedésére próbálunk egy összefüggést származtatni. Vonjuk le mindkét oldalból a minimumhelyet adó x^*-ot és legyen $\varepsilon_i = x_i - x^*$, ezzel

$$\varepsilon_3 = \frac{\varepsilon_0 + \varepsilon_1}{2} - \frac{f[x_0, x_1]}{2f[x_0, x_1, x_2]}.$$

A számlálóban lévő osztott differenciát átalakítjuk, kihasználva, hogy $f[x^*,x^*]=0$ és az alappontok sorrendje az osztott differenciákban tetszőleges: $f[x_0,x_1]=f[x_0,x_1]-f[x_1,x^*]+f[x_1,x^*]-f[x^*,x^*]=$ $=\varepsilon_0 f[x_0,x_1,x^*]+\varepsilon_1 f[x_1,x^*,x^*]$. Beírva a fenti formulába és közös nevezőre hozva:

$$\varepsilon_3=\frac{\varepsilon_0\left(f[x_0,x_1,x_2]-f[x_0,x_1,x^*]\right)+\varepsilon_1\left(f[x_0,x_1,x_2]-f[x_1,x^*,x^*]\right)}{2f[x_0,x_1,x_2]}.$$

A számláló első két tagja továbbírva $\varepsilon_0\varepsilon_2 f[x_0,x_1,x_2,x^*]$. A utolsó két tag átalakítása kicsit hosszabb: $\varepsilon_1\left(f[x_0,x_1,x_2]-f[x_0,x_1,x^*]+f[x_0,x_1,x^*]-f[x_1,x^*,x^*]\right)=\varepsilon_1\varepsilon_2 f[x_0,x_1,x_2,x^*]+\varepsilon_0\varepsilon_1 f[x_0,x_1,x^*,x^*]$.
Ezekkel

$$\varepsilon_3=\frac{(\varepsilon_0+\varepsilon_1)\varepsilon_2 f[x_0,x_1,x_2,x^*]}{2f[x_0,x_1,x_2]}+\frac{\varepsilon_0\varepsilon_1 f[x_0,x_1,x^*,x^*]}{2f[x_0,x_1,x_2]}.$$

Legyen $\delta_2=\max\{|\varepsilon_0|,|\varepsilon_1|,|\varepsilon_2|\}$ és

$$M=\frac{\max\limits_{x\in[a,b]}\left|f^{(3)}(x)\right|}{2\min\limits_{x\in[a,b]}\left|f''(x)\right|}. \tag{16.17}$$

Felhasználva, hogy az osztott differenciák kifejezhetők a nekik megfelelő rendű deriváltakkal, kapjuk:

$$|\varepsilon_3|\le\frac{3}{2}\delta_2^2\frac{2!}{3!}2M=\delta_2^2 M. \tag{16.18}$$

Így $|\varepsilon_3|$ biztosan kisebb a három megelőző ε abszolút maximumánál, ha $\delta_2 M<1$, vagy másképp $\delta_2<1/M$. Tehát a kapott módszer biztosan konvergens, ha a három induló pont a minimumhely $1/M$ -sugarú környezetében van.

16.6. Gyakorlatok

1. Bizonyítsuk be, hogyha $f\in C^1[a,b]$, $f(a)f(b)<0$ és $f'(x)$ nem vált előjelet $[a,b]$-ben, akkor ott az $f(x)$ függvénynek csak egy gyöke van.
2. A fixponttétel alkalmazásával mutassuk meg, hogy a $\cos x-4x+2=0,\ x\in\mathbb{R}$ egyenletnek egy zérushelye van és x-et $4x$ felől kifejezve a fixpont iteráció minden kezdőértékre konvergens!
3. Az előző feladatban a gyök milyen környezetéből konvergál biztosan a Newton-iteráció?
4. Oldjuk meg az $f(x)=1/x-a=0$ egyenletet Newton-iterációval! Milyen kezdőértékekre van konvergencia? A kapott formula érdekesége, hogy nincs benne osztás, aminek régebben külön jelentősége volt az osztás műveletével nem rendelkező gépi aritmetikákban.
5. Oldjuk meg az $f(x)=x^2-a=0,\ a>0$ egyenletet Newton-iterációval és tisztázzuk a konvergenciát!
6. Az előző feladat megoldása alapján készítsünk módszert $a^{1/k}$ meghatározására, ahol a pozitív valós szám.
7. Mutasssuk meg, hogy a 16.4.3 Tételt módosíthatjuk úgy, hogy az $f(x_0)f''(x_0)>0$ feltételt elhagyjuk és helyette azt követeljük meg, hogy az első lépés után $x_1\in[a,b]$.
8. Mutassuk meg, hogy az $F(x_n)=x_n-\dfrac{\left(f(x_n)\right)^2}{f(x_n)-f(x_n-f(x_n))}$ iteráció konvergenciája másodrendű!

9. Ellenőrízzük, hogy a Newton-módszer többszörös gyök esetén csak elsőrendben konvergál.
10. Igazoljuk, hogy r-szeres multiplicitású gyöknél a kvadratikus konvergencia megmarad, ha a Newton-módszer formuláját a következőre módosítjuk: $x_{n+1} = x_n - rf(x_n)/f'(x_n)$.
11. Adott ε pontosság elérése érdekében dolgozzuk ki annak feltételét, hogy mikor állítsuk le a Newton-módszert.
12. Mi történik a szelőmódszernél, ha a 16.4.3 Tétel feltételeitől csak annyi a különbség, hogy $\varepsilon_0 > 0$, de $\varepsilon_1 < 0$?
13. Mikor állítsuk le a szelőmódszert, hogy a megoldás előírt pontosságú legyen?

17. Nemlineáris egyenletek megoldása II.

Néhány speciális esettel folytatjuk.

17.1. Az intervallumfelezés módszere

Tegyük fel, az $[a,b]$ intervallum tartalmaz 1 db gyököt: $f(a)f(b) < 0$ és a függvény folytonos $[a,b]$-ben. Az intervallumfelezés módszere szerint ekkor megfelezzük az intervallumot és a két intervallum közül megtartjuk azt, ahol az előjelváltás megmarad. Így az algoritmus:

1. $\exists f \in C[a,b],\ f(a)f(b) < 0$ és adott $\in$ előírt pontosság.
2. Indulás: $[a_0, b_0] = [a,b],\ x_1 = (a+b)/2$.
3. $$[a_n, b_n] = \begin{cases} [a_{n-1}, x_n], & \text{ha } f(a_{n-1})f(x_n) < 0, \\ [x_n, b_{n-1}], & \text{egyébként,} \end{cases}$$
   $x_{n+1} = (a_n + b_n)/2.$
4. Megállás: ha $f(x_n) = 0$, vagy $|b_n - a_n| < \in$.

Ez nem túl gyors, de biztos módszer. Az előjelváltásból nem mindig következik a gyök léte. Gondoljunk az $1/x$ függvényre, amikor az algoritmust -1 és 2 között indítjuk.

17.1.1 Tétel

Az intervallumfelezéssel kapott x_n, $n = 1,2,\ldots$ sorozat elsőrendben konvergens és

$$|\varepsilon_n| \le \frac{b-a}{2^n}, \quad n = 0,1,\ldots \tag{17.1}$$

<u>Bizonyítás.</u> A konvergencia abból következik, hogy mindig a gyököt tartalmazó intervallumot tartjuk meg. A hibára minden lépésben teljesül:

$$|\varepsilon_{n+1}| \le \frac{1}{2}|\varepsilon_n|,$$

ez pedig elsőrendű konvergenciát jelent. ■

17.2. A húrmódszer (regula falsi)

Itt csak annyi az eltérés az intervallumfelezés módszerétől, hogy nem az intervallum közepét vesszük, hanem az $(a_n, f(a_n))$ és $(b_n, f(b_n))$ pontokra illesztett egyenes, más névvel: *húr* zérushelye a következő közlítés:

$$x_{n+1} = a_n - f(a_n)\frac{b_n - a_n}{f(b_n) - f(a_n)}. \tag{17.2}$$

Megfelelő feltételek mellett bizonyítható, hogy a húrmódszer konvergencája lineáris, így nem gyorsabb, mint az intervallumfelezés. Még az is megeshet, hogy annál lassúbb. Ez történik például olyan esetben, amikor a függvény értékei az x-tengelyhez közel vannak és az egyik végponthoz (a_n vagy b_n) nagyon közeli a gyök.

17.3. A Newton-iteráció többváltozós esetben

Legyen $f: \mathbb{R}^n \to \mathbb{R}^n$ egy n-változós leképezés, amelynek keressük azt a vektorát, amelyre $f(x) = 0$. Tételezzük fel a differenciálhatóságot, ekkor az $x_k \in \mathbb{R}^n$ körüli sorfejtésből közelítve

$$f(x_k) + f'(x_k)(x - x_k) = 0, \tag{17.3}$$

ahol most $f'(x) = \left[\partial f_i(x)/\partial x_j\right] \in \mathbb{R}^{n\times n}$ mátrix – a rendszer un. Jacobi-mátrixa - , amelyről feltesszük, hogy invertálható. (17.3)-et x-re megoldva a következő iterációt kapjuk:

$$x_{k+1} = x_k - \left[f'(x_k)\right]^{-1} f(x_k). \tag{17.4}$$

Ha van megoldás és elég közel vagyunk hozzá, akkor remélhetjük, hogy a többváltozós Newton-iteráció konvergens lesz.

A módszer azt követeli meg, hogy minden lépésben elkészítsük a deriváltak mátrixát és megoldjunk vele egy lineáris egyenletrendszert. Mivel ez nagyon munkaigényes lehet, szokás alkalmazni a következő egyszerűsítést: Elkészítjük az $f'(x_k) = LU$ faktorizációt és utána az egyszerűbb

$$x_{k+1} = x_k - (LU)^{-1} f(x_k) \tag{17.5}$$

iterációt alkalmazzuk. Ez 1-dimenzióban annak felel meg, hogy lépésenként a derivált értékét nem változtatjuk. Az ilyen módszereket kvázi-Newton módszereknek nevezzük.

17.4. Polinomok gyökei

A polinomok gyökeinek meghatározása talán leggyakrabban a mátrixok sajátértékeinek keresésekor jön elő, de ekkor nem érdemes a hatványösszeg alakot használni, mert a lineáris algebrai algoritmusok numerikusan előnyösebb megoldásokat kínálnak. Valóban magasabbfokú polinomok esetén a hatványösszeg reprezentáció

$$p(x) = \sum_{i=0}^{n} a_i x^i \tag{17.6}$$

nem előnyös, mert gépi számként ábrázolva az együtthatók n növekedésével egyre bizonytalanabb információt nyújtanak a gyökök pontos értékéről. Példának álljon itt Wilkinson kisérlete, aki az 1,2,...,19,20 gyökökkel rendelkező huszadfokú polinomot (17.6) alakban előállította, majd visszaszámolta a gyököket. Az eredmény annyira más volt, hogy több komplex gyökpárt is kapott. A jelenséget magyarázza a gyökök és együtthatók összefüggése: például a nulladfokú tag a gyökök szorzata: 20!, aminek a pontos ábrázolására messzi nem elegendő 15 decimális jegy. Így a gépi számábrázolás folytán sok fontos információ elvész.

A (17.6) alak összefüggésbe hozható az un. *Frobenius-féle kisérő mátrix*-szal, amellyel már találkoztunk a 7.3 szakaszban:

$$F=\begin{bmatrix} 0 & 0 & \dots & \dots & 0 & -a_0/a_n \\ 1 & 0 & \dots & \dots & 0 & -a_1/a_n \\ & 1 & \ddots & & \vdots & \vdots \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & \ddots & 0 & -a_{n-2}/a_n \\ & & & & 1 & -a_{n-1}/a_n \end{bmatrix}. \tag{17.7}$$

Az utolsó oszlopa mentén kifejtve könnyen igazolható, hogy $\det(\lambda I - F) = \frac{1}{a_n}\sum_{i=0}^{n} a_i \lambda^i$. Ennek a mátrixnak ismerete két szempontból is hasznos. Egyrészt mutatja, hogy a polinom x_k gyökei lineáris algebrai módszerekkel kereshetők, amelyek a legstabilabbnak tekinthető módszerek közé tartoznak. Másrészt rögtön lehetőségünk van egy olyan körlemez megadására a komplex síkon, amelyben a polinom összes gyöke benne van:

$$\|F\|_\infty = \max_{0\le i<n}\left(1-\delta_{i0} + |a_i/a_n|\right) = R \ge |x_k|,$$

ahol δ_{ij} a Kronecker delta. R nagyobb vagy egyenlő F spektrál sugaránál, ami most a polinom gyökök abszolút értékeinek maximuma.

Megadhatunk egy másik kisebb körlemezt is, amelyen kívül van az összes gyök. Vezessük be az $x = 1/y$ transzformációt és írjuk át a polinomot y szerint. Eredményül egy olyan polinomot kapunk, ahol az együtthatók fordított sorrendben vannak és ennek a polinomnak a gyökei az eredeti gyökök reciprokai. Az új polinomhoz tartozó Frobenius-féle mátrix sornormáját véve kapjuk: $1/|x_k| \le \max_{0<i\le n}\left(1-\delta_{in} + |a_i/a_0|\right) = 1/r$, ahol természetesen feltételeztük, hogy $a_0 \ne 0$. A két eredményt egybevetve látjuk, hogy a polinom gyökei az

$$r \le |x_k| \le R, \quad k = 1,2,\dots,n \tag{17.8}$$

körgyűrű tartományba esenek.

A (17.6)-tal adott polinomoknál előnyösen alkalmazható a Newton-módszer, mert a polinom értéke és a deriváltja egy lépésben egyszerűen számolható. Ha például a ξ helyen szeretnénk ezeket kiszámolni, nem kell mást tennünk, mint a polinomot maradékos osztással elosztani $(x-\xi)^2$-tel:

$$p(x) = q(x)(x-\xi)^2 + \alpha x + \beta\,. \tag{17.9}$$

Könnyen meggyőződhetünk róla, hogy a helyettesítési érték $\alpha\xi + \beta$, a derivált pedig α lesz a ξ helyen.

A többszörös gyökök kiszűrésére alkalmazhatjuk az Euklidészi algoritmust. Ekkor a két induló polinom $p_0(x) = p(x),\;\; p_1(x) = -p'(x)$, az $i+1$-edik polinomot pedig úgy készítjük, hogy $p_{i-1}(x)$-et osztjuk $p_i(x)$-szel és a maradékot képezzük:

$$p_{i-1}(x) = q_i(x)p_i(x) - c_i p_{i+1}(x), \quad i = 1,2,\dots \tag{17.10}$$

A sorozatban a polinomok fokszáma csökkenő, $c_i > 0$, egyébként tetszőleges. Az algoritmus $m \le n$ lépés után befejeződik:

$$p_{m-1}(x) = q_m(x)p_m(x), \quad p_m(x) \ne 0.$$

Az utolsó polinom a két kezdő polinom legnagyobb közös osztója. Mivel a derivált polinom az 1-nél nagyobb multiplicitású gyököket tartalmazza, így ezek a gyökök megjelennek $p_m(x)$-ben.

Abban az esetben, amikor minden gyök valós és egyszeres, akkor az Eulidészi algoritmus olyan polinomsorozatot készít, amely *Sturm-sorozat tulajdonságú*. Legyen az a helyen a sorozat előjelváltásainak száma $V(a)$, a b helyen pedig $V(b)$, ekkor megmutatható, hogy az $[a,b]$ intervallumban a gyökök száma $V(b)-V(a)$.

17.5. Gyakorlatok

1. Készítsük el a (17.9) osztás algoritmusát!
2. Adjuk meg a $4x^5-3x^4+6x^3-5x^2-8x+2$ polinom gyökeit tartalmazó körgyűrűt!

18. Numerikus integrálás (kvadratúra) I.

Az integrálok kiszámításakor nem mindig ismert a primitív függvény, vagy ha igen, némely esetben nagyon bonyolult, nehezen számítható. Ilyenkor a numerikus módszerek a kívánt pontosságú eredmény előállítására egyszerűbb alternatívát kínálnak. A továbbiakban az interpolációból nyerhető kvadratúra-formulákkal fogunk foglalkozni.

Láttuk, a függvény az $[a,b]$ intervallumban a következő módon állítható elő:

$$f = L_n + r_n, \tag{18.1}$$

ahol L_n a Lagrange-interpolációs polinom és r_n a hibatag. (Feltesszük, az alappontok nagyság szerint rendezettek: $x_{i-1} < x_i$ és $x_0 = a$, $x_n = b$.) A kvadratúra-formulák származ-tatási elve:

$$\int_a^b f = \int_a^b L_n + \int_a^b r_n = \sum_{i=0}^{n} a_i f(x_i) + R_n, \tag{18.2}$$

ahol az

$$a_i = \int_a^b l_i(x)dx \tag{18.3}$$

súlyok a Lagrange alappolinomok integrálásából adódnak.

Következmény. Az így nyert formulák legfeljebb n-edrendű polinomig pontosak.

Ekvidisztáns alappontok esetén nyerjük a Newton-Cotes formulákat.

18.1. Zárt és nyilt Newton-Cotes kvadratúra formulák

18.1.1 Definició

Az alappontok halmaza legyen $\Omega_n = \{x_0, \ldots, x_n\}$. *Zárt* a kvadratúra-formula, ha $a,b \in \Omega_n$, $h = (b-a)/n$, $x_k = a + k \cdot h,\ k = 0,1,\ldots,n$. *Nyilt* a formula, ha $a,b \notin \Omega_n$, $h = (b-a)/(n+2)$, $x_k = a + (k+1)\cdot h$, $k = 0,1,\ldots,n$, $x_{-1} = a,\ x_{n+1} = b$.

A továbbiakban rátérünk a zárt Newton-Cotes formulák együtthatóinak előállítására.

$$a_k = \int_a^b l_k(x)dx = \int_a^b \frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)}dx.$$

Vegyük észre: $x_k - x_j = (k-j)h$ és vezessünk be új változót: $t = (x-a)/h$, ahonnan $x = a + th$ és $x - x_j = (t-j)h$, s ezzel

$$\begin{aligned} a_k &= \int_0^n \frac{t(t-1)\ \overbrace{\ldots\ldots\ldots}^{(t-k)\text{ hiányzik}}\ (t-n)h^n}{k(k-1)\ldots 1(-1)(-2)\ldots(-n+k)h^n} h\,dt = \\ &= (b-a)\frac{(-1)^{n-k}}{nk!(n-k)!}\int_0^n \frac{t(t-1)\ldots(t-n)}{t-k}dt = (b-a)B_{k,n}^{\text{zárt}}, \end{aligned} \tag{18.4}$$

ahol a $B_{k,n}^{\text{zárt}}$ együtthatók az intervallumtól függetlenül egyszer s mindenkorra kiszámíthatók. Hasonló módon nyerhetjük a *nyilt Newton-Cotes* formulák együtthatóit:

$$a_k = (b-a)\frac{(-1)^{n-k}}{(n+2)k!(n-k)!}\int_0^{n+2}\frac{(t-1)(t-2)\ldots(t-n-1)}{t-k-1}dt = (b-a)B_{k,n}^{\mathrm{ny}}, \tag{18.5}$$

Az első néhány Newton-Cotes együttható:

Zárt						Nyílt				
1	1				Trapéz	1				Érintő formula
1	4	1			Simpson	1	1			
1	3	3	1			2	-1	2		
7	32	12	32	7		11	1	1	11	

A táblázatban minden sort osztani kell az együtthatók összegével, mert az együtthatók összegének 1-nek kell lenni. Például az 1 4 1 súlyok az 1/6 4/6 1/6 valódi súlyokra utalnak.

18.1.2 Tétel

$$1.\ \sum_{k=0}^{n} B_{k,n} = 1, \qquad 2.\ B_{k,n} = B_{n-k,n}. \tag{18.6}$$

Bizonyítás. Az első állítás az $f(x) \equiv 1$ függvény integrálásából adódik, kihasználva, hogy az integrál 0-adfokú polinomra pontos. A második állítást az $y = n - t$ új változóra való áttéréssel nyerjük. ■

18.2. Néhány egyszerű integráló formula

1. Az *érintőformula* (nyílt Newton-Cotes): $n = 0$, $B_{0,0}^{\mathrm{ny}} = \frac{1}{2\cdot 1\cdot 1}\int_0^2 1\cdot dt$, tehát

$$I_0(f) = (b-a)f\left(\frac{a+b}{2}\right). \tag{18.7}$$

Az érintőformula úgy is értelmezhető, hogy a függvényt $[a,b]$-ben a középponthoz húzott érintő egyenessel közelítjük, és az egyenes alatti területet vesszük. Ez mutatja, hogy legfeljebb elsőfokú polinomig pontos.

18.2.1 Tétel, érintő formula hibája

Legyen $c = (a+b)/2$, $f \in C^2[a,b]$, *ekkor az érintő formulával*

$$\int_a^b f(x)dx = (b-a)f(c) + \frac{(b-a)^3}{24}f''(\eta), \quad \eta \in [a,b]. \tag{18.8}$$

Bizonyítás. A c körüli sorfejtésből

$$f(x) = f(c) + (x-c)f'(c) + \frac{(x-c)^2}{2}f''(\xi_x).$$

Integráláskor az első tag adja a közelítő formulát, a második tag eredménye zérus, így elegendő a hibatagot vizsgálni:

$$R_1(f) = \frac{1}{2}\int_a^b (x-c)^2 f''(\xi_x)dx.$$

Az integrálszámítás középértéktétele szerint

$$R_1(f)=\frac{f''(\eta)}{2}\int_a^b (x-c)^2\,dx=\frac{f''(\eta)}{2}\left[\frac{(x-c)^3}{3}\right]_a^b=\frac{(b-a)^3}{24}f''(\eta). \qquad \blacksquare$$

A gyakorlatban ezt a formulát nem az egész $[a,b]$ intervallumra alkalmazzuk, hanem azt m részre osztjuk, és az egyes részintervallumokban az érintőformulával integrálunk. Például, ha $m=3$: $h=(b-a)/3$ és a három részintervallumra alkalmazzuk a (18.7) szabályt.

A részintervallumon nyert eredmények felösszegzésével jutunk az *érintőszabályhoz:*

$$\int_a^b f=\frac{b-a}{m}\sum_{i=1}^m f(a-h/2+ih)+\frac{(b-a)^3}{24m^2}f''(\eta), \tag{18.9}$$

ahol most $f''(\eta)=\frac{1}{m}(f''(\eta_1)+f''(\eta_2)+\ldots+f''(\eta_m))$, mert a Darboux-tulajdonság miatt f'' ezt az átlagértéket is felveszi valahol a teljes intervallumban.

2. A *trapézformula*. Elsőfokú polinom interpolációból nyert zárt formula: $n=1$, $B_0^z=B_1^z=1/2$:

$$I_1(f)=\frac{b-a}{2}(f(a)+f(b)). \tag{18.10}$$

18.2.2 Tétel, trapézformula hibája

Legyen $f\in C^2[a,b]$, *ekkor*

$$\int_a^b f=\frac{b-a}{2}(f(a)+f(b))-\frac{(b-a)^3}{12}f''(\eta),\quad \eta\in[a,b]. \tag{18.11}$$

Bizonyítás. Az interpoláció hibatagjának integrálja az integrálszámítás középérték tételének felhasználásával:

$$R_1(f)=\int_a^b\frac{f''(\xi_x)}{2}(x-a)(x-b)dx=-\int_a^b\frac{f''(\xi_x)}{2}\underbrace{(x-a)(b-x)}_{\geq 0}dx=-\frac{f''(\eta)}{12}(b-a)^3. \qquad \blacksquare$$

A teljes intervallumot m részre osztva, a részintervallumok eredményét felösszegezve nyerjük a *trapéz szabályt*:

$$\int_a^b f=\frac{b-a}{2m}[f(x_0)+2f(x_1)+\ldots+2f(x_{m-1})+f(x_m)]-\frac{(b-a)^3}{12m^2}f''(\eta). \tag{18.12}$$

18.2.3 Definició

Egy kvadratúra formulát akkor mondunk k*-adrendűnek,* ha k-adfokú az a legkisebb fokszámú polinom, amelyre a formula már nem pontos.

Eszerint az érintő- és trapézformula *másodrendű.*

3. A *Simpson formula*: másodfokú polinom interpolációból nyert zárt formula, $n=2$, $B_0^z=1/6$, $B_1^z=4/6$, $B_2^z=1/6$ és

$$I_2(f)=\frac{b-a}{6}(f(a)+4f(\frac{a+b}{2})+f(b)).$$

Az interpoláció maradéktagjában $\omega_2(x)=(x-a)(x-\frac{a+b}{2})(x-b)$ szerepel, ennek az integrálja $[a,b]$-ben zérus. Ezt legegyszerűbben úgy tudjuk belátni, hogy $[a,b]$-t a $[-1,1]$ intervallumba transzformáljuk. Ekkor $\omega_2(x)$ páratlan függvény, amelynek az integrálja zérus. Emiatt a hibatételt az Hermite-interpolációból származtatjuk,

$$f(x)=H_3(x)+\frac{f^{(4)}(\xi_x)}{4!}(x-a)(x-\frac{a+b}{2})^2(x-b), \tag{18.13}$$

ahol az $(a+b)/2$ középpontban az első deriváltat is interpoláljuk. Az általánosított osztott differenciák táblázatára gondolva tudjuk, hogy az interpoláló polinom a következő alakú: $H_3(x)=L_2(x)+C(x-a)(x-\frac{a+b}{2})(x-b)$. A második tag együtthatójának értéke nem fontos, mert az integrálja az előbbiek alapján zérus, s ezzel $\int_a^b H_3=\int_a^b L_2$.

18.2.4 Tétel, Simpson-formula hibája

Legyen $f\in C^4[a,b]$. *Ekkor létezik* $\eta\in[a,b]$, *amelyre*

$$\begin{aligned}\int_a^b f &= \frac{b-a}{6}\left(f(a)+4f(\frac{a+b}{2})+f(b)\right)-\frac{f^{(4)}(\eta)}{90}\left(\frac{b-a}{2}\right)^5= \\ &=I_2(f)-\frac{f^{(4)}(\eta)}{90}h^5,\end{aligned} \tag{18.14}$$

ahol $h=(b-a)/2$.

<u>Bizonyítás.</u> Kiindulunk az Hermite-interpoláció (18.13) alakjából, ahonnan integrálással kapjuk:

$$I(f)-I_2(f)=\int_a^b \frac{f^{(4)}(\xi_x)}{4!}(x-a)(x-\frac{a+b}{2})^2(x-b)dx.$$

Ahhoz, hogy az integrálszámítás középértéktételét alkalmazhassuk, az $f^{(4)}$ mellett álló tényező nem lehet negatív. Ezt úgy biztosíthatjuk, hogy $(x-b)$ helyett $(b-x)$-et írunk, s így

$$I(f)-I_2(f)=-\frac{f^{(4)}(\eta)}{4!}\int_a^b (x-a)(x-\frac{a+b}{2})^2(b-x)dx=-\frac{f^{(4)}(\eta)}{90}\left(\frac{b-a}{2}\right)^5. \qquad \blacksquare$$

Simpson-szabály. A teljes $b-a$ intervallumot páros számú m részintervallumra osztva és a Simpson-formulát a szomszédos intervallumpárokra alkalmazva kapjuk a Simpson-szabályt, mint összetett formulát. Ekkor három pontonként fogjuk össze a formulákat és az összetett formula:

$$\begin{aligned}\int_a^b f &= \sum_{k=1,3,\ldots}\left(\frac{2h}{6}(f(x_{k-1})+4f(x_k)+f(x_{k+1}))-\frac{f^{(4)}(\eta_k)}{90}h^5\right)= \\ &=\frac{h}{3}\left(f(x_0)+2\sum_{\substack{k \text{ páros} \\ \text{belső pont}}} f(x_k)+4\sum_{k \text{ ptlan}} f(x_k)+f(x_m)\right)-\frac{h^5}{90}\sum_{k \text{ ptlan}} f^{(4)}(\eta_k).\end{aligned} \tag{18.15}$$

A hibatag még tovább írható:

$$-\frac{h^5}{90}\sum_{k\ \mathrm{ptlan}} f^{(4)}(\eta_k) = -\frac{(b-a)^5}{180m^4}\left(\frac{\sum f^{(4)}(\eta_k)}{m/2}\right) = -\frac{(b-a)^5}{180m^4} f^{(4)}(\eta), \tag{18.16}$$

mivel a Darboux-tulajdonság miatt van egy η , amelyre a negyedik derivált az átlagértéket felveszi.

18.3. Példák

1) Az $\int_{-1}^{1}\frac{dx}{2+x}$ integrált az érintőszabállyal közelítjük. Hány osztópontot kell választanunk, hogy az integrált 10^{-2}-nél kisebb hibával kapjuk?

Megoldás. Azt kell biztosítani, hogy $\frac{(b-a)^3}{24m^2}M_2 \le 10^{-2}$ teljesüljön, ahol $b-a=2$ és $M_2 = 2\max_{x\in[-1,1]}\left|(2+x)^{-3}\right| = 2$. A számokat helyettesítve: $200/3 \le m^2$, így $m=9$ megfelel.

2) Határozzuk meg az A_0, A_1, A_2 paramétereket úgy, hogy a $\int_0^2 \sqrt{x} f(x)dx \approx$ $\approx I_2(f) = A_0 f(0) + A_1 f(1) + A_2 f(2)$ kvadratúra legfeljebb másodfokú polinomokra pontos legyen!

Megoldás. Két megoldás is létezik. Az egyik, hogy kiszámítjuk a kijelölt integrálokat a Lagrange-alappolinomokkal: $A_i = \int_0^2 l_i(x)\sqrt{x}dx$, ahol $\sqrt{x}$ -et súlyfüggvénynek tekintjük. A másik módszer szerint felírjuk azt a lineáris egyenletrendszert, ami a pontossági követeléseket tartalmazza. A következő egyenletrendszer első sora azt fejezi ki, hogy a kvadratúra az 1 polinomra pontos, a második sor szerint az x polinomra pontos, a harmadik sor szerint pedig az x^2 polinomra pontos:

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 1 & 4 \end{bmatrix}\begin{bmatrix} A_0 \\ A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} \int_0^2 x^{1/2}dx = 4\sqrt{2}/3 \\ \int_0^2 x\cdot x^{1/2}dx = 8\sqrt{2}/5 \\ \int_0^2 x^2 x^{1/2}dx = 16\sqrt{2}/7 \end{bmatrix}.$$

Ennek a megoldása: $A_0 = \frac{8\sqrt{2}}{105},\ A_1 = \frac{32\sqrt{2}}{35},\ A_2 = \frac{12\sqrt{2}}{35}$.

19. Numerikus integrálás, Gauss-kvadratúrák II.

Az eddigi, interpolációból származtatott kvadratúra-formulák legalább annyiad fokú polinomra pontosak, ahányad fokú polinomból származtattuk őket. A Gauss-kvadratúrák abból az észrevételből származnak, hogy az alappontok speciális megválasztásával a kavadratúra-formula rendje növelhető. Ismét szükségünk lesz az ortogonális polinomokra.

19.1. Tétel, ortogonális polinom gyökei

Legyen $\{p_k(x)\}$ egy ortogonális polinom rendszer. Ekkor bármely n-re $p_{n+1}(x)$ gyökei valósak, egyszeresek és az $[a,b]$ intervallumban vannak, ahol $[a,b]$ a skalárszorzat integrálási tartománya.

Bizonyítás. Legyenek $x_0, x_1, \ldots, x_k$ $p_{n+1}(x)$ páratlan multiplicitású gyökei $[a,b]$-ben, azaz ott $p_{n+1}(x)$ előjelet vált. Ha $k=n$, akkor a tétel állítása rendben van. Ha nem, akkor indirekt úton feltesszük: $k<n$ és megmutatjuk, hogy az állítás ellentmondásra vezet. Ehhez tekintsük a $q(x)=(x-x_0)(x-x_1)\ldots(x-x_k)$ $k+1$-edfokú polinomot. Mivel $k+1<n+1$, az ortogonalitás miatt $(p_{n+1},q)=0$. De ezzel ellentmondásra jutunk, mert $p_{n+1}(x)q(x)$ nem vált előjelet $[a,b]$-ben, mivel p_{n+1} minden előjelváltását $q(x)$ megszünteti és így $\int p_{n+1}q\alpha \neq 0$ volna. Vegyük észre, a gondolatmenet akkor is jó, ha egyetlen páratlan multiplicitású gyök sincs, mert ekkor $q(x)$ 0-adfokú.
■

Az $n+1$-pontos Gauss-kvadratúrát úgy kapjuk, hogy a $p_{n+1}(x)$ ortogonális polinom gyök-helyein készítjük az interpolációból származtatott kvadratúra-formulát. A séma a következő:

$$\int_a^b f\alpha = \int_a^b L_n\alpha + \int_a^b r_n\alpha = \sum_{i=0}^{n} a_i f(x_i) + R_n, \quad a_i = \int_a^b l_i(x)\alpha(x)dx. \tag{19.1}$$

19.2. Tétel, Gauss-kvadratúra pontossága

Legyenek a $p_{n+1}(x)$ ortogonális polinom gyökei $x_0, x_1, \ldots, x_n$, $a_i = \int l_i\alpha$, ahol l_i az i-edik Lagrange alappolinom a fenti alappontokon. Ekkor a Gauss-kvadratúra

$$G_n(f) = \sum_{i=0}^{n} a_i f(x_i)$$

pontos minden legfeljebb $2n+1$-edfokú polinomra: $f \in \mathcal{P}_{2n+1} \to \int f\alpha = G_n(f)$.

Bizonyítás. Az interpolációból való származatatás miatt $G_n(f)$ biztosan pontos a legfeljebb n-edfokú polinomokra. Tegyük fel, $f \in \mathcal{P}_{2n+1}$, $f = p_{n+1}\cdot q + r$, $q, r \in \mathcal{P}_n$, így

$$G_n(f) = \sum_{i=0}^{n} a_i f(x_i) = \sum_{i=0}^{n} a_i [\underbrace{p_{n+1}(x_i)}_{=0 \text{ minden } i-\text{re}} \cdot q(x_i) + r(x_i)] =$$

$$= \sum_{i=0}^{n} a_i r(x_i) = G_n(r) = \int r\alpha = \qquad \text{(mert } n\text{-edfokig pontos)}$$

$$= \int (p_{n+1}\cdot q + r)\alpha = \qquad \text{(mert } q \in \mathcal{P}_n, \text{ így ortogonális } p_{n+1}\text{-re)}$$

$$= \int f\alpha.$$

19.2.1 Következmény

Az a_i együtthatók pozitívak.

Bizonyítás. Tudjuk, $l_i(x_j)=l_i^2(x_j)=\delta_{ij}$, ahol δ_{ij} a Kronecker-delta, $l_i^2(x)\ge 0$ és $l_i^2(x)\in\mathcal{P}_{2n}$, következésképp a Gauss-kvadratúra pontos :

$$0<\int l_i^2\alpha=G_n(l_i^2)=\sum_{j=0}^{n}a_i l_i^2(x_j)=a_i.$$

Az $f(x)=1$ függvény integrálásával most a következőt kapjuk az együtthatók összegére:

$$\sum_{i=0}^{n}a_i=\int\alpha=\mu_0,\quad \mu_i=\int x^i\alpha,$$

ahol μ_0 a nulladik momentum. Vegyük észre, ez egyenlő $b-a$ -val, ha a súlyfüggvény 1.

19.2.2 Tétel, Gauss-kvadratúra hibaformulája

Legyen $f\in C^{2n+2}[a,b]$ és $G_n(f)=\sum_{k=0}^{n}a_k f(x_k)$, ahol az alappontok $p_{n+1}(x)$ gyökei. Akkor

$$I(f)-G_n(f)=\frac{f^{(2n+2)}(\eta)}{(2n+2)!}\left(p_{n+1},p_{n+1}\right), \tag{19.2}$$

itt $p_{n+1}(x)$ 1-főegyütthatós ortogonális polinom.

Bizonyítás. Hermite-Fejér interpolációból (amikor az interpolációban a függvényértékek és az első deriváltak vesznek részt) kapjuk a következő hiba-előállítást:

$$f(x)=H_{2n+1}(x)+\frac{f^{(2n+2)}(\xi_x)}{(2n+2)!}\underbrace{(x-x_0)^2(x-x_1)^2\dots(x-x_n)^2}_{=p_{n+1}^2(x)}.$$

Innen az integrálás középérték-tételének alkalmazásával

$$I(f)-G_n(f)=\int_a^b\frac{f^{(2n+2)}(\xi_x)}{(2n+2)!}\underbrace{p_{n+1}^2(x)}_{\ge 0}\alpha(x)dx$$

nyerjük az állítást, mert $H_{2n+1}(x)$ -re a Gauss-kvadratúra pontos. ■

Megadunk néhány 1-főegyütthatós ortogonális polinomot:

Név	$[a,b]$	$\alpha(x)$	μ_0	α_{n+1}	β_n	p_0	p_1	p_2
Legendre	$[-1,1]$	1	2	0	$n^2/(4n^2-1)$	1	x	$x^2-1/3$
Csebisev	$[-1,1]$	$\frac{1}{\sqrt{1-x^2}}$	π	0	1/4, de $\beta_1=1/2$	1	x	$x^2-1/2$
Laguerre	$[0,\infty]$	e^{-x}	1	$2n+1$	n^2	1	$x-1$	x^2-4x+2
Hermite	$[-\infty,\infty]$	e^{-x^2}	$\sqrt{\pi}$	0	$n/2$	1	x	$x^2-1/2$

19.3. Példák

1. Három-pontos Gauss-Csebisev kvadratúrával közelítjük a $\int_{-1}^{1}\left(1-x^2\right)^{-1/2} e^{-x}dx$ integrált. Becsüljük meg a hibát!

Megoldás. A három-pontos kvadratúránál $n=2$, ezt alkalmazzuk a (19.2) formulánál: $M_6 = e$ és mivel 1-főegyütthatós polinomoknak kell szerepelni, emiatt $p_3(x) = T_3(x)/4$. Így a hiba kisebb, mint $e\cdot(T_3,T_3)/(16\cdot 6!) = e\pi/(32\cdot 720)$, mert $(T_3,T_3)=\pi/2$, (lásd a 7.4 feladatot).

2. Készítsük el a két-pontos Gauss-Hermite kvadratúra súlyait! Ellenőrízzük, hogy az így kapott kvadratúra legfeljebb harmadfokú polinomokra pontos!

Megoldás. A két-pontos kvadratúránál a másodfokú Hermite-polinom gyökei $-x_0 = x_1 = 2^{-1/2}$. Így $a_0 = \int_{-\infty}^{\infty}\frac{(x-x_1)}{(x_0-x_1)}\exp(-x^2)dx = \frac{x_1\mu_0}{2x_1} = \mu_0/2$, mert az elsőfokú tag integrálja zérus, lévén az integrandus páratlan függvény. Hasonlóan adódik, hogy $a_1 = a_0$. A kapott kvadaratúra pontos az 1 függvényre, mert az eredmény μ_0. Az x és x^3 függvényre is pontos, mert a két tag a gyökhelyeken a páratlanság miatt kiejti egymást. Így már csak azt kell igazolunk, hogy a pontosság a másodfokú x^2 polinomra is teljesül. Ennek az integrálja a (9.8) formula alapján nem más mint $\left(p_1,p_1\right) = \beta_1\left(p_0,p_0\right)$, ami az előző oldalon látható táblázat szerint egyenlő $\mu_0/2$-vel. A Gauss-Hermite kvadratúra eredménye pedig $\frac{\mu_0}{2}(\frac{1}{2}+\frac{1}{2})$, tehát igaz az egyezés.

3. Határozzuk meg a következő integrál értékét:

$$\int_{-1}^{1}\frac{(2x^2+x)dx}{\sqrt{1-x^2}}.$$

Megoldás. A számlálóban szereplő polinomot előállítjuk az első három Csebisev polinom lineáris kombinációjaként: $2x^2 + x = c_0T_0 + c_1T_1 + c_2T_2$. Ezt felhasználva az integrálunk így is írható: $\left(T_0, c_0T_0 + c_1T_1 + c_2T_2\right) = c_0\left(T_0,T_0\right) = c_0\pi$. Az első három Csebisev polinom együtthatóival a következő lineáris egyenletrendszert írhatjuk fel (9.5) alapján:

$$\begin{bmatrix} 1 & 0 & -1 \\ & 1 & 0 \\ & & 2 \end{bmatrix}\begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix},$$

amiből $c_0 = 1$, tehát az integrál értéke π .

20. Közönséges differenciálegyenletek

20.1. Alapfogalmak

Az

$$F(x,y,y',y'',\dots,y^{(m)})=0 \qquad (20.1)$$

alakú egyenletet *közönséges differenciálegyenletnek* nevezzük, ahol keresendő az $y=y(x)$ függvény, mely a fenti kifejezésben deriváltjaival együtt szerepel. Ha y m-szer differenciálható $[0,1]$-ben és kielégíti (20.1)-et, akkor y egy *megoldás*. Ismeretes, a differenciálegyenleteknek sok megoldásuk van. A megoldást úgy tudjuk *egyértelművé* tenni, hogy a peremen még feltételeket teszünk a függvény vagy deriváltjainak értékére. Ha ezen feltételek mindegyike a kezdőpontban (általánosabban fogalmazva: csak egy pontban) van megadva, akkor *kezdetiérték feladatról* beszélünk, ha pedig több ponthoz kapcsolódnak a feltételek, akkor *peremérték feladatunk* van.

A differenciálegyenlet *lineáris,* ha y és deriváltjainak lineáris kombinációja szerepel (20.1)-ben és a differenciálegyenlet *explicit* alakú, ha a legmagasabb derivált explicit módon kifejezhető:

$$y^{(m)}=f(x,y,y',\dots)\,. \qquad (20.2)$$

Világos, minden lineáris differenciálegyenlet ebbe a kategóriába tartozik, és a gyakorlatban előforduló nemlineáris differenciálegyenletek zöme is ilyen. A továbbiakban explicit elsőrendű differenciálegyenletek megoldásának numerikus közelítéseivel fogunk foglalkozni, amikor kezdetiérték feladatról van szó. Ekkor a kezdetiérték feladat

$$y(0)=y_0,\quad y'=f(x,y),\quad x\in[0,1], \qquad (20.3)$$

ahol $f:\mathbb{R}^2\to\mathbb{R}$, x_0 és y_0 adottak, és keressük azt az $y:[0,1]\to\mathbb{R}$ függvényt, amely felveszi x_0-ban az y_0 értéket és az egyenletet kielégíti. Az egyszerűség kedvéért szoritkozunk a $[0,1]$ intervallumra, hiszen tudjuk, az $[a,b]$ intervallum lineáris transzformációval ide átvihető. A megoldás létezésére és egyértelműségére vonatkozik a következő tétel, melyet bizonyítás nélkül idézünk.

20.1.1 Tétel, a kezdetiérték feladat egyértelműsége

Ha f folytonos egy $(x,y)\in[0,1]\times[c,d]$ téglán és f a második változója szerint eleget tesz a Lipschitz-feltételnek, azaz létezik L:

$$\left|f(x,y_1)-f(x,y_2)\right|\le L\left|y_1-y_2\right|,\ \ \forall x\in[0,1],\ \ y_1,y_2\in[c,d]\,, \qquad (20.4)$$

akkor a kezdetiérték feladatnak egyértelmű megoldása van.

A numerikus eljárás során a megoldást az $x_n=nh$ osztópontokban közelítjük, ahol $h=1/N$ és a numerikus közelítést x_n-ben y_n-nel jelöljük. Természetesen azt szeretnénk, hogy y_n minél közelebb legyen a pontos értékhez, $y(x_n)$-hez.

20.1.2 Az Euler-módszer

Ez a legegyszerűbb módszer, amit a differencia-hányadosból származtatunk a következő közelítő érték előállítására:

$$\frac{y_{n+1}-y_n}{h}=f(x_n,y_n)\ \ \to\ \ y_{n+1}=y_n+hf(x_n,y_n). \qquad (20.5)$$

Példa: $y(0)=1,\ y'=y$. Ennek a megoldása $y=e^x$ és (20.5)-ből $y_n=(1+h)^n$. Ha $n\to\infty$, akkor az analízisből tudjuk, hogy $y_n=(1+h)^n=(1+x_n/n)^n \ \to\ e^{x_n}$.

20.1.3 Definíció, konvergencia

Egy numerikus módszer *lokálisan konvergens* az $x\in[0,1]$ pontban, ha $h=x/n,\ x=x_n$ mellett

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

teljesül. Ha a módszer konvergens $\forall x\in[0,1]$ -re, akkor azt mondjuk, a *módszer konvergens.*

20.1.4 Definíció, konvergencia-sebesség

Egy konvergens numerikus módszer sebessége p -edrendű, $(1\le p)$, ha $h=x/n,\ x=x_n$ mellett $\exists M$ úgy, hogy az

$$\left|y(x)-y_n\right|\le Mh^p \tag{20.6}$$

hibabecslés teljesül, ahol M független h-tól és n-től.

20.1.5 Definíció, lokális hiba v. képlethiba

Ez annak a képletnek a hibája, amellyel a következő függvényértéket közelítjük. Ilyenkor a képletbe mindenütt a pontos értéket írjuk. Például az Euler-módszernél a

$$g_{i-1}=y(x_i)-y(x_{i-1})-hf(x_{i-1},y(x_{i-1})),\ \ i=1,2,\ldots,N \tag{20.7}$$

mennyiségek a *lokális hibák vagy képlethibák.*

20.1.6 Definíció, konzisztencia

Ha $\exists p\ge 1$ és M konstans úgy, hogy

$$\left|g_i\right|\le Mh^{p+1},\ \ i=1,2,\ldots,N \tag{20.8}$$

akkor a módszer *p -edrendben konzisztens.*

A definició alapján világos, nagyobb p -re pontosabb megoldás várható.

20.1.7 Definició, globális hiba

Jelölje a pontos és numerikus megoldás eltérését

$$e_i=y(x_i)-y_i,\ \ i=1,2,\ldots,N,\ \ h=1/N,\ \ \ x_i=ih. \tag{20.9}$$

Az e_i mennyiségeket *globális hibának* nevezzük.

20.1.8 Definició, stabilitás

A numerikus módszer stabil, ha van olyan K konstans, amellyel

$$\left|e_i\right|\le K\left(\left|e_0\right|+\sum_{j=1}^{i}\left|g_j\right|\right),\ \ i=1,2\ldots,N \tag{20.10}$$

teljesül, vagyis a globális hiba felülről becsülhető a lokális hibák abszolút összegével.

20.1.9 Tétel, numerikus módszer konvergenciája

Ha egy numerikus módszer stabil és p-edrendben konzisztens, akkor p-edrendben konvergens.

Bizonyítás. Ha $x=0$, akkor a tétel igaz. Ha $x \in (0,1]$, legyen $h = x/n,\ x_n = x\ \forall n$-re. Először a stabilitás, majd a konzisztencia definicióját felhasználva

$$|e_i| \le K\left(|e_0| + \sum_{j=1}^{n} |g_j|\right) \le K\sum_{j=1}^{n} ch^{p+1} \le$$
$$\le Kcnh^{p+1} \le Kc(nh)h^p \le (Kc)h^p,$$

ahol $e_0 = y(x_0) - y_0 = 0$. Ezzel p-edrendű konvergencia-sebességre jutottunk. ■

20.2. Az Euler-módszer vizsgálata

Látjuk, az Euler-módszerre a konzisztenciát és stabilitást kéne megmutatni ahhoz, hogy a konvergenciát igazoljuk.

20.2.1 Tétel, konzisztencia

Az Euler-módszer elsőrendben konzisztens, vagyis $g_i = O(h^2)$ teljesül, ha a megoldás kétszer folytonosan differenciálható.

Bizonyítás. Tekintsük a lokális hibát az i-edik pontban és $y(x_{i+1})$-et fejtsük sorba az x_i hely körül:

$$g_i = y(x_{i+1}) - y(x_i) - hf(x_i, y(x_i)) =$$
$$= y(x_i) + hy'(x_i) + \frac{h^2}{2!}y''(\xi_i) - y(x_i) - hy'(x_i),$$

emiatt

$$|g_i| = \frac{h^2}{2!}|y''(\xi_i)| \le \|y''\|_\infty \frac{h^2}{2!},$$

tehát $p+1=2$, $p=1$, elsőrendű a konzisztencia. ■

20.2.2 Tétel

Az Euler-módszer stabil, ha f eleget tesz a második változója szerint a Lipschitz-feltételnek.

Bizonyítás. Vegyük az i-edik pontban a lokális hiba képletét, a módszer képletét és vonjuk ki őket egymásból:

$$y(x_{i+1}) = y(x_i) + hf(x_i, y(x_i)) + g_i \qquad /(+)$$
$$y_{i+1} = y_i + hf(x_i, y_i) \qquad /(-)$$
$$e_{i+1} = e_i + h[f(x_i, y(x_i)) - f(x_i, y_i)] + g_i$$

Mivel f eleget tesz a Lipschitz-feltételnek:

$$|e_{i+1}| \le |e_i| + h|f(x_i, y(x_i)) - f(x_i, y_i)| + |g_i| \le |e_i|(1+hL) + |g_i|.$$

Fejtsük vissza a rekurziót e_0-ig:

$$|e_{i+1}| \le |e_i|(1+hL) + |g_i| \le (|e_{i-1}|(1+hL) + |g_{i-1}|)(1+hL) + |g_i| \le$$
$$\le (1+hL)^2 |e_{i-1}| + (1+hL)|g_{i-1}| + |g_i| \le (1+hL)^{i+1}|e_0| + \sum_{k=0}^{i}(1+hL)^{i-k}|g_k|,$$

ebből megkapjuk a kívánt becslést, ha felhasználjuk az $(1+hL)^j \le e^{jhL} = e^{x_j L} \le e^L$ relációt:

$$|e_{i+1}| \le e^L |e_0| + \sum_{k=0}^{i} e^L |g_k| = e^L \left(|e_0| + \sum_{k=0}^{i} |g_k| \right). \qquad \blacksquare$$

A továbbiakban rátérünk néhány fontosabb módszercsalád rövid ismertetésére.

20.3. Taylor-polinomos módszerek

Az Euler-módszer nem egyéb, minthogy vesszük a függvény elsőfokú Taylor-polinomját x_n körül és annak segítségével lépünk a következő pontba. Ebből az ötletből kiindulva készíthetünk magasabbrendű módszereket is. A következő módszert m-edrendű Taylor-polinomos módszernek nevezzük:

$$y_{n+1} = y_n + y'(x_n) + \ldots + y^{(m)}(x_n)h^m / m!, \quad n = 0,1,\ldots,N. \tag{20.11}$$

Fontos kérdés, egyáltalán kiszámíthatók-e ezek a deriváltak. Ha f elegendően sokszor differenciálható, ennek elvi akadálya nincs. Sokszor azonban súlyos gyakorlati nehézség, hogy a deriváltak nagyon hosszú, nehezen kezelhető formulákat eredményeznek. A konstrukcióból látható, m-edrendben konzisztens módszerre vezet a fenti eljárás.

20.4. Runge-Kutta módszerek

Ezek a Taylor-polinomos módszerek fenti nehézségét küszöbölik ki: nem kell magasrendű deriváltakat számolni, *a magasabb konzisztencia-rend rekurzív függvényhívásokkal is elérhető.* Az általános alak:

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + ha_2, y_n + hb_{21}k_1), \\ &\vdots \\ k_j &= f(x_n + ha_j, y_n + h\sum_{l=1}^{j-1} b_{jl}k_l), \quad j = 1,2,\ldots,s, \\ y_{n+1} &= y_n + h\sum_{j=1}^{s} c_j k_j. \end{aligned} \tag{20.12}$$

Az előre kiszámolt a_i, b_{ij}, c_i paraméterek határozzák meg a konkrét módszert. Az s szám a rekurzió mélysége, más szóval: az egy lépés megtételéhez szükséges függvényhívások száma. A következő, ún. *módosított Euler-módszer* Runge-től származik:

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + h/2, y_n + (h/2)k_1), \\ y_{n+1} &= y_n + hk_2. \end{aligned} \tag{20.13}$$

Deriválással megmutatható, hogy másodrendben konzisztens. Hasonlóképp másodrendű a következő Runge-Kutta módszer, amelyet az egyszerűsége miatt egy sorban írunk fel:

$$y_{n+1} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_n + hf(x_n, y_n))]. \tag{20.14}$$

Az igen népszerű negyedrendű Runge-Kutta módszer negyedrendben konzisztens és stabil, vagyis negyedrendű konvergens módszer:

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + h/2, y_n + hk_1/2), \\ k_3 &= f(x_n + h/2, y_n + hk_2/2), \\ k_4 &= f(x_n + h, y_n + hk_3), \\ y_{n+1} &= y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4). \end{aligned} \tag{20.15}$$

A Runge-Kutta módszerek lokális hibája:

$$g_n = y(x_n) - y(x_{n-1}) - h\sum_{j=1}^{s} c_j k_j(x_{n-1}, y(x_{n-1}), h), \quad n = 1,\ldots,N, \tag{20.16}$$

ahol $k_j(x_{n-1}, y(x_{n-1}), h)$ azt jelenti, hogy a k_j sorozatot a pontos $y(x_{n-1})$ értékkel indítva számoljuk végig. Egy adott rend eléréséhez a lokális hibát sorbafejtjük és a paramétereket úgy választjuk, hogy a tagok minél magasabb rendig eltűnjenek.

20.5. Lineáris többlépéses módszerek

Ezek is az Euler-módszer általánosításának tekinthetők. Az s -lépéses módszer általános alakja:

$$\sum_{k=0}^{s} \alpha_k y_{i+k} = h\sum_{k=0}^{s} \beta_k f_{i+k}, \quad f_{i+k} = f(x_{i+k}, y_{i+k}) \tag{20.17}$$

ahol $\alpha_k, \beta_k,\ k = 0,1,\ldots,s$ adottak, $\alpha_s = 1$ valamint $|\alpha_0| + |\beta_0 \neq 0|$ teljesül. Például az Euler-módszerre $s=1,\ \alpha_0 = -1,\ \alpha_1 = 1,\ \beta_0 = 1,\ \beta_1 = 0.$ A módszer indításához az y_0 értékén túl kellenek még az $y_1, y_2, \ldots, y_{s-1}$ értékek. Ha $\beta_s = 0$, akkor a módszer *explicit* és könnyen számolható. Ha $\beta_s \neq 0$, akkor a módszer implicit és a következő y_{i+s} érték az adódó nemlineáris egyenletből számítandó. Vegyük észre, az explicit módszernél minden lépésben egy új $f(x,y)$ típusú függvény-kiértékelés kell, (mivel a többi már korábbról megvan), ugyanakkor az implicit módszernél még ügyes esetben is legalább 2-3 kiértékelésre szükség van. A mondottakat két egyszerű példán szemléltejük.

Középpontszabály. Explicit, 2-lépéses módszer, a centrális differenciahányadosból származtatható:

$$y_{n+1} = y_{n-1} + 2hf_n, \quad n = 1,2,\ldots,N \tag{20.18}$$

tehát $\alpha_0 = -1,\ \alpha_1 = 0,\ \alpha_2 = 1,\ \beta_0 = 0,\ \beta_1 = -2,\ \beta_2 = 0$. Bebizonyítható, hogy a módszer másodrendben konzisztens és stabil, ha f a második változójában eleget tesz a Lipschitz-feltételnek. A középpontszabály indításához legalább másodrendűen pontos y_1 értéket kell előállítani, amit megtehetünk valamely Runge-Kutta vagy Taylor-polinomos módszerrel, különben a másodrendű konvergencia nem lesz igaz.

Trapézszabály. Ez implicit módszer. Úgy származtatható, hogy a differenciálegyenletet átírjuk integrál alakba és a jobb oldalon keletkező integrált a trapézmódszerrel közelítjük:

$$y_{n+1} = y_n + \frac{h}{2}\left(f_n + f_{n+1}\right), \quad n = 0,1,\ldots,N-1. \tag{20.19}$$

így $s=1,\ \alpha_0 = -1,\ \alpha_1 = 1,\ \beta_0 = \beta_1 = 1/2.$ Itt a következő pont előállításához meg kell oldanunk y -ra az

$$y = y_n + \frac{h}{2}\left(f_n + f(x_n + h, y)\right) = F(y)$$

fixpont egyenletet. $F(y)$ az általános feltételünk alapján eleget tesz a Lipschitz-feltételnek $hL/2$ állandóval, így $F(y)$ kontrakció, ha $h < 2/L$. Célszerű formája az iterációnak:

$$\begin{aligned} y_{n+1}^{0} &= y_n + hf_n, && \text{kezdőérték az Euler-módszerből} \\ y_{n+1}^{k+1} &= y_n + \frac{h}{2}\left(f_n + f(x_{n+1}, y_{n+1}^{k}\right), \quad k \geq 0. \end{aligned} \tag{20.20}$$

Leállás: ha $\left|y_{n+1}^{k+1} - y_{n+1}^{k}\right| < \varepsilon\left(1-q\right)$, ahol q a konvergencia-tényező és ε a kívánt hibakorlát.

A trapézmódszer másodrendben konzisztens és stabil, ha $f \in C^2([0,1]\times\mathbb{R})$ és $h < 1/L$, tehát ekkor másodrendű konvergenciára számíthatunk.

További többlépéses módszereket a zárt vagy a jobbról nyilt kvadratúra-formulák segítségével lehet származtatni.

20.6. Aszimptotikus stabilitás

A gyakorlati számítások szempontjából nem elegendő, ha egy módszer konzisztens és stabil. A kezdeti hiba (- ha y_0 is hibával terhelt, -) továbbterjedése szempontjából fontos egy további stabilitási tulajdonság. Egy differenciálegyenlet-megoldó numerikus módszer *aszimptotikusan stabil,* ha az

$$y(0) = 1, \;\; y'(x) = q\,y(x), \;\; q < 0 \tag{20.21}$$

tesztfeladatra alkalmazva a kapott numerikus közelítések sorozatára minden $h > 0$ lépésköz esetén fennáll $y_n \to 0$, ha $n \to \infty$. Ezt a stabilitási fogalmat a szakirodalom gyakran A_0-stabilitásnak nevezi.

A tesztfeladat megoldása x növekedésével zérushoz tart: $y(x) = e^{qx} \to 0, \;\; x \to \infty$, mivel q negatív. Így a definició azt követeli a módszertől, hogy a lépésköztől függetlenül a numerikus megoldás is tartsa meg ezt a lecsengő jelleget. Kimutatható, az Euler-módszer nem aszimptotikusan stabil, de a trapézmódszer az.