1. fejezet

Bevezetés

A földtudományi kutatások során gyakran merülnek fel olyan számítások, amelyeket csak közelítőleg tudunk elvégezni. Ez a jegyzet a közelítő számítási módszerek alapjaival foglalkozik.

1.1. Hibafogalmak

Jelöljük a-val valaminek a pontos értékét, amit szeretnénk kiszámítani. Legyen ez most egy valós szám. Jelölje számításunk eredményét $\tilde{a}$. Az $\tilde{a}$ valós szám többnyire nem egyezik meg pontosan az a-val, általában meg kell elégednünk azzal, ha kellően közel van hozzá, azaz az $|a - \tilde{a}|$ távolság kellően kicsi. Ezt $\tilde{a} \approx a$ jelölje. (Kiolvasva: $\tilde{a}$ közelíti a-t.)

1.1.1 Definíció. *A*

$$\Delta a := a - \tilde{a}$$

számot az $\tilde{a}$ közelítő érték **abszolút hibá***jának nevezzük.*

Természetesen az a pontos értékét általában nem ismerjük, így az abszolút hibát sem. Azt azonban mindig tudnunk kell, hogy közelítő számításunk eredményét legfeljebb mekkora hiba terheli, azaz legfeljebb mekkora az $|a - \tilde{a}|$ eltérés.

1.1.2 Definíció. *A $\Delta_a \geq 0$ számot az $\tilde{a}$ közelítő érték* **abszolút hibakorlát***jának nevezzük, ha*

$$|a - \tilde{a}| = |\Delta a| \leq \Delta_a.$$

Erre az $a = \tilde{a} \pm \Delta_a$ jelölést is alkalmazzuk.
Nyilvánvalóan egy Δ_a-nál nagyobb szám is abszolút hibakorlát.

1.1.3 Megjegyzés. *A közelítés és az abszolút hibakorlát fogalma minden olyan struktúrában értelmes, ahol van távolság, tehát a és $\tilde{a}$ vehető tetszőleges (X, ρ) metrikus térből. Ekkor az $|a - \tilde{a}|$ távolságot $\rho(a, \tilde{a})$ helyettesíti.*

Az abszolút hiba önmagában nem elegendő a közelítés jóságának jellemzéséhez. Ha csak annyit tudunk például, hogy egy hőmérsékletszámítás során 50^oC abszolút hibát vétettünk, akkor mást jelent ez, ha a napfelszín, és mást, ha Budapest átlaghőmérsékletének kiszámításáról van szó. Az első esetben számításunk egészen elfogadható pontosságú, míg a második esetben rendkívül pontatlan. Ezért ha a és $\tilde{a}$ szám, akkor azt is hasznos lehet tudnunk, hogy hogyan viszonyul a Δa hiba vagy a Δ_a hibakorlát az $\tilde{a}$-hoz.

1.1.4 Definíció. *A $\delta a := \Delta a/|\tilde{a}|$ számot az $\tilde{a}$ közelítő érték* **relatív hibá***jának nevezzük.*

Megjegyezzük, hogy amennyiben a nevezőben szereplő $|\tilde{a}|$ szám nullához igen közel esik, a relatív hiba annak ellenére nagy lehet, hogy az abszolút hiba esetleg nagyon kicsi. Ezért ilyen esetben pusztán a relatív hibát nézve tévesen támadhat az a benyomásunk, hogy rossz eredményt kaptunk. Célszerű tehát az abszolút és a relatív hibát együtt vizsgálni.

1.1.5 Definíció. *A δ_a számot az $\tilde{a}$ közelítő érték egy* **relatív hibakorlát***jának nevezzük, ha $|\delta a| \leq \delta_a$.*

Nyilvánvalóan, ha Δ_a abszolút hibakorlát, akkor $\frac{\Delta_a}{|\tilde{a}|}$ egy relatív hibakorlát lesz. A relatív hibát ill. hibakorlátot az a pontos értékhez viszonyítva is lehet értelmezni. A gyakorlatban azonban a többnyire nem ismert, ezért ezt nem mindig tudjuk kiszámítani. (Ha $\tilde{a}$ elég közel van a-hoz, akkor persze a kétféle relatív hiba is közel egyenlő.)

Feladat. Közelítsük a π számot a két tizedesjegyre kerekített értékével, azaz 3,14-dal. Adjunk a közelítő értékre abszolút és relatív hibakorlátot.
Megoldás: $\Delta\pi = \pi - 3,14$ a közelítés abszolút hibája. Mivel $|\Delta\pi| \leq 5 \cdot 10^{-3}$, ezért $\Delta_\pi = 5 \cdot 10^{-3}$ a közelítés egy abszolút hibakorlátja. A $\delta\pi = \frac{\pi - 3,14}{3,14}$ szám a közelítés relatív hibája. Mivel $|\delta\pi| = \frac{|\pi - 3,14|}{3,14} \leq \frac{5 \cdot 10^{-3}}{3} \leq 1,667 \cdot 10^{-3}$, ezért $\delta_\pi = 1,667 \cdot 10^{-3} = 0,1667\%$ a közelítés egy relatív hibakorlátja.

1.2. A számítás során fellépő problémák

Számításunk eredménye szinte mindig eltér a keresett pontos értéktől. Ennek oka az, hogy a számítás során különféle hibák léphetnek fel. Tekintsük át a lehetséges hibaforrásokat!

1. **Tévedés, géphiba.** Előfordul néha, hogy egy papíron végzett számítást véletlenül elrontunk, vagy számítógépes programunkba valamilyen hiba csúszik, és ezért hibás eredményt kapunk. Az ilyen természetű hibák matematikai szempontból érdektelenek, nem tartoznak a numerikus módszerek tárgykörébe. Bennünket sokkal inkább a 2. és a 3. hibaforrás érdekel.

2. **Képlethiba.** A számítás eredménye (a kimenő adat vagy adatok) a számításhoz felhasznált (bemenő) adatok valamilyen függvénye, amit szokásosan képletnek nevezünk. Ha a képlet nem építhető fel azokból a műveletekből, amelyeket a gép ismer, akkor hiba keletkezik. Például az e szám egyik képlete a jól ismert határérték:

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n. \tag{1.1}$$

A gép azonban nem tud határértéket számítani, ezért az (1.1) képlet helyett másikat használ, pl. a határértéket az $\left(\left(1 + \frac{1}{n}\right)^n\right)$ sorozat valamely elég nagy indexű tagjával helyettesíti. Ez azonban csak közelítése lesz az e számnak. A különbség oka a pontos képletnek egy másik képlettel való helyettesítése, ami képlethibát eredményez.

3. **Öröklött hiba** (vagy a **függvényérték hibája**). Ez a hibaforrás akkor lép fel, ha a bemenő adatok hibával terheltek. (Ennek oka lehet pl. a mérés pontatlansága, ha a bemenő adatok mérésből származnak.) Ilyenkor fontos tudnunk, hogy amennyiben az a szám a bemenő adat, és ez Δ_a hibával ismert, azaz $|a - \tilde{a}| \leq \Delta_a$, akkor az a-tól függő $f(a)$ kimenő adathoz mennyire lesz közel $f(\tilde{a})$, azaz legfeljebb mekkora lehet az $|f(a) - f(\tilde{a})|$ távolság. Az öröklött hiba becslési módszereivel külön alfejezetben foglalkozunk.

4. **A számábrázolásból eredő hiba.** Ez, a számítástechnika tárgykörébe tartozó hibafajta abból adódik, hogy a számítógépen a valós számoknak csak egy véges részhalmazát tudjuk ábrázolni. Ennek tárgyalásával mi nem foglalkozunk.

Természetesen az említett hibaforrások egyszerre is felléphetnek.

Példák képlethibára

1. Tekintsük az e szám már említett közelítő képletét:

$$e \approx \left(1 + \frac{1}{n}\right)^n, \quad \text{ahol } n \in \mathbb{N} \text{ kellően nagy, rögzített index.}$$

a) Adjunk meg erre a közelítésre egy abszolút hibakorlátot!
b) Hányadik tagig kell elmennünk a sorozatban ahhoz, hogy a hiba legfeljebb 10^{-10} legyen?

Megoldás:
a) Feladatunk felső becslést találni az $|e - (1 + \frac{1}{n})^n|$ kifejezésre. Ismeretes, hogy az $((1 + \frac{1}{n})^n)$ sorozat alulról tart e-hez, míg az $((1 + \frac{1}{n})^{n+1})$ sorozat felülről szintén e-hez tart. Ezért minden $n \in \mathbb{N}$-re érvényes a következő becslés:

$$\left|e - \left(1 + \frac{1}{n}\right)^n\right| = e - \left(1 + \frac{1}{n}\right)^n < \left(1 + \frac{1}{n}\right)^{n+1} - \left(1 + \frac{1}{n}\right)^n = \left(1 + \frac{1}{n}\right)^n \left(1 + \frac{1}{n} - 1\right) < \frac{e}{n}$$

Mivel $\frac{e}{n} < \frac{3}{n}$, ezért az $\frac{e}{n}$ felső becslés helyett a racionális $\frac{3}{n}$ becslés is használható.
b) Pl. a $\frac{3}{n}$ becsléssel számolva a sorozat azon n indexű tagjai, amelyekre $\frac{3}{n} < 10^{-10}$, 10^{-10}-nél már közelebb vannak e-hez. Ebből az $n > 3 \cdot 10^{10}$ küszöbindexet kapjuk. Mivel a $(\frac{3}{n})$ sorozat lassan tart a nullához, ezért igen nagy küszöbszámot kaptunk.

2. Az e szám meghatározására más képlet is létezik, pl. a következő numerikus sorössszeg:

$$e = \sum_{k=0}^{\infty} \frac{1}{k!} = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots$$

Ennek a sornak a részletösszegeit is felhasználhatjuk e közelítésére:

$$e \approx \sum_{k=0}^{n} \frac{1}{k!} = 1 + \frac{1}{1!} + \frac{1}{2!} + \dots + \frac{1}{n!},$$

ahol $n \in \mathbb{N}$ rögzített. Adjunk abszolút hibakorlátot erre a közelítésre!

Megoldás:

$$\left| e - \sum_{k=0}^{\infty} \frac{1}{k!} \right| = \sum_{k=n+1}^{\infty} \frac{1}{k!} = \frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \dots = \frac{1}{(n+1)!} \left(1 + \frac{1}{n+2} + \frac{1}{(n+2)(n+3)} + \dots \right).$$

A jobb oldalt növeljük, ha a zárójelben szereplő törtek nevezőjét csökkentjük úgy, hogy minden tényező helyére $(n+1)$-et írunk:

$$\frac{1}{(n+1)!} \left(1 + \frac{1}{n+2} + \frac{1}{(n+2)(n+3)} + \dots \right) < \frac{1}{(n+1)!} \left(1 + \frac{1}{n+1} + \frac{1}{(n+1)^2} + \dots \right).$$

A jobb oldal pedig egyszerű átalakítással az

$$\frac{1}{(n+1)!} \frac{1}{1 - \frac{1}{n+1}} = \frac{1}{n!n}$$

alakra hozható. Az $\frac{1}{n!n}$ hányados már igen gyorsan tart a nullához, ezért az 1. példában tárgyaltnál hatékonyabb közelítéshez jutottunk.

3. Legyen $x > -1$ adott szám. Mekkora képlethibát okoz, ha a $\sqrt{1+x}$ kifejezést az $1 + \frac{x}{2}$ összeggel helyettesítjük?

Megoldás: Vegyük észre, hogy az $1 + \frac{x}{2}$ nem más, mint az $f(x) = \sqrt{1+x}$ függvény 0 pont körüli elsőfokú Taylor-polinomja (jel.: $T_1(f(x), 0)$). A hiba becsléséhez kézenfekvő felhasználni a Taylor-formulát, amely szerint létezik olyan ξ pont a 0 és az x között, amelyre

$$f(x) - T_1(f(x), 0) = \frac{f''(\xi)}{2!}(x-0)^2.$$

Ebből

$$|f(x) - T_1(f(x), 0)| = \left| \frac{f''(\xi)}{2!}(x-0)^2 \right| = \left| \frac{x^2}{2} \left(-\frac{1}{4} \right) \frac{1}{\sqrt{(1+\xi)^3}} \right| = \frac{x^2}{8} \frac{1}{\sqrt{(1+\xi)^3}}.$$

Mivel a ξ pontot nem ismerjük, ezért erre a kifejezésre olyan felső becslést keresünk, amely nem

tartalmazza ξ-t. Nyilván ha $x > 0$, akkor a törtet csökkentjük, ha a számlálóban ξ helyére 0-t írunk. Azaz

$$\frac{x^2}{8}\frac{1}{\sqrt{(1+\xi)^3}} < \frac{x^2}{8}\frac{1}{\sqrt{(1+0)^3}} = \frac{x^2}{8}.$$

Ha pedig $-1 < x < 0$, akkor ξ-t x-re cserélhetjük, és így az $\frac{x^2}{8}\frac{1}{\sqrt{(1+x)^3}}$ becsléshez jutunk.

1.3. Az öröklött hiba becslése

1. Az öröklött hiba első becslési módszerét vizsgáljuk először az egyváltozós esetben, azaz legyen $f : \mathbb{R} \to \mathbb{R}, a, \tilde{a} \in D(f)$. Tegyük fel, hogy f folytonos az $[a, \tilde{a}]$ intervallumon, és differenciálható az $(a, \tilde{a})$-n. Az $|f(a) - f(\tilde{a})|$ kifejezést szeretnénk felülről becsülni annak ismeretében, hogy $|a - \tilde{a}| \leq \Delta_a$. Az $f(a) - f(\tilde{a})$ különbség a Lagrange-középértéktétel értelmében pontosan kifejezhető az

$$f(a) - f(\tilde{a}) = f'(\xi)(a - \tilde{a})$$

alakban, ahol ξ az $(a, \tilde{a})$ intervallum valamely – általában ismeretlen – pontja. A két oldal abszolút értékét véve az

$$|f(a) - f(\tilde{a})| = |f'(\xi)(a - \tilde{a})| = |f'(\xi)| \cdot |a - \tilde{a}| \leq |f'(\xi)| \cdot \Delta_a.$$

becsléshez jutunk. Mivel a jobb oldalon ξ nem ismert, ezért ezt a becslést ebben formában még nem tudjuk használni. Az azonban biztos, hogy akármelyik pontja is a ξ az $(a, \tilde{a})$ intervallumnak, az $|f'(\xi)|$ nem nagyobb, mint $\sup_{[a,\tilde{a}]} |f'|$. Azaz

$$|f(a) - f(\tilde{a})| \leq \sup_{[a,\tilde{a}]} |f'| \cdot \Delta_a.$$

Ez az összefüggés könnyen általánosítható $f : \mathbb{R}^n \to \mathbb{R}$ függvények esetére. Legyen $f : \mathbb{R}^n \to \mathbb{R}, a, \tilde{a} \in D(f), f \in D([a, \tilde{a}])$ (az $[a, \tilde{a}]$ szakasz mentén folytonos, és annak belsejében differenciálható függvény). Tegyük fel, hogy az $\|a - \tilde{a}\|_2$ euklideszi norma legfeljebb Δ_a, az i-edik koordinátában fellépő abszolút hiba pedig legfeljebb Δ_{ai} (azaz $|a_i - \tilde{a}_i| \leq \Delta_{ai}, i = 1, \ldots, n$). A Lagrange-középértéktétel ebben az általánosabb esetben is úgy szól, hogy

$$f(a) - f(\tilde{a}) = f'(\xi)(a - \tilde{a}),$$

ahol most ξ az $[a, \tilde{a}]$ szakasz belsejének valamely pontja, és $f'(\xi) = \nabla f(\xi) \in \mathbb{R}^{1\times n}$. A két oldal abszolút értékét véve az

$$|f(a) - f(\tilde{a})| = |f'(\xi)(a - \tilde{a})|$$

egyenlőséghez jutunk.
A jobb oldal kétféleképpen is becsülhető.
a) Kihasználva a Cauchy–Schwarz–Bunyakovszkij-egyenlőtlenséget

$$|f'(\xi)(a-\tilde{a})| \le \|f'(\xi)\|_2 \cdot \|a-\tilde{a}\|_2 \le \sup_{[a,\tilde{a}]} \|f'\|_2 \cdot \Delta_a.$$

b) A skaláris szorzatot koordinátákkal kiírva:

$$|f'(\xi)(a-\tilde{a})| = \left|\sum_{i=1}^{n} \partial_i f(\xi)(a_i-\tilde{a}_i)\right| \le \sum_{i=1}^{n} (\sup_{[a,\tilde{a}]} |\partial_i f|)\cdot|a_i-\tilde{a}_i| \le \sum_{i=1}^{n} (\sup_{[a,\tilde{a}]} |\partial_i f|)\cdot\Delta_{ai}. \quad (1.2)$$

Ennek a becslési módszernek az a hátránya, hogy az f függvény deriváltjára ill. parciális deriváltjaira sokszor nehéz felső korlátot adni.

Példa. Tekintsük azt a kétváltozós függvényt, amely két számot összead:

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x_1,x_2) := x_1 + x_2.$$

Legyen $a = (a_1, a_2)$ tetszőleges számpár. Mekkora hibát követhetünk el az összeg képzésekor, ha a_1-et Δ_{a1}, a_2-t pedig Δ_{a2} abszolút hibával ismerjük?

Megoldás: Jelölje ismét $\tilde{a}_i$ az a_i közelítését, $i = 1, 2$. Ekkor (1.2) alapján

$$|(a_1+a_2)-(\tilde{a}_1+\tilde{a}_2)| = |f(a_1,a_2)-f(\tilde{a}_1,\tilde{a}_2)| \le (\sup_{[a,\tilde{a}]} |\partial_1 f|)\cdot\Delta_{a1} + (\sup_{[a,\tilde{a}]} |\partial_2 f|)\cdot\Delta_{a2}.$$

Mivel $\partial_1 f(x_1,x_2) = 1 = \partial_2 f(x_1,x_2)$, ezért a szuprémumok 1-gyel egyenlők. Azaz

$$|(a_1+a_2)-(\tilde{a}_1+\tilde{a}_2)| = \Delta_{a1} + \Delta_{a2}.$$

Ezt az eredményt úgy is megfogalmazhatjuk, hogy az összeg egy abszolút hibakorlátja az összeadandók abszolút hibakorlátjainak összege.

2. Az öröklött hiba másik lehetséges becslési módszerét, az ún. **lineáris becslés**t is vizsgáljuk először az egyváltozós esetben. Ha $f : \mathbb{R} \to \mathbb{R}$ n-szer folytonosan differenciálható egy olyan intervallumban, amely a-t és $\tilde{a}$-ot a belsejében tartalmazza, akkor a Taylor-formula szerint alkalmas, $\tilde{a}$ és a közötti ξ érték esetén fennáll az

$$f(a) = f(\tilde{a}) + f'(\tilde{a})(a-\tilde{a}) + \ldots + \frac{f^{(n-1)}(\tilde{a})}{(n-1)!}(a-\tilde{a})^{n-1} + \frac{f^{(n)}(\xi)}{n!}(a-\tilde{a})^n$$

egyenlőség. Feltéve, hogy f magasabbrendű deriváltjainak értéke $|f'(\tilde{a})|$-hoz képest nem túl nagy, és $|a-\tilde{a}|$ elég kicsi, az

$$f(a) - f(\tilde{a}) \approx f'(\tilde{a})(a-\tilde{a})$$

közelítő egyenlőséget nyerjük. Ez hasonlít az (1.3) pontos egyenlőséghez, azonban itt a jobb oldalon $f'(\xi)$ helyett $f'(\tilde{a})$ szerepel. Abszolút értéket véve az

$$|f(a) - f(\tilde{a})| \approx |f'(\tilde{a})| \cdot |a-\tilde{a}| \leq |f'(\tilde{a})| \cdot \Delta_a$$

közelítő becslés adódik.
Ez az út a többváltozós esetben is járható. Legyen tehát most $f : \mathbb{R}^n \to \mathbb{R}$, és $\tilde{a} \in \mathbb{R}^n$ az $a \in \mathbb{R}^n$ vektor egy közelítése. Ha f magasabbrendű deriváltjai az $\tilde{a}$ helyen nem túl nagyok, és a elég közel van $\tilde{a}$-hoz, akkor a többváltozós Taylor-formula alapján

$$|f(a) - f(\tilde{a})| \approx \sum_{i=1}^{n} |\partial_i f(\tilde{a})| \, |a_i - \tilde{a}_i| \leq \sum_{i=1}^{n} |\partial_i f(\tilde{a})| \cdot \Delta_{ai}.$$

Látható, hogy a lineáris becslés egyszerűbb, mint a Lagrange-középértéktételen alapuló (hiszen az f függvény deriváltját ill. parciális deriváltjait csak az $\tilde{a}$ helyen kell venni), de csupán közelítő érvényű.

Feladatok.

1. Vizsgáljuk meg a négy alapművelet hibáját a lineáris becslési módszerrel!

   1^o Összeadás
   Az $f(x_1, x_2) = x_1 + x_2$ függvény parciális deriváltjait már kiszámítottuk, és ez alapján

   $$|(a_1 + a_2) - (\tilde{a}_1 + \tilde{a}_2)| \leq 1 \cdot \Delta a_1 + 1 \cdot \Delta a_2 = \Delta_{a1} + \Delta_{a2},$$

   ami megegyezik korábbi eredményünkkel.

   2^o Kivonás
   Az $f(x_1, x_2) = x_1 - x_2$ függvény parciális deriváltjaira $\partial_1 f(x_1, x_2) = 1, \partial_2(x_1, x_2) = -1$, ebből

   $$|(a_1 - a_2) - (\tilde{a}_1 - \tilde{a}_2)| \leq 1 \cdot \Delta a_1 + 1 \cdot \Delta a_2 = \Delta_{a1} + \Delta_{a2},$$

   azaz az abszolút hibák a kivonásnál is összeadódnak! A különbség hibája tehát általában nem becsülhető felül a hibák különbségével.

   3^o Szorzás
   Az $f(x_1, x_2) = x_1 \cdot x_2$ függvény parciális deriváltjai: $\partial_1 f(x_1, x_2) = x_2$, $\partial_2 f(x_1, x_2) = x_1$, amiből

   $$|a_1 a_2 - \tilde{a}_1 \tilde{a}_2| \leq |\tilde{a}_2| \cdot \Delta a_1 + |\tilde{a}_1| \cdot \Delta a_2 \leq |\tilde{a}_2| \cdot \Delta_{a1} + |\tilde{a}_1| \cdot \Delta_{a2}.$$

   4^o Osztás

Az $f(x_1, x_2) = \frac{x_1}{x_2}$ függvény parciális deriváltjai: $\partial_1 f(x_1, x_2) = \frac{1}{x_1}$, $\partial_2 f(x_1, x_2) = -\frac{x_1}{x_2^2}$, amiből

$$\left|\frac{a_1}{a_2} - \frac{\tilde{a}_1}{\tilde{a}_2}\right| \leq \frac{1}{|\tilde{a}_2|}\Delta a_1 + \frac{|\tilde{a}_1|}{\tilde{a}_2^2}\Delta a_2 = \frac{|\tilde{a}_2|\Delta a_1 + |\tilde{a}_1|\Delta a_2}{\tilde{a}_2^2} \leq \frac{|\tilde{a}_2|\Delta_{a1} + |\tilde{a}_1|\Delta_{a2}}{\tilde{a}_2^2}.$$

Látható, hogy kis abszolút értékű számmal való osztásnál nagy lehet az elkövetett hiba!

2. Legyen egy henger sugara $R = 1 \pm 0,01$, magassága $h = 2 \pm 0,02$. Számítsuk ki közelítőleg a térfogatát! Határozzuk meg lineáris becsléssel, hogy mekkora lesz az eredmény abszolút hibája. A π értékét vegyük 3,14-nak, és vegyük figyelembe ennek a közelítésnek a hibáját is. (Segítségül: ekkor $\pi = 3,14 \pm 0,005$.)
Megoldás: A közelítő értékeket használva a henger térfogata közelítőleg

$$V = R^2\pi h \approx 1^2 \cdot 3,14 \cdot 2 = 6,28.$$

Jelöljük a számítás hibáját ΔV-vel. Az $R^2\pi h$ szorzatot bontsuk szét az R^2 és πh tényezőkre, és alkalmazzuk az 1. feladat szorzásra kapott eredményét!

$$\Delta V = \Delta(R^2\pi h) \leq \tilde{\pi}\tilde{h} \cdot (\Delta R^2) + \tilde{R}^2 \cdot \Delta(\pi h) \leq \tilde{\pi}\tilde{h}(\tilde{R} \cdot \Delta_R + \tilde{R} \cdot \Delta_R) + \tilde{R}^2(\tilde{h} \cdot \Delta_\pi + \tilde{\pi} \cdot \Delta_h) = 0,1984.$$

3. Számoljuk ki zsebszámológépen az $a = \sqrt{20001}$ és a $b = \sqrt{20000}$ szám hét tizedesjegyre kerekített értékét. Közelítsük az $x := a - b$ számot ezen kerekített értékek különbségével, majd becsüljük az abszolút és relatív hibáját.
Az x-et más módon is kiszámíthatjuk: $x = y := \frac{1}{a+b} = \frac{1}{\sqrt{20001}+\sqrt{20000}}$. Melyik számolási mód biztosítja a kisebb hibát?
Megoldás: $\tilde{a} = 141,4248917$, $\tilde{b} = 141,4213562$, $\tilde{x} = 3,535500000 \cdot 10^{-3}$, $\tilde{y} = 3,535489713 \cdot 10^{-3}$. A kiindulási értékek abszolút hibakorlátjai: $\Delta_a = \Delta_b = 5 \cdot 10^{-8}$. Az x abszolút és relatív hibakorlátja: $\Delta_x = \Delta_a + \Delta_b = 10^{-7}$, $\delta_x = \frac{\Delta_x}{x} = \frac{10^{-7}}{3,53549 \cdot 10^{-3}} = 0,283 \cdot 10^{-4}$.
A másik módszerrel számolva: $\Delta_{a+b} = \Delta_a + \Delta_b = 10^{-7}$, aminek felhasználásával $\Delta_y = \Delta_{\frac{1}{a+b}} = \frac{|\tilde{a}+\tilde{b}|\cdot\Delta_1 + |1|\cdot\Delta_{a+b}}{(\tilde{a}+\tilde{b})^2} = \frac{10^{-7}}{242^2} = 1,7076 \cdot 10^{-12}$. (Itt felhasználtuk, hogy $\Delta_1 = 0$, mivel az 1 pontos érték.) A relatív hibakorlátra $\delta_y = \delta_{\frac{1}{a+b}} = \frac{\Delta_y}{y} = \frac{1,7076 \cdot 10^{-12}}{3,53549 \cdot 10^{-3}} = 0,483 \cdot 10^{-9}$ adódik. A példában tehát a második számítási módszer öt nagyságrenddel kisebb abszolút és relatív hibát biztosít, mint az első.
Tanulság: közeli számok kivonását lehetőleg kerüljük.

2. fejezet

Függvényközelítés (approximáció)

Függvények közelítésére számos esetben szükségünk lehet.
Előfordulhat, hogy egy folytonos függvényt egyszerűbbel kell helyettesítenünk.
Vizsgáljuk meg például a $\sin x$ értékeinek a kiszámítását! Nagyon speciális x-ek kivételével ezek nem számíthatók ki pontosan. A

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

definíciós Taylor-sor alkalmazása egy végtelen sor összegzését jelentené, amit nem lehet elvégezni. Olyan függvénnyel kell közelítenünk a $\sin x$-et, amely ténylegesen kiszámítható, és hibája a lehető legkisebb.

De valójában milyen függvényeket lehet kiszámítani? Tekintsünk most el attól, hogy minden számítógép számrendszere diszkrét, és tegyük fel, hogy végtelen sok tizedesjeggyel is tud dolgozni. Azonban még egy ilyen gép is csak véges sok utasítást tud elvégezni. Azt mondjuk, hogy egy függvény **racionális**, ha felépíthető véges számú összeadás, kivonás, szorzás és osztás segítségével. Nyilvánvalóan függvényközelítés céljaira csak ilyen függvények alkalmasak. Ezek közül is a **polinom**okra ill. **szakaszonként polinomiális függvény**ekre szorítkozunk, ezek ugyanis a legkönnyebben kezelhető és a legkedvezőbb analitikus tulajdonságokat követő függvények.

A gyakorlatban nem egyszer előfordul, hogy egy folytonos függvényt nem ismerünk a teljes értelmezési tartományán, csak néhány $x_0, x_1, \dots, x_n$ alappontban felvett $f_0, f_1, \dots, f_n$ értékét tudjuk. Ez a helyzet például, ha az adatok diszkrét pontokban végzett mérésekből származnak. A matematikai analízis módszereit ezekre a diszkrét pontokban adott függvényekre közvetlenül nem tudjuk alkalmazni. Helyette a pontokra jól kezelhető polinomot illesztünk. Az adott pontokra való jó illeszkedésük szempontjából a közelítések három típusát különböztetjük meg.

1. Megkeressük azt a lehető legalacsonyabb fokú polinomot, amelynek értéke vala-

mennyi alappontban az adott függvényértékkel egyenlő. Ezt a függvényközelítést **interpoláció**nak nevezzük. Mint látni fogjuk, az alappontok számának növelésével általában az interpolációs polinom fokszáma is növekszik.

Gyakran eleve tudjuk, hogy a függvény lineáris vagy parabolikus, vagyis ismert, hogy hányadfokú polinom. Az adott függvényértékek pontatlansága következtében adott fokszámú polinom (pl. lineáris függvény) nem illeszthető pontosan az adott pontokra. Így meg kell keresnünk azt az adott fokszámú polinomot, amely a pontokra valamilyen értelemben a legjobban illeszkedik. Ha több ilyen polinom is létezik, további meggondolások szükségesek ahhoz, hogy eldönthessük, melyiket válasszuk.

2. A **legkisebb négyzetek módszere** azt az adott fokszámú p polinomot határozza, meg, amelyre az adott függvényértékek és polinomértékek különbségeinek négyzetösszege, vagyis a

$$\sum_{k=0}^{n}[f_k - p(x_k)]^2$$

kifejezés értéke minimális. (Vegyük észre, hogy ez az összeg jól jellemzi a $p(x)$ polinom eltérését az f_k értékektől, hiszen az $f_k - p(x_k)$ különbség négyzete mindig nemnegatív, és annál nagyobb, minél nagyobb az $|f_k - p(x_k)|$ távolság. A négyzet helyett vehetnénk egyszerűen az abszolút értéket is, ám deriválhatósága miatt a négyzetfüggvényt használjuk.)

3. A **Csebisev-féle közelítés**nél azt az adott fokszámú p polinomot választjuk, amelyre a

$$\max_{0\le k\le n} |f_k - p(x_k)|$$

mennyiség, azaz a keresett polinom és az adott függvény alappontokban vett maximális eltérése a lehető legkisebb. Ezt a közelítési módot **egyenletes közelítés**nek is nevezik. Ezzel jegyzetünkben nem foglalkozunk.

Az interpolációs polinomok tehát az alappontokban ugyanazokat az értékeket veszik fel, mint az adott függvény, míg a legkisebb négyzetek és a Csebisev-féle közelítés módszerével nyert polinomok az alappontokban az adott függvényértékeknek csak közelítését adják.

2.1. Folytonos függvény közelítése Taylor-polinommal

Legyen f az $[a, b]$ intervallumon értelmezett folytonos függvény, és x_0 az $[a, b]$ intervallum egy belső pontja. Ha f $(n + 1)$-szer folytonosan differenciálható $[a, b]$-n, és az x_0 pont környezetében szeretnénk polinommal közelíteni, akkor erre használhatjuk az f x_0 körüli

n-edfokú (vagy n-nél alacsonyabb fokú) Taylor-polinomját:

$$p(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \ldots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n.$$

A már látott Taylor-formula szerint az f függvény és az n-edfokú Taylor-polinomjának eltérése

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}, \tag{2.1}$$

ahol ξ valamely pont az x és az x_0 között.
Például egy $f : \mathbb{R} \to \mathbb{R}$ függvény $x = x_0$ ponthoz tartozó elsőfokú Taylor-polinomja ($n = 1$ eset) a

$$p(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) = f(x_0) + f'(x_0)(x - x_0)$$

polinom (ami nem más, mint az f x_0-beli érintője), hibája pedig

$$f(x) - p(x) = f(x) - f(x_0) - f'(x_0)(x - x_0) = \frac{f''(\xi)}{2!}(x - x_0)^2,$$

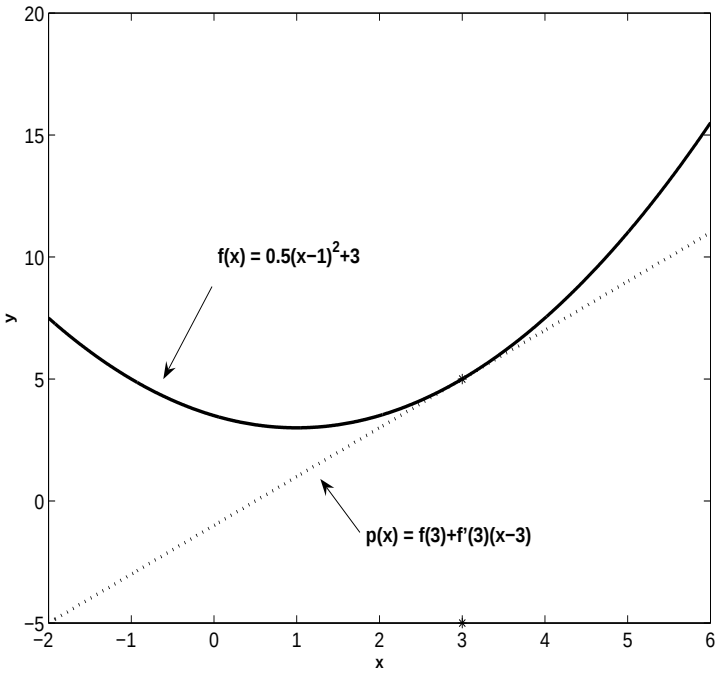
ahol $\xi \in (x_0, x)$ valamely pont.

Vegyük észre, hogy az n-edfokú Taylor-polinom értéke és első n deriváltja az $x = x_0$ pontban megegyezik az f függvényével.

A 2.1 ábrán az $f(x) = \frac{1}{2}(x-1)^2 + 3$ függvény $x_0 = 3$ ponthoz tartozó elsőfokú Taylor-polinomját láthatjuk. A Taylor-polinommal való közelítésnek a következő hátrányai vannak:

1. A (2.1) egyenlőségből látható, hogy a közelítés általában jobb azokban az x pontokban, amelyek közel vannak x_0-hoz, és rosszabb azokban, amelyek távol vannak x_0-tól. (Ez a 2.1 ábrán is megfigyelhető: a $p(x)$ elsőfokú Taylor-polinom az $x_0 = 3$ pont közelében halad a legközelebb az f függvény grafikonjához.) Azaz a közelítés nem egyenletesen jó az adott intervallumon. Ezért a Taylor-polinommal való közelítést csak az x_0-hoz kellően közeli x pontokban szokásos használni.

2. Olyan közelítést szeretnénk, amelyre egy adott intervallumban a maximális hiba tetszőlegesen kicsivé tehető. A Taylor-polinomnál ez csak akkor teljesül, ha a maradéktag nullához tart, miközben n tart a végtelenhez. Ehhez f összes deriváltjának léteznie kell, és x az x_0 pont körül Taylor-sorba fejthető kell, hogy legyen, ami nem mindig teljesül.



2.1. ábra. Az $f(x) = \frac{1}{2}(x-1)^2 + 3$ függvény $x_0 = 3$ ponthoz tartozó elsőfokú Taylor-polinomja.

2.2. Az interpolációs polinom

Az n-edfokú Taylor-polinom egy adott pontban megegyezik az adott f függvénnyel az első n deriváltban. Egy interpolációs polinom $n+1$ pontban egyezik meg f-fel, de a deriváltakkal esetleg sehol sem.

Legyen f egy I intervallumon értelmezve, és legyen $x_0, x_1, \ldots, x_n$ az I intervallum $n+1$ különböző pontja. Jelölje f_k az $f(x_k)$ függvényértéket, $k = 0, 1, \ldots, n$. Olyan p polinomot keresünk, amelynek fokszáma legfeljebb n, és teljesül, hogy $p(x_k) = f_k,\ k = 0, 1, \ldots, n$. Az x_k pontokat **interpolációs alappont**oknak, az f_k értékeket **interpolált érték**eknek, p-t pedig **interpolációs polinom**nak nevezzük. Az interpolációs polinom létezését és egyértelműségét biztosítja az alábbi tétel.

2.2.1 Tétel. *Pontosan egy olyan polinom létezik, amelynek fokszáma legfeljebb n, és az $n+1$ különböző pontban az f_k értékeket veszi fel.*

Biz.: a) Először a létezést bizonyítjuk a képlet bemutatásával.
Ezt először vizsgáljuk egy speciális esetben. Tegyük fel, hogy az összes interpolált érték nulla, egy kivételével, amely 1-gyel egyenlő. Legyen ez például az m-edik, azaz

$$f_m := 1.$$

Első lépésben olyan, n-nél nem nagyobb fokszámú l_m polinomot keresünk, amelyre

$$l_m(x_k) = \begin{cases} 1, & k = m \\ 0, & k \neq m. \end{cases}$$

Mivel az $x_0, x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_n$ pontokban el kell tűnnie, tartalmaznia kell az $x - x_k$ $(k \neq m)$ gyöktényezőket. Ez azt jelenti, hogy a keresett polinom

$$l_m(x) = c_m \prod_{\substack{k=0 \\ k \neq m}}^{n} (x - x_k)$$

alakú, ahol c_m állandó. Ezt az állandót az $l_m(x_m) = 1$ feltételből határozhatjuk meg, amiből

$$c_m = \frac{1}{\prod\limits_{\substack{k=0 \\ k \neq m}}^{n} (x_m - x_k)}.$$

Így az

$$l_m(x) = \prod_{\substack{k=0 \\ k \neq m}}^{n} \frac{x - x_k}{x_m - x_k}.$$

alakhoz jutunk. Világos, hogy ez a polinom pontosan n-edfokú. Ebből következik, hogy tetszőleges f_k-ra a

$$p(x) := \sum_{m=0}^{n} l_m(x) f_m$$

függvény legfeljebb n-edfokú polinom. Már csak azt kell belátni, hogy p az x_k pontokban az f_k értékeket veszi fel. Az x_k-t behelyettesítve $p(x)$ képletébe

$$p(x_k) = \sum_{m=0}^{n} l_m(x_k) f_m.$$

Ebben az összegben csak az $m = k$ indexű tag nem nulla, és az éppen f_k-val egyenlő. Tehát

$$p(x_k) = f_k, \quad k = 0, 1, \ldots, n.$$

b) Az egyértelműséget indirekt módon bizonyítjuk. Tegyük fel, hogy két polinom is rendelkezik a kívánt tulajdonságokkal, azaz

$$p(x_k) = f_k \ \text{ és } \ q(x_k) = f_k, \ k = 0, 1, \ldots, n.$$

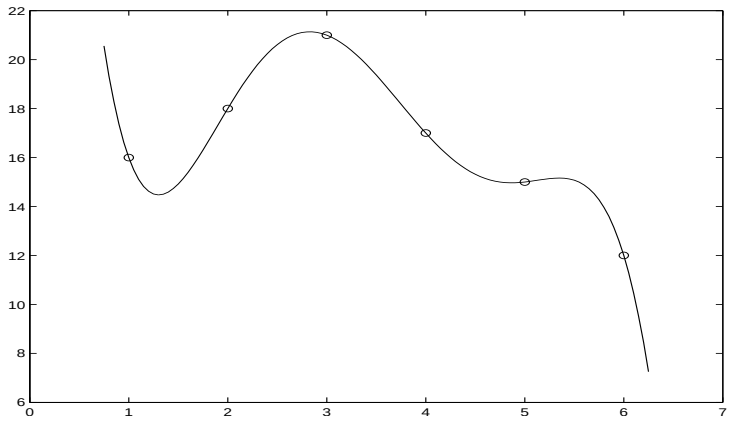
Jelölje d a $p-q$ különbséget. Ekkor

$$d(x_k) = p(x_k) - q(x_k) = f_k - f_k = 0, \ k = 0, 1, \ldots, n.$$

Így d egy n-nél nem magasabb fokszámú polinom, amelynek $n+1$ különböző zérushelye van. Ez csak úgy lehetséges, ha d azonosan nulla, azaz $p = q$.

Az l_m polinomokat **Lagrange-féle interpolációs alappolinomok**nak, p-t pedig **Lagrange-féle interpolációs polinom**nak nevezzük. Megjegyezzük, hogy helyesebb volna az **interpolációs polinom Lagrange-féle reprezentációja** elnevezés, mert az interpolációs polinom egyértelműen meghatározott, és a Lagrange-alappolinomokkal kifejezett alak csak az egyik lehetséges felírási módja.

Illusztrációként az ábrán megtekinthetünk egy hat alappontra illeszkedő interpolációs polinomot.



2.2. ábra. Hat alappontra illeszkedő interpolációs polinom.

Feladat.
Legyenek egy függvény értékei az $x_0 = 0, x_1 = 1, x_2 = 2$ és $x_3 = 3$ alappontokban rendre $f_0 = -5, f_1 = -6, f_2 = -1$ és $f_3 = 16$. Írjuk fel az ezen pontokra illeszkedő interpolációs polinom Lagrange-féle alakját.
Megoldás: Az alappolinomok:

$$\begin{aligned}
l_0(x) &= \prod_{k=0, k\neq 0}^{3} \frac{x - x_k}{x_0 - x_k} = \frac{x - x_1}{x_0 - x_1} \cdot \frac{x - x_2}{x_0 - x_2} \cdot \frac{x - x_3}{x_0 - x_3} = \frac{(x-1)(x-2)(x-3)}{(-1)\cdot(-2)\cdot(-3)} \\
l_1(x) &= \frac{x(x-2)(x-3)}{(1)\cdot(-1)\cdot(-2)} \\
l_2(x) &= \frac{x(x-1)(x-3)}{(2)\cdot(1)\cdot(-1)} \\
l_3(x) &= \frac{x(x-1)(x-2)}{(3)\cdot(2)\cdot(1)}
\end{aligned}$$

Tehát az interpolációs polinom:

$$\begin{aligned} p(x) &= \frac{(x-1)(x-2)(x-3)}{(-6)}(-5) + \frac{x(x-2)(x-3)}{2}(-6) + \\ &+ \frac{x(x-1)(x-3)}{(-2)}(-1) + \frac{x(x-1)(x-2)}{6}(16). \end{aligned}$$

2.2.1. Newton-féle alak

A Lagrange-interpolációnak az a nagy hátránya, hogy újabb alappontot felvéve a számításokat elölről kell kezdenünk, nem tudjuk az eddigi eredményeket közvetlenül felhasználni. A továbbiakban az interpolációs polinom olyan alakjával foglalkozunk, amelynél ez a hátrány kiküszöbölhető.

Legyen $f : \mathbb{R} \to \mathbb{R}$, és $x_0, x_1, x_2, \ldots \in D(f)$ páronként különböző alappontok.

2.2.2 Definíció. *(Osztott differenciák)*
Nulladrendű osztott differenciáknak nevezzük az

$$f[x_i] := f(x_i)$$

függvényértékeket.
Elsőrendű osztott differenciáknak nevezzük az

$$f[x_i, x_{i+1}] := \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i}$$

hányadosokat.
Továbbmenve, ha $k \in \mathbb{N}^+$, akkor a k-adrendű osztott differenciákat az

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] := \frac{f[x_{i+1}, \ldots, x_{i+k}] - f[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}$$

rekurzív képlettel definiáljuk.

Kiszámításukhoz érdemes ún. differenciatáblázatot készíteni. Írjuk fel egymás alá az alappontokat, és mindegyik mellé a hozzá tartozó nulladrendű osztott differenciát, majd a további osztott differenciákat a 2.1 szerinti elrendezésben.
A következő tétel az interpolációs polinomot az osztott differenciákkal fejezi ki.

2.2.3 Tétel. *(Az interpolációs polinom Newton-féle alakja)*
A p interpolációs polinom előáll a

$$\begin{aligned} p(x) = f[x_0] + f[x_0, x_1] \cdot (x - x_0) + f[x_0, x_1, x_2] \cdot (x - x_0)(x - x_1) + \ldots \\ \ldots + f[x_0, \ldots, x_n] \cdot (x - x_0) \cdots (x - x_{n-1}) \end{aligned} \tag{2.2}$$

$$
\begin{array}{llllll}
x_0 & \underline{f[x_0]} & & & & \\
 & & \underline{f[x_0,x_1]} & & & \\
x_1 & f[x_1] & & \underline{f[x_0,x_1,x_2]} & & \\
 & & f[x_1,x_2] & & \underline{f[x_0,x_1,x_2,x_3]} & \\
x_2 & f[x_2] & & f[x_1,x_2,x_3] & & \\
 & & f[x_2,x_3] & & & \\
x_3 & f[x_3] & & & &
\end{array}
$$

2.1. táblázat. A differenciatáblázat négy alappont esetén.

alakban.

Az interpolációs polinom felírásához tehát a differenciatáblázatban aláhúzott elemek kellenek. Látható, hogy egy újabb alappont hozzávételekor csak egy új tagot kell hozzáadni a korábbi alakhoz, és ez a tag a már elkészített differenciatáblázat kibővítésével könnyen kiszámítható.

Feladatok. 1. Írjuk fel az interpolációs polinom Newton-féle alakját, ha $x_0 = -1$, $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, és $f_0 = 1$, $f_1 = -1$, $f_2 = -1$, $f_3 = 1$.

Megoldás: Készítsük el a differenciatáblázatot:

$$
\begin{array}{llllll}
-1 & \underline{1} & & & & \\
 & & \underline{-2} & & & \\
0 & -1 & & \underline{1} & & \\
 & & 0 & & \underline{0} & \\
1 & -1 & & 1 & & \\
 & & 2 & & & \\
2 & 1 & & & &
\end{array}
$$

Ebből $p(x) = 1 + (-2)(x+1) + 1(x+1)x$.

2. Vegyünk fel egy újabb alappontot: $x_4 = 3$, $f_4 = 2$, és írjuk fel az új $p(x)$-et!

Megoldás: A kibővített differenciatáblázat alapján $p(x) = 1 + (-2)(x+1) + 1(x+1)x + 0 + \frac{1}{6}(x+1)x(x-1)(x-2)$.

$$
\begin{array}{llllll}
-1 & \underline{1} & & & & \\
 & & \underline{-2} & & & \\
0 & -1 & & \underline{1} & & \\
 & & 0 & & \underline{0} & \\
1 & -1 & & 1 & & \underline{\frac{1}{6}} \\
 & & 2 & & -\frac{1}{2} & \\
2 & 1 & & -\frac{1}{2} & & \\
 & & 1 & & & \\
3 & 2 & & & &
\end{array}
$$

2.2.2. Az interpolációs polinom hibája

2.2.4 Tétel. *Legyen $f : I \to \mathbb{R}$ $n+1$-szer folytonosan differenciálható függvény, p az a legfeljebb n-edfokú polinom, amely az I intervallum $n+1$ különböző x_k $(k = 0, 1, \ldots, n)$ pontjában interpolálja f-et, és $x \in I$ tetszőleges pont. Ekkor az x-et és az összes x_k pontot tartalmazó legszűkebb intervallum tartalmaz a belsejében egy olyan ξ pontot, amelyre*

$$f(x) - p(x) = \frac{1}{(n+1)!}\omega_n(x)f^{(n+1)}(\xi), \tag{2.3}$$

ahol

$$\omega_n(x) := (x - x_0)(x - x_1)\cdots(x - x_n). \tag{2.4}$$

Biz.:

- $x = x_k$-ra nyilvánvaló az állítás, hiszen ekkor (2.3) mindkét oldala tetszőleges ξ-re nulla.

- Tegyük fel, hogy $x \neq x_k$ minden k-ra, és rögzítsük x-et. Tekintsük a

$$g(t) := f(t) - p(t) - c\omega_n(t) \tag{2.5}$$

segédfüggvényt, ahol c egyelőre tetszőleges állandó. Nyilván

$$g(x_k) = f_k - f_k - 0 = 0, \quad k = 0, 1, \ldots, n.$$

Válasszuk meg c-t úgy, hogy a g függvény $t = x$ esetén is tűnjön el, azaz $f(x) - p(x) - c\omega_n(x) = 0$ legyen. Ebből

$$c = \frac{f(x) - p(x)}{\omega_n(x)}.$$

Ezen c-vel g-nek legalább $n+2$ gyöke lesz: $x_0, x_1, \ldots, x_n$ és x, és valamennyi az I intervallumra esik. Rolle tétele szerint g'-nek legalább $n+1$ gyöke van, g''-nek legalább n és így tovább, a $g^{(n+1)}$-nek legalább egy. Ezek abban a legszűkebb intervallumban vannak, amely tartalmazza az $x_0, x_1, \ldots, x_n$ és x pontokat. Legyen ξ a $g^{(n+1)}$ függvény egy ilyen gyöke. A (2.5) egyenlőséget $n+1$-szer deriválva

$$g^{(n+1)}(t) = f^{(n+1)}(t) - c(n+1)!$$

A $t = \xi$ helyettesítéssel

$$f^{(n+1)}(\xi) = c(n+1)!$$

Figyelembe véve c megválasztását, éppen a (2.3) egyenlőséget kapjuk.

Alkalmazzuk a tételt $n = 0$-ra (egy darab alappont)! Ekkor p nulladfokú polinom, azaz konstans függvény, értéke minden pontban $f(x_0)$. Ekkor a (2.3) egyenlőség speciálisan

$$f(x) - f(x_0) = (x - x_0)f'(\xi), \text{ ahol } \xi \text{ } x \text{ és } x_0 \text{ közötti,}$$

és ez az analízisből ismert Lagrange-középértéktétel.
Tekintsük most az $n = 1$ esetet (két alappont). Ekkor lineáris interpolációról beszélünk. A (2.3) egyenlőség:

$$f(x) - p(x) = \frac{(x - x_0)(x - x_1)}{2} f''(\xi)$$

Megjegyzés: A 2.2 ábrán látható az interpolációs polinom használatának fő hátránya. Az alappontok között – az ábrán ez különösen az első és az utolsó részintervallumon szembetűnő – a függvény túlságosan ingadozik. Minél nagyobb fokszámú az interpolációs polinom (azaz minél több alappont van), általában annál nagyobbak az illesztett görbe ingadozásai. (Ennek oka az, hogy egy n-edfokú polinom deriváltja $(n-1)$-edfokú polinom, amelynek akár $n - 1$ zérushelye is lehet. Ahol pedig a derivált nulla, ott az interpolációs polinomnak lokális szélsőértéke vagy inflexiós pontja van.) Ezért az interpolációnak ezt a legalapvetőbb módját a gyakorlatban ritkán használják görbék illesztésére.

Példák

1. Legyen $I \subset \mathbb{R}$ intervallum, $f \in C^2(I)$. Interpoláljunk lineárisan az $x_0 \in I$ és $x_1 \in I$ pontok között. Mekkora az elkövethető legnagyobb hiba az (x_0, x_1) intervallumon, ha $\sup_{[x_0,x_1]} |f''| = M$?
   Megoldás: Az $|\frac{(x-x_0)(x-x_1)}{2}|$ kifejezés a maximumát az $[x_0, x_1]$ intervallum közepén veszi fel, értéke $\frac{(x_1-x_0)^2}{8}$. A keresett hibakorlát tehát

$$|f(x) - p(x)| \leq \frac{(x_1 - x_0)^2}{8} M.$$

2. Mekkora lépésközzel készítsük el a sin függvény függvénytáblázatát, hogy utána két szomszédos alappont között lineárisan interpolálva a hiba ne legyen nagyobb, mint 10^{-4}?
   Megoldás: Az $f(x) = \sin x$ függvény második deriváltja $f''(x) = -\sin x$, így $|f''(x)| \leq 1$ minden x-re. Vagyis az $M = 1$ korlát mellett alkalmazhatjuk az 1. feladat eredményét. Ha két szomszédos alappont távolsága h, akkor bármely x pontban fennáll az

$$|f(x) - p(x)| \leq \frac{h^2}{8} \cdot 1$$

   becslés. Ezért ha $\frac{h^2}{8} < 10^{-4}$, azaz $h < \sqrt{8} \cdot 10^{-2}$, akkor megfelelő lesz a függvénytáblázat.

2.3. A legkisebb négyzetek módszere

Jelölje továbbra is $x_0, x_1, \ldots, x_n$ az alappontokat, és $f_0, f_1, \ldots, f_n$ a hozzájuk tartozó függvényértékeket. Ahogy a bevezetőben említettük, a legkisebb négyzetek módszere azt az adott fokszámú p polinomot határozza meg, amelyre a

$$\sum_{k=0}^{n}[f_k - p(x_k)]^2$$

kifejezés értéke minimális.
Nézzük először azt az esetet, amikor a pontokra $p(x) = c_0 + c_1 x$ alakú elsőfokú polinomot, azaz egyenest illesztünk. Mely c_0 és c_1 esetén lesz minimális a fenti összeg, vagyis a

$$\sum_{k=0}^{n}[f_k - c_0 - c_1 x_k]^2$$

kifejezés? Itt x_k és f_k, $k = 0, 1, \ldots, n$ adottak, és a (c_0, c_1) számpártól függő

$$F : (c_0, c_1) \mapsto \sum_{k=0}^{n}[f_k - c_0 - c_1 x_k]^2$$

$\mathbb{R}^2 \to \mathbb{R}$ függvény minimumhelyét keressük. Az analízisből ismeretes, hogy egy ilyen függvénynek ott lehet minimumhelye, ahol a deriváltja (gradiense) nulla. Deriváljuk tehát c_0 és c_1 szerint ezt a függvényt:

$$\begin{aligned}\frac{\partial F}{\partial c_0} &= \sum_{k=0}^{n} 2[f_k - c_0 - c_1 x_k] \cdot (-1)\\ \frac{\partial F}{\partial c_1} &= \sum_{k=0}^{n} 2[f_k - c_0 - c_1 x_k] \cdot (-x_k).\end{aligned}$$

Ezeket nullával egyenlővé téve – egyszerűsítések után – a

$$\begin{aligned}\sum_{k=0}^{n}[f_k - c_0 - c_1 x_k] &= 0\\ \sum_{k=0}^{n}[f_k x_k - c_0 x_k - c_1 x_k^2] &= 0.\end{aligned}$$

lineáris algebrai egyenletrendszert kapjuk. Ezt kell megoldanunk a c_0, c_1 ismeretlenekre. A megoldás valóban minimumhelye lesz az F függvénynek, ugyanis az F második de-

riváltjának mátrixa

$$F''(c_0, c_1) = \begin{bmatrix} 2(n+1) & 2\sum_{k=0}^{n} x_k \\ 2\sum_{k=0}^{n} x_k & 2\sum_{k=0}^{n} x_k^2 \end{bmatrix},$$

amelynek mind a bal felső eleme, mind pedig $\det F'' = 4(n+1)\sum_{k=0}^{n} x_k^2 - 4(\sum_{k=0}^{n} x_k)^2$ determinánsa pozitív. (Az utóbbi állítás abból következik, hogy minden $n \in \mathbb{N}, n \geq 1$ számra fennáll az $(n+1)(x_0^2 + x_1^2 + \ldots + x_n^2) > (x_0 + x_1 + \ldots + x_n)^2$ egyenlőtlenség.)

Általános, esetben, ha N-edfokú polinomot akarunk a pontokra illeszteni, akkor $N+1$-ismeretlenes lineáris algebrai egyenletrendszert kapunk. Az ilyen egyenletrendszerek megoldásával bővebben foglalkozunk a 7. fejezetben.

Példa

Keressük meg a $x_0 = 0$, $x_1 = 1$, $x_2 = 2$ alappontokhoz és az $f_0 = 1$, $f_1 = 2$, $f_2 = 3/2$ függvényértékekhez tartozó négyzetesen legjobban közelítő elsőfokú függvényt.

Megoldás: Esetünkben

$$F(c_0, c_1) = (1 - c_0)^2 + (2 - c_0 - c_1)^2 + \left(\frac{3}{2} - c_0 - 2c_1\right)^2.$$

Ebből a

$$\frac{\partial F}{\partial c_0}(c_0, c_1) = 6c_0 + 6c_1 - 9 = 0,$$

és

$$\frac{\partial F}{\partial c_1}(c_0, c_1) = 6c_0 + 10c_1 - 10 = 0$$

egyenleteket kapjuk. A kétismeretlenes rendszer megoldása: $c_0 = \frac{5}{4}$, $c_1 = \frac{1}{4}$, így a keresett függvény a $p(x) = \frac{5}{4} + \frac{1}{4}x$.

3. fejezet

Közelítő integrálás

Legyen $f : [a, b] \to \mathbb{R}$ egy nemnegatív függvény. Sok esetben kíváncsiak vagyunk a függvény grafikon alatti területére. Az analízisből ismeretes, hogy ha f Riemann-integrálható az $[a, b]$ intervallumon, akkor ezt a területet az $\int_a^b f(x)\mathrm{d}x$ határozott integrál adja meg. Továbbá, amennyiben f-nek létezik F primitív függvénye, akkor a Newton–Leibniz-szabály értelmében

$$\int_a^b f(x)\mathrm{d}x = F(b) - F(a).$$

A primitív függvényt azonban nem mindig tudjuk meghatározni. Ezért van szükség olyan módszerekre, amelyekkel az $\int_a^b f(x)\mathrm{d}x$ integrált legalább közelítőleg ki tudjuk számítani.

A közelítő integrálás formuláit kvadratúraformuláknak is szokásos nevezni. Ezek a formulák a függvény grafikon alatti területét valamilyen egyszerű alakzat területével közelítik, ami a függvény néhány pontban felvett értékéből kiszámítható. Ennek különféle lehetőségeit tekintjük át a 3.1. alfejezetben. (A „kvadratúra” szó arra az elemi területszámítási módszerre utal, amikor a függvény grafikonját négyzethálós papírra visszük, és összeadjuk a görbe alatti kis négyzetek területét.)

3.1. Az alapvető kvadratúraformulák

Jelölje $h := b - a$ az intervallum hosszát. Ismerkedjünk meg először két alapvető kvadratúraformulával.

1. **Középponti szabály:** annak a téglalapnak a területével közelítjük az integrált, amelynek alapja h, magassága pedig az intervallum $c = \frac{a+b}{2}$ felezőpontjában felvett függvényérték, azaz

$$k(f) := h \cdot f(c).$$

2. **Trapézszabály:** annak a trapéznak a területével közelítjük az integrált, amelynek alapja h, oldalainak hosszai pedig az intervallum két végpontjában felvett függvényértékek, azaz
$$t(f) := h \cdot \frac{f(a) + f(b)}{2}.$$

Egy kvadratúraformulát jól jellemez az, hogy hogyan viselkedik különböző fokszámú polinomokra. Egy kvadratúraformula rendje annak a legalacsonyabb fokú polinomnak a fokszáma, amelyre a formula már nem adja meg pontosan az integrált. Pl. a középponti és a trapézszabály minden f elsőfokú polinomra pontos, de a másodfokú polinomokra már nem, ezért másodrendű kvadratúraformuláknak nevezzük őket. Általában ha egy p-edrendű kvadratúraformulát megfelelően sima függvényre alkalmazunk egyre csökkenő hosszúságú intervallumokon, akkor a területszámítás hibája h p-edik hatványával arányos.

Feladat. Hányadrendűek az alábbi kvadratúraformulák?

a.) $a(f) := h \cdot f(a)$ (lásd a 3.1. ábra jobb alsó grafikonját)

b.) $b(f) := h \cdot f(b)$

Megoldás: Elsőrendűek, mert csak a konstans függvényekre pontosak, más elsőfokú függvényre nem.

Alkalmazzuk a középponti és a trapézszabályt az $f(x) = x^2$ függvényre a $[0, 1]$ intervallumon!
Az integrál pontos értéke
$$\int_0^1 x^2 \mathrm{d}x = \left.\frac{x^3}{3}\right|_0^1 = \frac{1}{3}.$$
A középponti szabály eredménye
$$k(f) = 1 \cdot \left(\frac{1}{2}\right)^2 = \frac{1}{4},$$
a trapézszabályé pedig
$$t(f) = 1 \cdot \frac{0+1}{2} = \frac{1}{2}.$$
Így $k(f)$ hibája $\frac{1}{3} - \frac{1}{4} = \frac{1}{12}$, $t(f)$ hibája pedig $\frac{1}{3} - \frac{1}{2} = -\frac{1}{6}$. Látjuk, hogy a hibák ellentétes előjelűek, és a középponti szabály kétszer olyan pontos, mint a trapézszabály. Ez a megállapítás nagyjából helytálló más sima függvényekre is. Ezt az észrevételt felhasználhatjuk arra, hogy még pontosabb formulát gyártsunk.
Nyilván, hogyha $t(f)$ hibája pontosan -2-szerese $k(f)$ hibájának, akkor az
$$x - t(f) = -2(x - k(f))$$

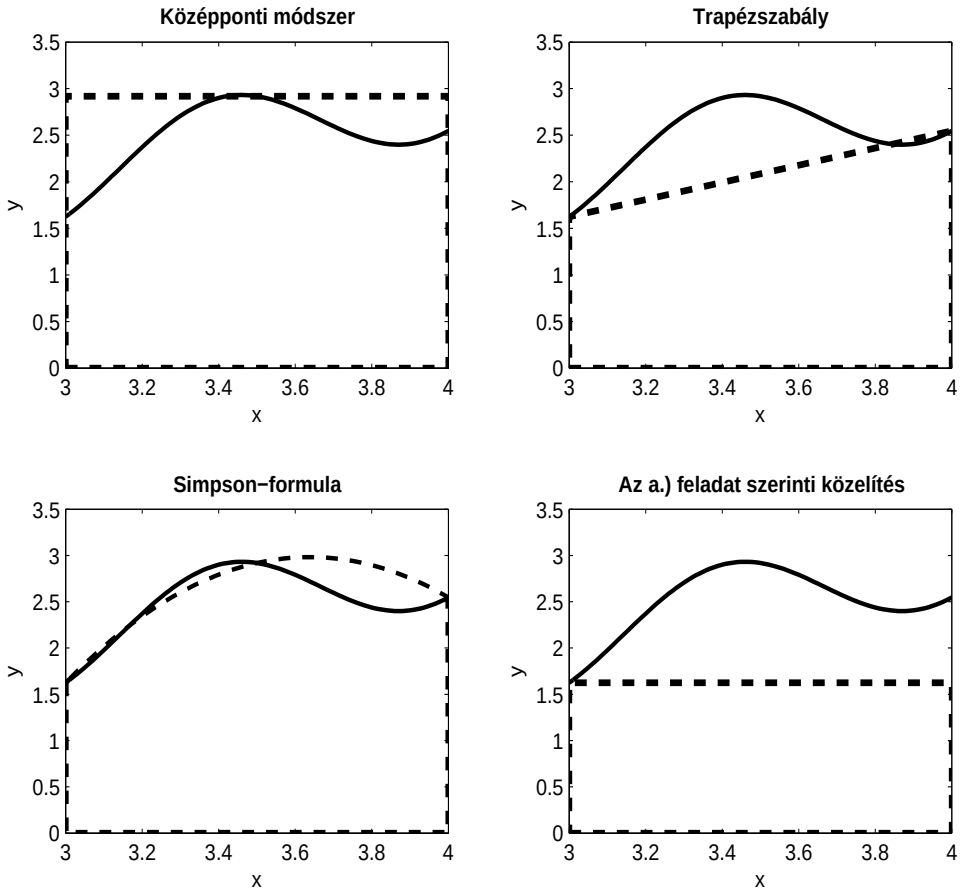
egyenlet x megoldása az integrál pontos értékét adja. Ebből x-et kifejezve az

$$x = \frac{2}{3}k(f) + \frac{1}{3}t(f)$$

képletet kapjuk, amelyet kezelhetünk újabb területszámítási formulaként. Az így nyert

$$s(f) := \frac{2}{3}k(f) + \frac{1}{3}t(f) = \frac{h}{6}(f(a) + 4f(c) + f(b))$$

formulát **Simpson-formulá**nak vagy **parabolaformulá**nak nevezzük.
A különböző kvadratúraformulák szemléltetésére szolgál a 3.1. ábra.



3.1. ábra. Kvadratúraformulák. A folytonos vonal az $f(x) = \frac{1}{2}\sin 6x + x - 1$ függvény grafikonja. A függvény grafikon alatti területét a $[3, 4]$ intervallumon a szaggatott vonallal bekeretezett tartományok területével közelítjük.

Kvadratúraformulákat másképpen is származtathatunk, pl. úgy, hogy f-et helyettesítjük

valamely interpolációs polinomjával, és ezen interpolációs polinom grafikon alatti területét számítjuk ki. Ezzel foglalkozik a következő alfejezet, amelyben már részletesebben megvizsgáljuk a kvadratúraformulák hibáját. Megtudjuk azt is, hogy a Simpson-formulát miért nevezik másképpen parabolaformulának.

3.2. Interpolációs típusú kvadratúraformula

Legyen $[a,b] \subset \mathbb{R}$ zárt intervallum, és $x_0, x_1, \ldots, x_n \in [a,b]$ tetszőleges alappontok. Az $f : [a,b] \to \mathbb{R}$ függvény integrálját közelítsük az őt interpoláló polinom integráljával! Láttuk, hogy amennyiben $f \in C^{n+1}[a,b]$,

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(x), \tag{3.1}$$

ahol $p(x) = \sum_{k=0}^{n} l_k(x) f(x_k)$ az interpolációs polinom, és a jobb oldali tag az interpoláció hibája, ahol ξ az x-et és az alappontokat tartalmazó legszűkebb intervallum valamely pontja. Integráljuk a (3.1) egyenletet az $[a,b]$ intervallumon:

$$\int_a^b f(x)\mathrm{d}x - \int_a^b \left(\sum_{m=0}^{n} l_m(x) f(x_m) \right) \mathrm{d}x = \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(x)\mathrm{d}x. \tag{3.2}$$

A bal oldali második tag az integrál közelítése, amely másképpen a

$$\sum_{m=0}^{n} \left(\int_a^b l_m(x)\mathrm{d}x \right) f(x_m) \tag{3.3}$$

alakba írható, a jobb oldali tag pedig a közelítés képlethibája.

3.2.1 Definíció. *A 3.3 kifejezést* **interpolációs típusú kvadratúraformulá***nak hívjuk.*

3.2.2 Definíció. *Általánosan a* $\sum_{m=0}^{n} A_m f(x_m)$ *alakú kifejezést* **lineáris kvadratúraformulá***nak, az* A_m *számokat pedig a kvadratúraformula együtthatóinak nevezzük.*

Az interpolációs típusú kvadratúraformula tehát olyan lineáris kvadratúraformula, amelynek együtthatóira speciálisan

$$A_m := \int_a^b l_m(x)\mathrm{d}x.$$

Sokféleképpen lehet származtatni lineáris kvadratúraformulákat. Az interpolációs típusú azonban rendelkezik a következő előnyös tulajdonsággal.

3.2.3 Tétel. *Egy lineáris kvadratúraformula akkor és csak akkor pontos minden legfeljebb* n*-edfokú polinomra, amikor interpolációs típusú.*

Biz.: ($\Leftarrow$) Visszafelé az állítás nyilvánvalóan következik abból, hogy egy legfeljebb n-edfokú polinom interpolációs polinomja önmaga.
($\Rightarrow$) Mivel az $l_j, j = 0, 1, \ldots, n$ Lagrange-féle alappolinomok n-edfokú polinomok, ezért mindegyikükre pontos a kvadratúraformula. Azaz

$$\int_a^b l_j(x)\mathrm{d}x = \sum_{m=0}^{n} A_m l_j(x_m) = A_j.$$

Az utolsó lépésben felhasználtuk, hogy $l_j(x_m) = 1$, ha $m = j$, és 0, ha $m \neq j$.

3.2.1. Newton–Cotes-formulák

3.2.4 Definíció. *A (3.3) interpolációs típusú kvadratúraformulát Newton–Cotes-formulának nevezzük, ha $[a, b]$ felosztása egyenlőközű, azaz $x_i := a + i\frac{b-a}{n}, (i = 0, 1, \ldots, n)$.*

Speciális esetek

- $n = 1$, azaz két alappont van: $x_0 = a$ és $x_1 = b$. Ekkor az interpolációs polinom lineáris függvény. Számítsuk ki a formula együtthatóit!

$$A_0 = \int_a^b l_0(x)\mathrm{d}x = \int_a^b \frac{x - x_1}{x_0 - x_1}\mathrm{d}x = \int_a^b \frac{x - b}{a - b}\mathrm{d}x = \frac{1}{a-b}\left[\frac{x^2}{2} - bx\right]_a^b =$$
$$= \frac{1}{a-b}\left(\frac{b^2}{2} - b^2 - \frac{a^2}{2} + ab\right) = \frac{1}{a-b}\frac{-b^2 - a^2 + 2ab}{2} = -\frac{(a-b)^2}{2(a-b)} = \frac{b-a}{2} = \frac{h}{2}.$$

  Hasonlóan,

$$A_1 = \int_a^b l_1(x)\mathrm{d}x = \int_a^b \frac{x - x_0}{x_1 - x_0}\mathrm{d}x = \frac{b-a}{2} = \frac{h}{2}.$$

  Így a jól ismert

$$t(f) := \frac{h}{2}(f(a) + f(b)) \tag{3.4}$$

  trapézformulát kapjuk. A képlethiba (3.2)-ből (az integrálás részleteit mellőzve)

$$\int_a^b f - t(f) = \int_a^b \frac{f''(\xi)}{2!}(x-a)(x-b)\mathrm{d}x = -\frac{h^3}{12}f''(\kappa)$$

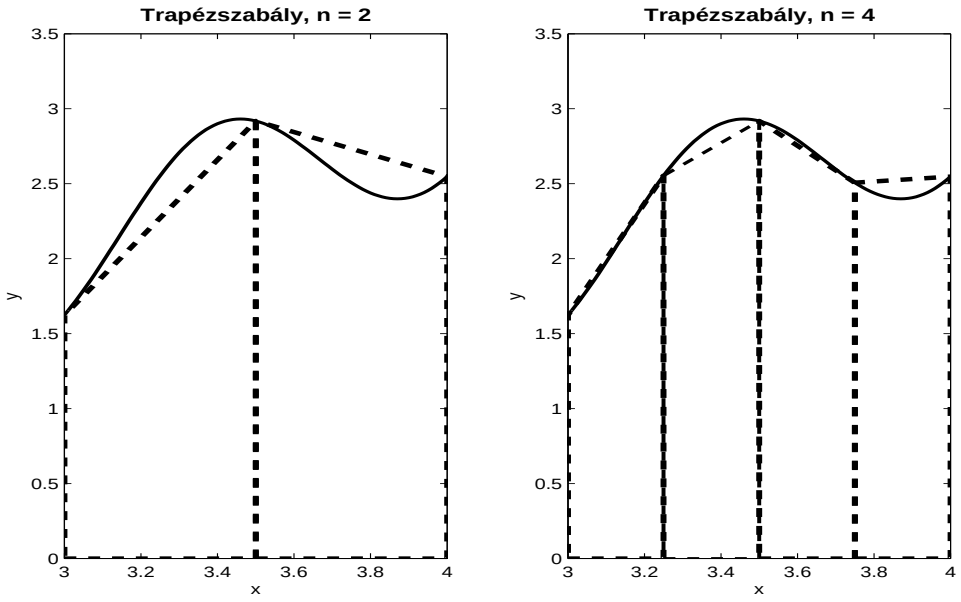
  ahol $\kappa \in [a, b]$ valamely pont, és $f \in C^2[a, b]$.

  Pontosabb integrálközelítő formulát nyerünk, ha szakaszonként végzünk lineáris interpolációt, és minden egyes szakaszon alkalmazzuk a trapézformulát, majd az ered-

ményeket összeadjuk. Ehhez jelölje Δh a $(b-a)/n$ hányadost. Ekkor

$$\int_a^b f = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f \approx \sum_{i=1}^{n} \frac{\Delta h}{2}(f(x_{i-1})+f(x_i)) = \frac{\Delta h}{2}(f(x_0)+2\sum_{i=1}^{n-1} f(x_i)+f(x_n)) =: t_n(f).$$

Ez, az ún. **összetett trapézformula** szintén lineáris kvadratúraformula, de nem interpolációs típusú. Az összetett trapézformulát szemlélteti $n = 2$ és $n = 4$ esetén a 3.2. ábra.



3.2. ábra. Az $f(x) = \frac{1}{2}\sin 6x + x - 1$ függvény grafikon alatti területének kiszámítása a $[3, 4]$ intervallumon az összetett trapézszabállyal $n = 2$ és $n = 4$ részintervallum esetén. Az integrált a szaggatott vonallal határolt tartományok területének összege annál jobban megközelíti, minél több részre daraboljuk az intervallumot.

Az összetett trapézformula képlethibája $f \in C^2[a, b]$ függvényekre

$$\int_a^b f - t_n(f) = \sum_{i=1}^{n} \frac{-\Delta h^3}{12} \cdot f''(\kappa_i) = -\frac{(b-a)^3}{12n^2} \cdot \frac{1}{n} \sum_{i=1}^{n} f''(\kappa_i), \tag{3.5}$$

ahol $\kappa_i \in [x_{i-1}, x_i]$. Az $\frac{1}{n}\sum_{i=1}^{n} f''(\kappa_i)$ kifejezés n darab függvényérték számtani közepe. Tegyük fel, hogy $f \in C^2[a, b]$. Ekkor f'' valamely $\eta \in [a, b]$ helyen felveszi ezt a számtani közepet. Ebből a képlethiba a

$$-\frac{(b-a)^3}{12n^2} f''(\eta)$$

alakba írható. Mivel f'' folytonos az $[a, b]$-n, így korlátos is, azaz $|f''| \leq K$. Így a

hiba abszolút értékben felülről becsülhető a következőképpen:

$$\left| -\frac{(b-a)^3}{12n^2} f''(\eta) \right| \leq \frac{(b-a)^3}{12n^2} K = \tilde{K} \Delta h^2,$$

ahol $\tilde{K} = \frac{b-a}{12} K$ szintén konstans. A $\tilde{K}\Delta h^2$ felső korlát $\Delta h \to 0$ esetén nullához tart, így a lépésköz finomításával tetszőleges pontosság elérhető.
Annak jellemzésére, hogy egy közelítő módszer hibája hogyan viselkedik egyre kisebb lépésközökre, bevezetjük a következő fogalmat.

3.2.5 Definíció. *Tegyük fel, hogy* $g : K(0) \to \mathbb{R}$ *olyan függvény, amelyhez van olyan* $\tilde{p} \in \mathbb{N}^+$ *szám és* $K \in \mathbb{R}$ *korlát, hogy a 0-hoz kellően közeli* t *pontokban*

$$|g(t)| \leq K|t|^{\tilde{p}}.$$

Jelölje p *a legnagyobb ilyen tulajdonságú* $\tilde{p}$ *számot! Ekkor azt mondjuk, hogy a* g *függvény* p*-edrendben tart 0-hoz a 0 pontban. Ezt jelölje* $g(t) = \mathcal{O}(t^p)$ *(ejtsd: "ordó* t^p *".)*

Például a $g(t) = t^2$ függvényre $|t^2| \leq |t|$ a $[-1, 1]$-en, $|t^2| \leq |t|^2$ (mindenhol), de $|t^2| \leq K|t|^3$ már semmilyen K-ra nem teljesül a 0 egyetlen környezetén sem, ugyanis $\frac{|t^2|}{|t|^3} = \frac{1}{|t|}$ $t \to 0$-ra kinő a végtelenbe. Tehát a $g(t) = t^2$ függvény a 0-ban másodrendben tart 0-hoz.
Az összetett trapézszabály hibakorlátja szintén másodrendben tart nullához. Ez azt jelenti, hogy ha Δh-t q-ad részére csökkentjük, akkor a hibakorlát q^2-ed részére csökken. (Hiszen valamely Δh_1 lépésközhöz $\tilde{K}\Delta h_1^2$ hibakorlát tartozik, míg a $\Delta h_1/q$ lépésközhöz $\tilde{K}(\frac{\Delta h_1}{q})^2 = \frac{\tilde{K}\Delta h_1^2}{q^2}$.)

- $n = 2$, azaz három alappont van: $x_0 = a, x_1 = \frac{a+b}{2}$ és $x_2 = b$. Ekkor az interpolációs polinom másodfokú, az együtthatók:

$$A_0 = \int_a^b \frac{(x - \frac{a+b}{2})(x-b)}{(a - \frac{a+b}{2})(a-b)} \mathrm{d}x = \frac{h}{6} = A_2, \quad A_1 = \frac{4h}{6}.$$

Ezekből kapjuk a már szintén ismert

$$s(f) = \frac{h}{6}\left(f(a) + 4f(c) + f(b)\right), \quad c = \frac{a+b}{2}$$

Simpson- vagy **parabolaformulát**, amely tehát nem más, mint a függvényt közelítő, $f(a)$, $f(c)$ és $f(b)$ értékekre illeszkedő másodfokú interpolációs polinom grafikon

alatti területe, lásd a 3.1. ábra bal alsó grafikonját. Képlethibája $f \in C^4[a,b]$ függvényekre (bizonyítás nélkül)

$$\int_a^b f - s(f) = \int_a^b \frac{f'''(\xi)}{3!}(x-a)\left(x - \frac{a+b}{2}\right)(x-b)\mathrm{d}x = -\frac{(b-a)^5}{2880} f^{IV}(\kappa),$$

ahol $\kappa \in [a,b]$ valamely pont.

3.2.6 Következmény. *A Simpson-formula a harmadfokú polinomokra is pontos, mert azok negyedik deriváltja nulla.*

3.2.7 Megjegyzés. *Az összetett trazépformulához hasonlóan konstruálhatjuk meg az összetett Simpson-formulát. Legyen* $x_i = a + i\Delta h$ $(i = 0, 1 \ldots, n), \Delta h = \frac{b-a}{n}, y_i := \frac{x_{i-1}+x_i}{2}$ $(i = 1, \ldots, n)$.

$$\int_a^b f \approx \sum_{i=1}^{n} \frac{\Delta h}{6}(f(x_{i-1}) + 4f(y_i) + f(x_i)) =$$
$$\frac{\Delta h}{6}(f(x_0) + f(x_n) + 2\sum_{i=1}^{n-1} f(x_i) + 4\sum_{i=1}^{n} f(y_i)) =: s_n(f).$$

Képlethibája $f \in C^4[a,b]$ *függvényekre*

$$-\frac{(b-a)^5}{2880n^4} \cdot f^{IV}(\eta), \quad \eta \in [a,b] \text{ valamely pont.}$$

Mivel $|f^{IV}| \le K$, *ez a hiba abszolút értékben felülről becsülhető a*

$$\frac{(b-a)^5}{2880n^4} K \le \tilde{K}\Delta h^4$$

kifejezéssel. Ez a hibakorlát $\Delta h \to 0$ *esetén negyedrendben tart nullához, így az összetett Simpson-formula pontosabb közelítést nyújt, mint az összetett trapézformula.*

Feladat. Számítsuk ki az $\int_0^1 x^3\mathrm{d}x$ integrál közelítő értékét a trapézformulával. Becsüljük meg a hibát, majd hasonlítsuk össze a pontos értékkel. Végezzük el a számításokat a Simpson-formulával is. Ha az integrált az összetett trapéz- ill. az összetett Simpson-formulával, 10^{-4} pontossággal szeretnénk közelíteni, akkor hány részre daraboljuk ehhez a $[0,1]$ intervallumot?

Megoldás: A trapézszabály eredménye $t(f) = \frac{1}{2}$. A 3.2.1 hibabecslő képlet használatához kiszámítjuk az $f(x) = x^3$ függvény második deriváltját: $f''(x) = 6x$. Ebből $|f''(x)| \le 6\ \forall x \in [0,1]$, így $|\int_0^1 x^3\mathrm{d}x - t(f)| \le \frac{h^3}{12} \cdot 6 = \frac{1}{2}$. Az integrál pontos értéke $\frac{1}{4}$, amelytől az $\frac{1}{2}$ közelítő érték távolsága valóban kisebb, mint $\frac{1}{2}$. A Simpson-formula eredménye $s(f) = \frac{1}{4}$, ami pontos, és ez nem is meglepő, mivel a hibabecslő formulában

szereplő negyedik derivált az $f(x) = x^3$ függvényre nulla. Az összetett trapézszabály hibája a (3.5) formula szerint $-\frac{1}{12n^2} \cdot f''(\eta)$, ami abszolút értékben kisebb vagy egyenlő, mint $\frac{1}{12n^2} \cdot 6 = \frac{1}{2n^2}$. Ha egy n számra ez kisebb, mint 10^{-4}, akkor a hiba is kisebb. Tehát minden olyan n jó, amelyre $\frac{1}{2n^2} \leq 10^{-4}$. Ebből $n \geq \frac{100}{\sqrt{2}} = 70,71$, azaz legalább 71 részre kell darabolni az intervallumot. Az összetett Simpson-formula hibabecslésében szereplő negyedik derivált 0, ezért az minden n-re (már $n = 1$-re is) pontos.

4. fejezet

Közelítő differenciálás

Függvények deriváltjának közelítő kiszámítására szükségünk lehet például akkor, ha a függvényt csak diszkrét pontokban ismerjük, esetleg ott is csak közelítőleg. Például egy meteorológiai mérőállomáson a hőmérséklet folytonosan függ az időtől, de mérni csak bizonyos időközönként (diszkrét időpontokban) tudjuk. A 4.1. ábrán folytonos vonal mutatja a valódi hőmérséklet-idő függvényt egy konkrét esetben, és a csillagok az óránként mért értékeket. (Az egyszerűség kedvéért a mérések legyenek most teljesen pontosak.) Tegyük fel, hogy kíváncsiak vagyunk arra, hogy a hőmérséklet egy kiválasztott t^* időpontban milyen gyorsan változott. Ezt a $T(t)$ függvény t^*-beli deriváltja mutatja meg. Ezt nem tudjuk kiszámolni, mert a teljes függvényt nem ismerjük. Vajon hogyan lehet a diszkrét pontokban felvett értékekből közelíteni a keresett deriváltat?
A valós függvény deriváltjának legelterjedtebb közelítő formuláját a differenciálhányados definíciója szolgáltatja. Ismeretes, hogy az $f : \mathbb{R} \to \mathbb{R}$ függvény $x_0 \in intD(f)$-beli deriváltját az

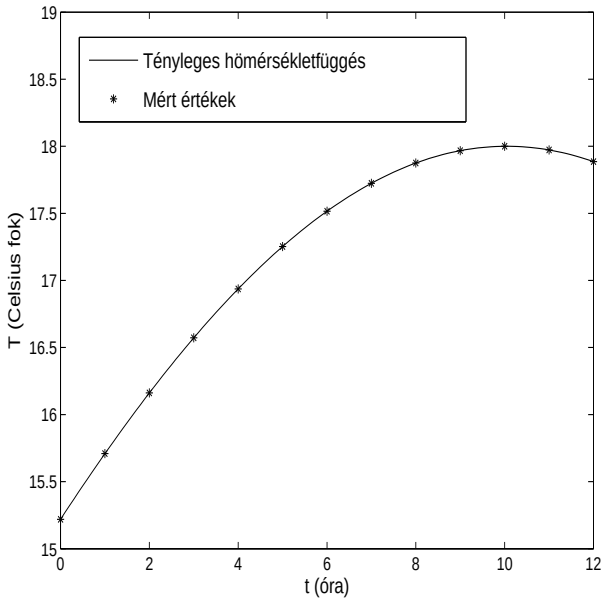
$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

határértékkel definiáljuk. Válasszunk ki például egy, az x_0-tól jobbra lévő x_1 pontot. Ha x_1 kellően közel van x_0-hoz, akkor az $f'(x_0)$ derivált értékét jól közelíti az

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

ún. jobb oldali különbségi hányados. A derivált definíciójából nem következik semmi ezen becslés hibájára nézve. Ha azonban a függvény kétszer folytonosan differenciálható, akkor a Taylor-tétel értelmében létezik olyan $\xi \in (x_0, x_1)$ pont, amelyre

$$f(x_1) = f(x_0) + \frac{1}{1!} f'(x_0)(x_1 - x_0) + \frac{1}{2!} f''(\xi)(x_1 - x_0)^2.$$



4.1. ábra. A hőmérséklet az időben folytonosan változik, de a meteorológiai állomásokon csak óránként mérik.

Ezt az egyenletet $(x_1 - x_0)$-val osztva az

$$f'(x_0) - \frac{f(x_1) - f(x_0)}{x_1 - x_0} = -\frac{1}{2!} f''(\xi)(x_1 - x_0) \tag{4.1}$$

hibaképlet adódik.

A függvény deriváltját természetesen egy x_0-tól balra elhelyezkedő x_{-1} pont segítségével, az

$$\frac{f(x_0) - f(x_{-1})}{x_0 - x_{-1}}$$

bal oldali különbségi hányadossal is közelíthetjük. A fenti hibabecslő formula ekkor is alkalmazható, ha x_1 helyére x_{-1}-et írunk.

Feladat. Közelítsük az $f(x) = x^2$ függvény deriváltját az $x_0 = 2$ pontban az x_0-beli és az $x_1 = 2.1$-beli függvényértékek felhasználásával. Alkalmazzuk a (4.1) formulát a hiba becslésére. Hogyan viszonyul ez a hiba pontos értékéhez?

Megoldás: A pontos derivált az $x_0 = 2$ pontban 4. Közelítése a differenciahányadossal

$$f'(2) \approx \frac{2,1^2 - 2^2}{2,1 - 2} = 4,1.$$

A közelítés hibája tehát 0,1. A (4.1) formula szerint a hiba $-\frac{1}{2} f''(\xi)(2,1-2)$, ahol $\xi \in (x_0, x_1)$ valamely pont. Mivel $f''(x) = 2$ mindenhol, így a ξ pontban is, így ez a kifejezés $0,1$-del egyenlő. A hibabecslő

formulával tehát ebben az esetben a hiba pontos értékét kaptuk.

Ha a függvény első deriváltját tudjuk közelíteni, akkor ennek felhasználásával megbecsülhető a második derivált, hiszen definíció szerint

$$f''(x_0) = \lim_{x \to x_0} \frac{f'(x) - f'(x_0)}{x - x_0}.$$

Ha tehát x_1 közel van x_0-hoz, akkor $f''(x_0)$-t jól közelíti az

$$\frac{f'(x_1) - f'(x_0)}{x - x_0}$$

hányados. Vegyünk fel x_0-tól balra is egy x_{-1} alappontot, és a fenti hányadosban szereplő első deriváltakat közelítsük az

$$f'(x_1) \approx \frac{f(x_1) - f(x_0)}{x_1 - x_0}, \quad f'(x_0) \approx \frac{f(x_0) - f(x_{-1})}{x_0 - x_{-1}}$$

különbségi hányadosokkal. Ha a felosztás egyenlőközű, azaz $x_1 - x_0 = x_0 - x_{-1} =: h$, akkor ebből az

$$f''(x_0) \approx \frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2}$$

közelítő képletet nyerjük, amely a második derivált leggyakoribb véges különbséges közelítése.

5. fejezet

Közönséges differenciálegyenletek megoldása

A földtani jelenségek matematikai leírásában rendkívül fontosak a közönséges ill. parciális differenciálegyenletek. Ebben a fejezetben a közönséges differenciálegyenletekre vonatkozó kezdetiérték-feladatok numerikus megoldásával foglalkozunk.

Tekintsük az

$$\begin{cases} y'(t) = f(t, y(t)) \quad t \in [t_0, T] \\ y(t_0) = y_0 \end{cases} \tag{5.1}$$

kezdetiérték-feladatot. A megoldás létezését és egyértelműségét feltételezzük. (Ezek biztosítottak például, ha f azonkívül, hogy folytonos, kielégíti a Lipschitz-feltételt.)

Az ilyen feladatok pontos megoldását nem mindig tudjuk meghatározni. Ismeretes, hogy ha az $f(t, y(t))$ jobb oldal csak t függvénye, akkor a pontos megoldás

$$y(t) = y_0 + \int_{t_0}^{t} f(t')\mathrm{d}t' \quad t \in [0, T]$$

alakú, tehát f integrálását igényli. Ezt sokszor nem tudjuk zárt alakban elvégezni. Ha pedig f az y-tól is függ, a helyzet még bonyolultabb. Ekkor ugyanis a pontos megoldás alakja

$$y(t) = y_0 + \int_{t_0}^{t} f(t', y(t'))\mathrm{d}t', \quad t \in [0, T]$$

és itt a jobb oldali integrálban nem is ismerjük $y(t)$-t. Ezért a legtöbb esetben közelítő módszereket kell alkalmaznunk.

5.1. Véges különbséges módszerek

Az (5.1) közelítő megoldására az egyik legelterjedtebb módszercsalád a véges különbséges módszerek, ezért mi is ezek ismertetésére szorítkozunk.

A véges különbséges módszerek alkalmazásakor a feladatot először **diszkretizáljuk**. A függvény értelmezési tartományán definiálunk egy diszkrét pontokból álló

$$\Omega := \{t_0, t_1, \ldots, t_n\}$$

rácsot, amely most az egyszerűség kedvéért legyen egyenlőközű, azaz tegyük fel, hogy a szomszédos pontok egymástól egyenlő Δt távolságra vannak. Az ismeretlen függvényt a rácspontokban akarjuk közelíteni. A közelítő megoldás tehát nem egy folytonos függvény, hanem egy ún. **rácspontfüggvény** lesz. Jelöljük y_i-vel a megoldás t_i pontbeli közelítését, ahol $i = 1, 2, \ldots, n$. Fontos, hogy ezt ne keverjük össze az $y(t_i)$ jelöléssel, amely a feladat pontos megoldását jelenti a t_i pontban! Az intervallum t_0 kezdőpontjában a megoldás a kezdeti feltételből ismert. Hogyan tudjuk közelíteni a megoldás t_1-beli értékét? Ehhez hívjuk segítségül a közelítő deriválás 4. fejezetben tanult módszerét! A t_0-beli deriváltat az

$$y'(t_0) \approx \frac{y(t_1) - y(t_0)}{\Delta t}$$

hányadossal közelítjük. A bal oldal az (5.1) feladat szerint $f(t_0, y(t_0))$-val egyenlő. Itt és a jobb oldalon $y(t_0) = y_0$ a megadott kezdeti feltétel, tehát ismert. Az egyenletből $y(t_1)$ közelítőleg kifejezhető, ez lesz a megoldás t_1-beli értékének közelítése:

$$y(t_1) \approx f(t_0, y_0) \cdot \Delta t + y(t_0) =: y_1.$$

Hasonlóan léphetünk tovább a t_2 pontra:

$$y'(t_1) \approx \frac{y(t_2) - y(t_1)}{\Delta t} \approx \frac{y(t_2) - y_1}{\Delta t}.$$

Itt $y'(t_1) = f(t_1, y(t_1)) \approx f(t_1, y_1)$. Ebből

$$y(t_2) \approx f(t_1, y_1) \cdot \Delta t + y_1 =: y_2$$

és így tovább. Így az algoritmus az

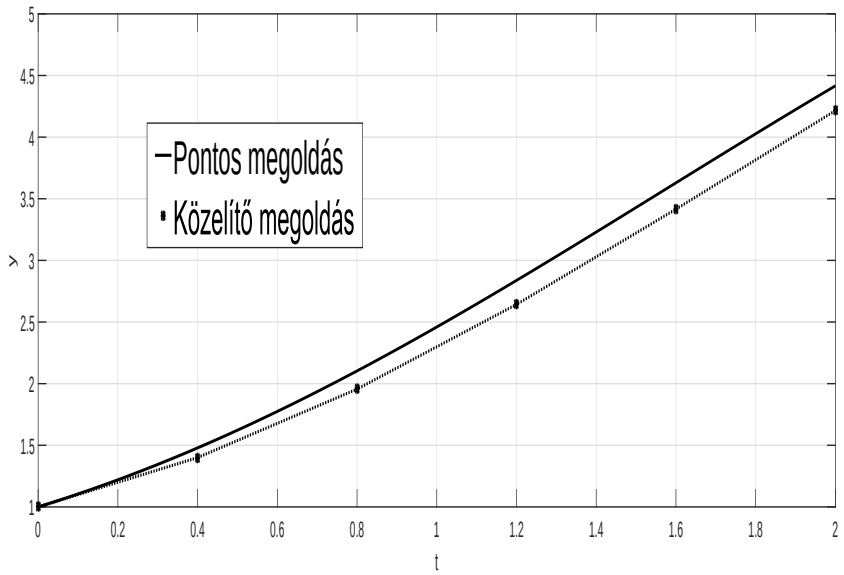
$$y_{i+1} = f(t_i, y_i) \cdot \Delta t + y_i, \;\; i = 0, \ldots, n-1.$$

egyenletrendszert adja a megoldás Ω-n való közelítésére. (Vegyük észre, hogy általában $y_{i+1} \neq y(t_{i+1})$, csupán azt várhatjuk el, hogy közel legyenek egymáshoz.) Ezt az eljárást

explicit Euler-módszernek nevezzük. Az explicit Euler-módszerrel kapott közelítő megoldást szemlélteti a 5.1. ábra, ahol a módszert az

$$\begin{cases} y'(t) = \sin t + 1 \quad t \in [0,2] \\ y(0) = 1 \end{cases} \tag{5.2}$$

feladatra alkalmaztuk.



5.1. ábra. Az $y'(t) = \sin t + 1,\ y(0) = 1$ feladat pontos megoldása (folytonos vonal) és az explicit Euler-módszerrel, $\Delta t = 0.4$-es lépésközzel kapott közelítő megoldása (a szürke vonallal összekötött pontok) a $[0,2]$ intervallumon.

Bármilyen megoldó módszertől elvárható, hogy a közelítő megoldás konvergáljon a feladat pontos megoldásához. Hogyan kell ezt érteni? Nyilván azt szeretnénk, hogy az igazi megoldást tetszőleges pontossággal közelítsük, ha Δt-t elég kicsinek választjuk. Építsük fel az egyre finomodó (Ω_n) rácshálósorozatot, és jelölje Δt_n az n-edik rácsháló lépésközét. Legyen továbbá t^* tetszőleges olyan pont amely valamely rácshálótól kezdve mindegyikben benne van, és jelölje k_n azt a lépésszámot, amellyel t_0-ból Δt_n lépésközzel lépve éppen t^*-ba jutunk, azaz $k_n := (t^* - t_0)/\Delta t_n$. A közelítő megoldás konvergenciája azt jelenti, hogy

$$\lim y_{k_n} = y(t^*). \tag{5.3}$$

Belátható, hogy az explicit Euler-módszer konvergens, amennyiben az f függvény folytonosan differenciálható. Itt ezt nem bizonyítjuk. Helyette tanulmányozzuk a módszert egy próbafeladaton!

A kezdetiérték-feladat gyakran használt mintafeladata a következő:

$$\begin{cases} y'(t) = \lambda y(t) \quad t \in [0,T] \\ y(0) = 1, \end{cases} \tag{5.4}$$

ahol $\lambda \in \mathbb{R}$ állandó. Ennek pontos megoldása az

$$y(t) = e^{\lambda t} \quad t \in [0,T]$$

függvény. Generáljunk a $[0,T]$ intervallumhoz egy (Ω_n) rácshálósorozatot, legyen $t^* \in \Omega_n$ tetszőleges közös eleme az összes rácshálónak (valamelyiktől kezdve), és $k_n = t^*/\Delta t_n$. Alkalmazzuk (5.4)-re az explicit Euler-módszert!

$$\begin{aligned} y_0 &= 1 \\ y_{i+1} &= \lambda y_i \Delta t_n + y_i = (1 + \lambda \Delta t_n) y_i, \end{aligned} \tag{5.5}$$

és indukcióval

$$y_{k_n} = (1 + \lambda \Delta t_n)^{k_n}.$$

Vizsgáljuk meg az y_{k_n} határértékét!

$$\lim y_{k_n} = \lim (1 + \lambda \Delta t_n)^{k_n} = \lim \left(1 + \frac{\lambda t^*}{k_n}\right)^{k_n} = e^{\lambda t^*}. \tag{5.6}$$

Vegyük észre, hogy $e^{\lambda t^*}$ éppen az $y(t^*)$ pontos megoldással egyenlő. Tehát a módszer a vizsgált próbafeladaton konvergens.
A próbafeladat alapján valamit a módszer konvergenciasebességéről is mondhatunk. Az $y_{k_n} = (1 + \lambda \Delta t_n)^{t^*/\Delta t_n}$ közelítő megoldásra egy jól ismert logaritmikus azonosságot alkalmazva, majd Δt szerint sorba fejtve

$$\begin{aligned} (1 + \lambda \Delta t_n)^{\frac{t^*}{\Delta t_n}} &= e^{\frac{t^*}{\Delta t_n} \ln(1 + \lambda \Delta t_n)} \\ &= e^{\frac{t^*}{\Delta t_n}(\lambda \Delta t_n - \frac{\lambda^2 \Delta t_n^2}{2} + \mathcal{O}(\Delta t_n^3))} \\ &= e^{\lambda t^* (1 - \frac{\lambda \Delta t_n}{2} + \mathcal{O}(\Delta t_n^2))} \\ &= e^{\lambda t^*} e^{-\frac{\lambda^2 t^* \Delta t_n}{2} + \mathcal{O}(\Delta t_n^2)} \\ &= e^{\lambda t^*} \{1 - \frac{\lambda^2 t^* \Delta t_n}{2} + \mathcal{O}(\Delta t_n^2)\}. \end{aligned}$$

Azaz rögzített t^* esetén a hibát a következő kifejezés adja meg:

$$\underbrace{y_{k_n}}_{\text{közelítő érték a } t^*\text{-ban}} - \underbrace{e^{\lambda t^*}}_{\text{pontos megoldás a } t\text{-ben}} = -\tfrac{1}{2}\lambda^2 t^* \Delta t_n e^{\lambda t^*} + \mathcal{O}(\Delta t_n^2) \tag{5.7}$$

Így $y_{k_n} - y(t^*) = \mathcal{O}(\Delta t_n)$, vagyis a hiba Δt_n-nel elsőrendben tart 0-hoz. (Ez általánosan is belátható az explicit Euler-módszerre.)

6. fejezet

Egyenletek közelítő megoldása

Az egyismeretlenes valós egyenletek egy része az egyenlet rendezésével könnyen megoldható. A legegyszerűbbek a lineáris egyenletek, de a másodfokúakat is gond nélkül meg tudjuk oldani a jól ismert megoldóképlet segítségével. A harmad- és negyedfokú egyenletekre is létezik megoldóképlet, de már jóval bonyolultabb. Magasabbfokú egyenletekre pedig már bizonyíthatóan nem lehet általános képletet felállítani. A logaritmikus, trigonometrikus és egyéb egyenletek megoldása esetenként szintén problémát jelenthet.

Világos, hogy minden egyismeretlenes valós egyenlet a tagok bal oldalra rendezésével felírható

$$f(x) = 0$$

alakban, ahol $f : \mathbb{R} \to \mathbb{R}$ valamilyen függvény. A továbbiakban az ilyen alakú egyenletek megoldásával foglalkozunk. Ezen felírás mellett az egyenlet megoldása egyben az f függvény **zérushelye**, vagy más szóval **gyöke**.

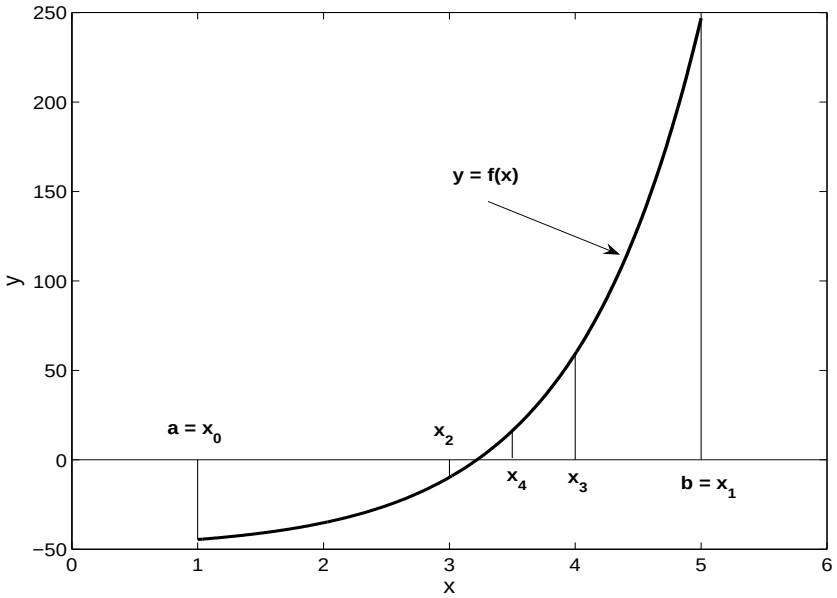
A megoldandó egyenletről sokszor belátható, hogy létezik gyöke. Ez a helyzet például akkor, ha f folytonos, és a határértéke a mínusz és a plusz végtelenben ellentétes előjelű. A nehézséget a megoldás megtalálása jelenti. Erre látunk most különféle módszereket.

6.1. Intervallumfelezés

Legyen f folytonos az $[a, b]$ intervallumon, és tegyük fel, hogy $f(a) \cdot f(b) < 0$, azaz f az a és a b pontban ellentétes előjelű. Ekkor Bolzano tétele értelmében (a, b)-ben található legalább egy olyan x^* pont, ahol $f(x^*) = 0$.

Felezzük meg az $[a, b]$ intervallumot. Ha f értéke a felezőpontban nulla, akkor f valamelyik gyökét már meg is találtuk. Ha nem nulla, akkor válasszuk ki a két részintervallum közül azt, amelyiknek a végpontjaiban az f előjele ellentétes, és az eljárást ismételjük meg erre a részintervallumra, és így tovább. A megtartott intervallumok met-

szete egyetlen számot tartalmaz, és ez az f függvény valamelyik gyöke (lásd a 6.1. ábrát).



6.1. ábra. Intervallumfelezés.

Az eljárás vagy véges sok lépésben befejeződik (amikor valamelyik felezőpontban az f értéke pontosan zérus), vagy elég sok lépés után tetszőlegesen kis intervallumba szorítjuk f valamelyik gyökét. A k-adik ($k \geq 1$) lépésben nyert intervallum hossza

$$\frac{b-a}{2^k},$$

így a részintervallum bármelyik végpontja adott $\varepsilon > 0$ hibakorlátnál pontosabban fogja közelíteni f $[a, b]$-beli egyik gyökét, ha

$$\frac{b-a}{2^k} < \varepsilon,$$

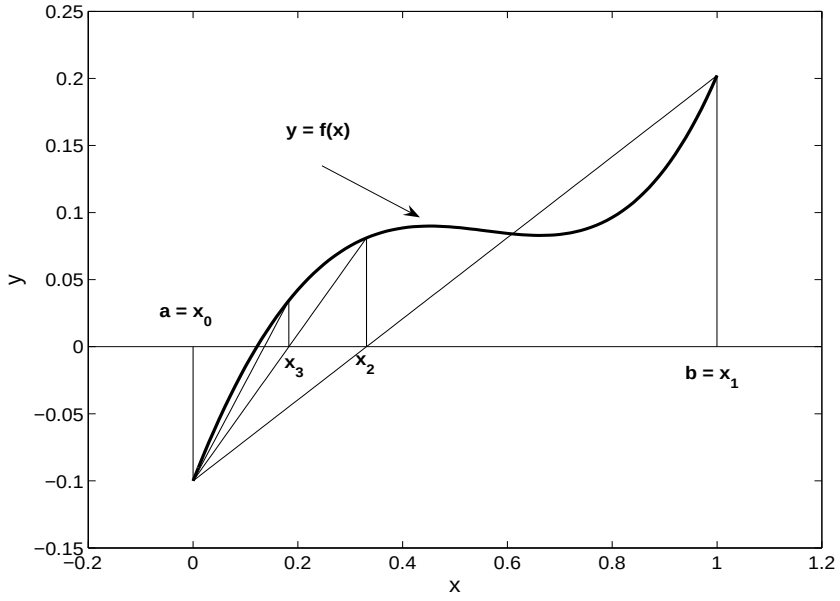
vagyis átrendezve

$$k > \log_2 \frac{b-a}{\varepsilon}.$$

6.2. Húrmódszer

Az intervallumfelezésnél a részintervallumokat mindig két egyenlő részre osztottuk. Célszerűnek látszik a felezési eljárást úgy módosítanunk, hogy a részintervallumokat a végpontokban felvett függvényértékek arányában osszuk fel. Ez geometriailag azt jelenti,

hogy a függvénynek a részintervallum végpontjaihoz tartozó húrjának és az x tengelynek a metszéspontját választjuk osztópontnak (lásd a 6.2. ábrát).



6.2. ábra. A húrmódszer.

Az eljárás algoritmusa tehát a következő. Legyen az f folytonos az $[a, b]$ intervallumon, és legyen $x_0 := a$, $x_1 := b$, vagy a és b között két olyan pont, amelyek közrefogják a gyököt. Tegyük fel, hogy $f(x_0) \cdot f(x_1) < 0$. Kössük össze az $(x_0, f(x_0))$ és az $(x_1, f(x_1))$ pontot egy egyenessel, és jelöljük annak x tengellyel alkotott metszéspontját x_2-vel. Ekkor $f(x_2)$ előjelétől függően vagy az $[x_0, x_2]$, vagy az $[x_2, x_1]$ részintervallumon folytatjuk tovább az eljárást. Ha az adott lépésben x_k és x_j $(j < k)$ jelöli a megtartott részintervallum végpontjait, akkor az összekötő húr egyenlete

$$y = \frac{f(x_k) - f(x_j)}{x_k - x_j} \cdot (x - x_k) + f(x_k),$$

amelyből $y = 0$ helyettesítéssel adódik a húrnak az x tengellyel alkotott metszéspontja:

$$x_{k+1} = x = x_k - \frac{x_k - x_j}{f(x_k) - f(x_j)} \cdot f(x_k). \tag{6.1}$$

6.2.1 Megjegyzés. *Ha a függvény konvex, azaz a húr mindig a függvény grafikonja fölött halad, akkor az osztópontbeli függvényérték minden lépésben negatív lesz, ezért x_j a fenti képletben mindig x_1-gyel egyenlő (azaz mindig a jobb oldali részintervallumot kell választani). Ha pedig a függvény konkáv, akkor ennek a fordítottja igaz, és x_j minden lépésben x_0-val egyenlő (mindig a bal oldali részintervallumot választjuk).*

6.2.2 Tétel. *(Bizonyítás nélkül) A megadott feltételek mellett a húrmódszerrel kapott* (x_k) *sorozatnak létezik véges* x^* *határértéke, és ez a határérték az egyenlet valamelyik gyöke.*

Ez a tétel nem ad felvilágosítást a konvergencia sebességéről.

6.2.3 Tétel. *(Bizonyítás nélkül) Tegyük fel, hogy* f *kétszer differenciálható, és minden* $x \in [a,b]$ *esetén* $|f'(x)| \geq m > 0$*, valamint* $|f''(x)| \leq M < \infty$*. Ekkor*

$$|x_{k+1} - x^*| \leq \frac{M}{2m}|x_k - x^*| \cdot |x_j - x^*|. \tag{6.2}$$

Vezessük be a $d_k := \frac{M}{2m}|x_k - x^*|$ jelölést. Ekkor (6.2) mindkét oldalát megszorozva $M/2m$-mel, a d_k mennyiségekre a

$$d_{k+1} \leq d_k d_j$$

egyenlőtlenséghez jutunk. Ha x_0 és x_1 olyan jó közelítése x^*-nak, hogy $d := \max\{d_0, d_1\} < 1$, akkor $j = 0$ választással teljes indukcióval belátható, hogy

$$d_{k+1} \leq d^{k+1} \quad (k \geq 0). \tag{6.3}$$

Ezzel a gyakorlatban is jól alkalmazható hibabecslő formulát nyertünk.

6.3. Szelőmódszer

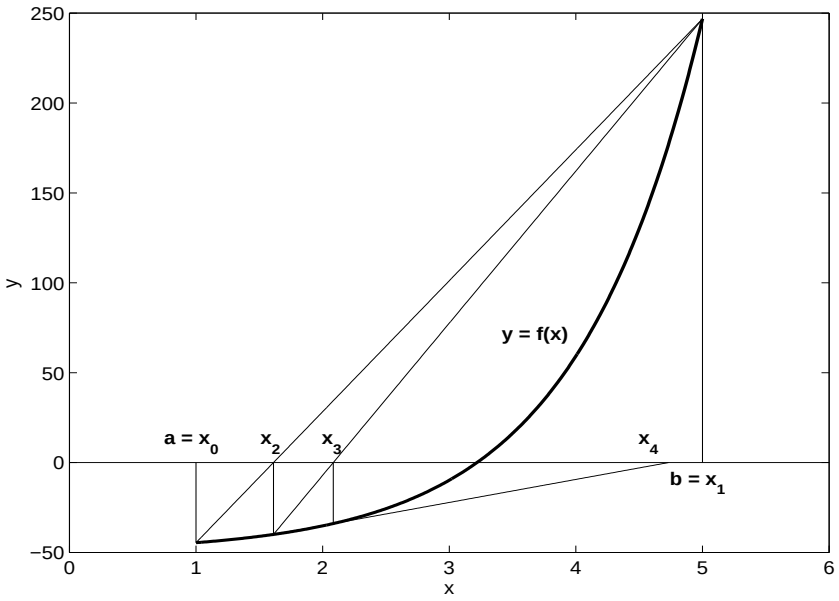
A húrmódszer esetén minden további lépés számításakor meg kell vizsgálni $f(x_k)$ előjelét, és ennek alapján döntjük el, hogy az x_{k+1} újabb közelítést melyik intervallumban vesszük fel. A húrmódszer másik hátránya az, hogy konvex vagy konkáv f függvényeknél a (6.1)-ben x_j állandó, mégpedig a kezdeti intervallum egyik végpontja. A (6.2) alapján gyorsabban konvergáló eljárást nyerhetünk, ha minden lépésben a $j = k - 1$ választással dolgozunk, hiszen ha a módszer konvergens, akkor (6.2) jobb oldalának a húrmódszer esetében gyakran konstans harmadik tényezője is nullához fog tartani. Az így nyert

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \cdot f(x_k) \tag{6.4}$$

formulát **szelőmódszer**nek nevezzük (lásd a 6.3. ábrát).

Mivel a szelőmódszer speciális húrmódszer, ezért hibájáról (6.2) alapján a következőt mondhatjuk.

6.3.1 Tétel. *Tegyük fel, hogy* f *kétszer differenciálható, és minden* $x \in [a,b]$ *esetén*



6.3. ábra. A szelőmódszer.

$|f'(x)| \geq m > 0$, *valamint* $|f''(x)| \leq M < \infty$. *Ekkor*

$$|x_{k+1} - x^*| \leq \frac{M}{2m}|x_k - x^*| \cdot |x_{k-1} - x^*|. \tag{6.5}$$

Az egyenlőtlenség mindkét oldalát $M/2m$-mel szorozva a $d_k := \frac{M}{2m}|x_k - x^*|$ mennyiségekre a

$$d_{k+1} \leq d_k d_{k-1}$$

összefüggés adódik. Tegyük fel, hogy x_0 és x_1 olyan jó közelítése x^*-nak, hogy $d := \max\{d_0, d_1\} < 1$. Ekkor teljes indukcióval belátható, hogy

$$d_{k+1} \leq d^{\gamma_k} \quad (k \geq 0), \tag{6.6}$$

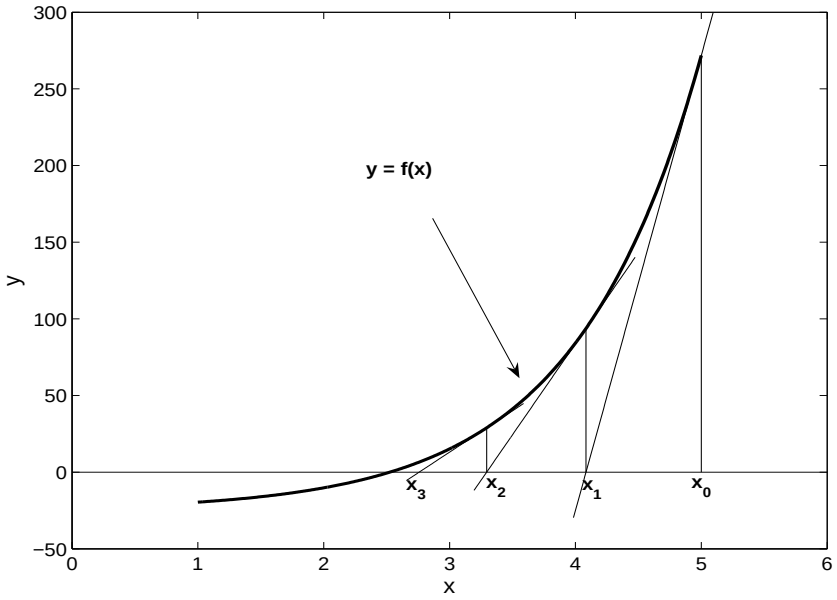
ahol γ_k a $\gamma_0 = 0$, $\gamma_1 = 1$, $\gamma_k = \gamma_{k-1} + \gamma_{k-2}$ rekurzióval megadott ún. Fibonacci-sorozat k-adik tagja. A (6.6)-t a (6.3) egyenlőtlenséggel összevetve a szelőmódszer gyorsabb konvergenciája közvetlenül leolvasható.

6.4. Newton-módszer (érintőmódszer)

Az előző módszerek az x^* gyökhely két elég jó közelítésének ismeretét igénylik. Newton módszerénél elég egyetlen elegendően pontos x_0 közelítést ismernünk. Az f függvénynek azonban deriválhatónak kell lennie, sőt, mint látni fogjuk, a módszer konvergenciájának

bizonyításához a második derivált létezését is feltesszük.

Húzzuk meg az f függvény x_0-beli érintőjét, és jelölje x_1 az érintő x tengellyel alkotott metszéspontját. Az eljárás folytatásával egy $x_0, x_1, x_2, \ldots$ sorozatot nyerünk, lásd a 6.4. ábrát. (Ha f-nek állandó a konvexitása, akkor ez a sorozat monoton lesz.) Az f függvény



6.4. ábra. A Newton-módszer.

x_k-pontbeli érintőjének egyenlete

$$y = f'(x_k)(x - x_k) + f(x_k), \tag{6.7}$$

amelyből $y = 0$ helyettesítéssel kapjuk a Newton-módszer formuláját:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \tag{6.8}$$

6.4.1 Tétel. *Tegyük fel, hogy az x^* megoldást és az (x_k) sorozatot tartalmazó valamilyen I intervallumban f'' létezik és folytonos, valamint $|f'| \geq m > 0$ és $|f''| \leq M < \infty$ az I-n. Ekkor*

$$|x_{k+1} - x^*| \leq \frac{M}{2m}|x_k - x^*|^2. \tag{6.9}$$

Biz.: A Taylor-formula alapján

$$0 = f(x^*) = f(x_k) + f'(x_k) \cdot (x^* - x_k) + \frac{1}{2}f''(\xi) \cdot (x^* - x_k)^2, \tag{6.10}$$

ahol ξ az x_k és az x^* között van. A Newton-módszer (6.8) egyenletéből

$$0 = f(x_k) + f'(x_k)(x_{k+1} - x_k).$$

Ezt az egyenletet a (6.10) összefüggésből kivonva a

$$0 = f'(x_k)(x^* - x_{k+1}) + \frac{1}{2}f''(\xi)(x^* - x_k)^2$$

egyenlet adódik. Ebből

$$|x_{k+1} - x^*| = \left|\frac{f''(\xi)}{2f'(x_k)}\right| |x_k - x^*|^2 \leq \frac{M}{2m}|x_k - x^*|^2.$$

A tételből következik, hogy $d_k := \frac{M}{2m}|x_k - x^*|$ jelöléssel

$$d_{k+1} \leq d_k^2,$$

amiből

$$d_k \leq d_0^{2^k},$$

tehát

$$|x_k - x^*| \leq \frac{2m}{M} \cdot \left(\frac{M}{2m}|x_0 - x^*|\right)^{2^k}. \tag{6.11}$$

Feladatok.

1. Oldjuk meg a Newton-módszerrel az $x = -e^x$ egyenletet, ha $x_0 = 0$. Hány lépés kell a 10^{-4}-es pontosság eléréséhez?
   Megoldás: Rendezzük át az egyenletet $x + e^x = 0$ alakba. Az $f(x) = x + e^x$ függvény szigorúan monoton növő, $f(0) = 1 > 0, f(-1) = -1 + \frac{1}{e} < 0$, így $x^* \in [-1, 0]$. Továbbá f konvex (mivel $f''(x) = e^x > 0$), ezért (x_k) monoton csökkenve tart az x^*-hoz. Így a $[-1, 0]$ intervallum tartalmazza az x^* megoldást és az (x_k) sorozatot. Ezen az intervallumon $|f'(x)| = 1 + e^x \geq 1$ és $|f''(x)| = e^x \leq 1$. Tehát $m = 1$ és $M = 1$ mellett alkalmazható a (6.11) hibabecslő képlet:

   $$|x_k - x^*| \leq \frac{2}{1} \cdot \left(\frac{1}{2}|x_0 - x^*|\right)^{2^k}.$$

   A jobb oldalon $|x_0 - x^*| \leq 1$, tehát azt a legkisebb k lépésszámot keressük, amelyre

   $$2 \cdot \left(\frac{1}{2}\right)^{2^k} \leq 10^{-4}.$$

   Átalakítások után ebből a

   $$2^k \geq -\frac{4}{\lg 0,5} + 1 \approx 14,29$$

   feltétel adódik. A legkisebb k hatvány, amelyre ez fennáll, a $k = 4$. Négy lépést teszünk a

Newton-módszerrel:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = -0,5;$$
$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = -0.5663;$$
$$\vdots$$
$$x_4 = -0.5671.$$

2. Az $e^x - x^2 - 2 = 0$ egyenlet gyöke 1,2 és 1,5 közé esik. Hány lépés kell a 10^{-6}-os pontosság eléréséhez a Newton-, ill. a húrmódszerrel számolva?
Megoldás: Mivel $f'(x) = e^x - 2x > 0$, ezért az $f(x)$ függvény szigorúan monoton növő. Az $f''(x) = e^x - 2 > 0$ második derivált függvény pedig az $I = [1,2; 1,5]$ intervallumon pozitív, ezért f konvex I-n. Az I jobb oldali végpontjából indítva az iterációt, $(x_k) \to x^*$ monoton csökkenve, és $x_k \in I\ \forall k$. Továbbá, f'' monoton növekedése miatt f'' maximuma az I-n a jobb oldali végpontban van, tehát $|f''(x)| \le e^{1,5} - 2 < 3 =: M$. Az $f'(x) = e^x - 2x$ szintén monoton nő, mert $f''(x) > 0$ az I-n. Ezért $|f'(x)| \ge e^{1,2} - 2 \cdot 1,2 = e^{1,2} - 2,4 > \frac{1}{2} =: m$. Tehát a Newton-módszer esetén

$$|x_k - x^*| \le \frac{1}{3} \cdot (3|x_0 - x^*|)^{2^k} \le \frac{1}{3} \cdot 0,9^{2^k} \le 10^{-6},$$

ha $k \ge 7$. A húrmódszerrel számolva (lásd a (6.3) formulát)

$$|x_k - x^*| \le \frac{1}{3} \cdot (3|x_k - x^*|)^k \le \frac{1}{3} \cdot 0,9^k \le 10^{-6},$$

ha $k \ge 121$.

7. fejezet

Lineáris algebrai egyenletrendszerek megoldása

A gyakorlati feladatoknál sűrűn előfordul, hogy lineáris egyenletrendszereket kell megoldanunk. Többek között a lineáris parciális differenciálegyenletek véges különbséges módszerrel történő megoldása is lineáris algebrai egyenletrendszerek megoldására vezet.

A lineáris egyenletrendszerek megoldására szolgáló módszerek két nagy csoportba oszthatók. **Direkt módszer**ről akkor beszélünk, ha a módszerben szereplő számolásokat kerekítés nélkül pontosan elvégezve az egyenletrendszer pontos megoldását kapnánk. Az **iterációs módszer**ek egy iterációs vektorsorozatot generálnak, amely a pontos megoldásvektorhoz tart. A sorozat elég nagy indexű tagja a megoldást tetszőlegesen megközelíti.

Leszögezzük, hogy itt csak olyan egyenletrendszerek megoldásával foglalkozunk, amelyek mátrixa négyzetes és reguláris. Ismeretes, hogy ekkor az egyenletrendszernek létezik egyetlen megoldása.
Írjuk fel az egyenletrendszert az

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots a_{2n}x_n &= b_2 \\ &\dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots a_{nn}x_n &= b_n \end{aligned}$$

alakban, amely az

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{7.1}$$

jelölések bevezetésével rövidítve

$$Ax = b$$

alakban is felírható.

7.1. Direkt módszerek

7.1.1. Kiküszöbölési (eliminációs) eljárások

A kiküszöbölési eljárások azon az észrevételen alapulnak, hogy egyes speciális egyenletrendszerek könnyen megoldhatók. A legegyszerűbb eset az, amikor az A együtthatómátrix diagonális. Tekintsük például az

$$\begin{aligned} 5x_1 &= 10 \\ 2x_2 &= 4 \\ 3x_3 &= -1 \end{aligned}$$

egyenletrendszert. Ebben az esetben

$$A = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \quad \text{és} \quad b = \begin{bmatrix} 10 \\ 4 \\ -1 \end{bmatrix},$$

és az ismeretlenek az egyenletekből közvetlenül kifejezhetők:

$$\begin{aligned} x_1 &= 2 \\ x_2 &= 2 \\ x_3 &= -\tfrac{1}{3}. \end{aligned}$$

Könnyű dolgunk van háromszögmátrixú egyenletrendszerek esetén is. Az

$$\begin{aligned} x_1 \quad +2x_2 \quad +2x_3 &= 7 \\ 4x_2 \quad +5x_3 &= 17 \\ 9x_3 &= 27 \end{aligned}$$

egyenletrendszer mátrixa felsőháromszög-mátrix. Az utolsó egyenletből kifejezhető az x_3 ismeretlen: $x_3 = 27/9 = 3$. Ezt behelyettesítjük a második egyenletbe, és kifejezzük x_2-t: $x_2 = 1/2$. Végül x_1 az első egyenletből adódik x_2 és x_3 behelyettesítésével: $x_1 = 7 - 2 \cdot 3 - 2 \cdot \frac{1}{2} = 1$.

Ha az egyenletrendszer se nem diagonális, se nem háromszögmátrixú, akkor ezeket a megoldási módszereket nem tudjuk közvetlenül alkalmazni. Szerencsére azonban minden egyéb reguláris négyzetes mátrixú egyenletrendszer átalakítható akár diagonális mátrixú, akár háromszögmátrixú rendszerré. Az átalakítás ekvivalens átalakítást jelent, azaz olyat, amely nem változtatja meg az egyenletrendszer megoldását. Így például

1. egyenleteket felcserélhetünk;
2. egyenleteket bármilyen nullától különböző számmal szorozhatunk;
3. bármelyik egyenlethez hosszáadhatjuk egy másik egyenlet tetszőleges számszorosát.

A következőkben két alapvető kiküszöbölési eljárást ismertetünk.

A Gauss-elimináció

A Gauss-elimináció során az egyenletrendszert az előző részben ismertetett 1., 2. ill. 3. műveletek segítségével felsőháromszög-mátrixú rendszerré transzformáljuk. Ehhez célszerű az együtthatókat és a szabad tagokat táblázatba foglalni, és az átalakításokat ezen a táblázaton nyomon követni, hiszen az ismeretleneket fölösleges minden egyes lépésben kiírnunk. Az elemek nullára transzformálása során felülről lefelé és balról jobbra haladunk. Így elkerülhetjük, hogy már lenullázott elem újra feltöltődjön.

Alkalmazzuk ezt a módszert a

$$\begin{array}{rrrcl} 2x_1 & +x_2 & +x_3 & = & 4 \\ x_1 & +3x_2 & +2x_3 & = & 6 \\ x_1 & +2x_2 & +2x_3 & = & 5 \end{array}$$

egyenletrendszerre! Készítsük el az együtthatók és a szabad tagok 7.1 táblázatát.

2	1	1	4
1	3	2	6
1	2	2	5

7.1. táblázat.

Lépések:

1. Először a bal felső elem alatti számokat transzformáljuk nullára. Ehhez a 2. ill. a 3.

2	1	1	4
0	5/2	3/2	4
0	3/2	3/2	3

7.2. táblázat.

egyenletből kivonjuk az első egyenlet $\frac{1}{2}$-szeresét (7.2. táblázat).
2. A főátló alatt már csak egy nemnulla elem van: a 3. sor 2. eleme. Ezért a 3. sorból kivonjuk a 2. sor $\frac{3}{5}$-szeresét (7.3. táblázat). Tehát a megoldandó egyenletrendszer

2	1	1	4
0	5/2	3/2	4
0	0	3/5	3/5

7.3. táblázat.

ekvivalens a

$$\begin{aligned} 2x_1 \quad +x_2 \quad +x_3 \quad &= \quad 4 \\ \tfrac{5}{2}x_2 \quad +\tfrac{3}{2}x_3 \quad &= \quad 4 \\ \tfrac{3}{5}x_3 \quad &= \quad \tfrac{3}{5} \end{aligned}$$

egyenletrendszerrel. Ebből az ismeretleneket alulról fölfelé kiküszöbölve az

$$x_1 = 1, \quad x_2 = 1, \quad x_3 = 1$$

megoldáshoz jutunk.

A Gauss-eliminációt időnként az ún. **főelem-kiválasztás** módszerével kombináljuk. Láthattuk, hogy a szabályos Gauss-eliminációs algoritmus során az aktuális táblázat i-edik oszlopában úgy tudjuk nullára transzformálni az a_{ii} alatti elemeket, hogy az i-edik sort először elosztjuk a_{ii}-vel, majd az így kapott sor megfelelő számszorosát kivonjuk vagy hozzáadjuk a lejjebb lévő sorokhoz. Mivel kis számmal való osztásnál nagy lehet a kerekítési hiba hatása, ezért kedvezőtlen, ha az a_{ii} elem kis abszolút értékű. Ennek elkerülésére szolgál a főelem-kiválasztás.

A **részleges főelem-kiválasztás** során megvizsgáljuk, hogy az adott oszlopban a főátló alatt van-e a főátlóbeli elemnél nagyobb abszolút értékű szám, és ha igen, akkor sorcserével a főátlóba hozzuk (lásd a 7.4. táblázatot).
Még jobban csökkenthető a számítási hiba a **teljes főelem-kiválasztás**sal. Ilyenkor a táblázatnak a főátlóbeli elemből jobbra és lefelé kiinduló legnagyobb négyzetes blokkjában keressük a legnagyobb abszolút értékű elemet (főelem). Ennek főátlóba hozásához esetleg sor- és oszlopcserét is végre kell hajtanunk. Az utóbbinál arra kell vigyázni, hogy megváltozik az oszlopok számozása. (Pl. ha a 4. oszlopot áthozzuk a 2. oszlopba, akkor

2	3	6	2	7
0	**1**	5	4	1
0	**5**	5	10	-15
0	**3**	2	5	-9
2	3	6	2	7
0	5	5	10	-15
0	1	5	4	1
0	3	2	5	-9

7.4. táblázat. Példa részleges főelemkiválasztásra. Mielőtt nullára transzformálnánk a 2. oszlopban lévő, főátló alatti elemeket, felcseréljük a 2. és a 3. sort. Ezzel a főátlóba kerül az oszlop legnagyobb abszolút értékű főátló alatti eleme.

onnantól kezdve a 2. oszlopban lévő számok az x_4 ismeretlen együtthatói lesznek!) Egy példát mutatunk be a 7.5. táblázatban.

2	3	6	2	7
0	**1**	**5**	**4**	1
0	**5**	**5**	**10**	-15
0	**3**	**2**	**5**	-9
2	3	6	2	7
0	5	5	10	-15
0	1	5	4	1
0	3	2	5	-9
2	2	6	3	7
0	10	5	5	-15
0	4	5	1	1
0	5	2	3	-9

7.5. táblázat. Példa teljes főelem-kiválasztásra. Megkeressük a legnagyobb abszolút értékű elemet a vastagon szedett blokkban. Ezt úgy hozhatjuk a főátlóba, hogy felcseréljük a 2. és a 3. sort, majd a 2. és a 4. oszlopot. Mostantól a 2. oszlop tartalmazza az x_4 együtthatóit, és a 4. oszlop az x_2 együtthatóit.

A Gauss–Jordan-elimináció

Ennél az eljárásnál a Gauss-féle kiküszöbölés lépései után a táblázat átalakítását tovább folytatjuk egészen addig, amíg az együtthatómátrix helyén az identitásmátrixot nem kapjuk. Ekkor a szabad tagok helyén leolvasható a megoldás.

Példa. Az előző példában szereplő egyenletrendszert oldjuk meg a Gauss–Jordan-féle eljárással.
Megoldás: A 7.6 táblázatot kapjuk. Az utolsó résztáblázat jobb oldali oszlopa szerint az összes ismeretlen 1-gyel egyenlő.

2	1	1	4
1	3	2	6
1	2	2	5
2	1	1	4
0	5/2	3/2	4
0	3/2	3/2	3
2	0	2/5	12/5
0	5/2	3/2	4
0	0	3/5	3/5
2	0	0	2
0	5/2	0	5/2
0	0	3/5	3/5
1	0	0	1
0	1	0	1
0	0	1	1

7.6. táblázat.

7.1.2. Faktorizációs eljárások

Faktorizációról akkor beszélünk, ha egy mátrixot két mátrix szorzatára bontunk. Tegyük fel, hogy az egyenletrendszer $A \in \mathbb{R}^{n\times n}$ együtthatómátrixa felírható

$$A = B \cdot C, \quad B, C \in \mathbb{R}^{n\times n}$$

alakban. Ekkor az $Ax = b$ egyenletrendszer megoldása ekvivalens a

$$By = b$$

és a

$$Cx = y$$

egyenletrendszer egymás utáni megoldásával. Nyilvánvalóan ennek csak akkor van értelme, ha a két rendszert könnyebb megoldani, mint az eredeti egyenletrendszert, azaz ha B és C „jó struktúrájú" mátrixok. Ez a helyzet például, ha háromszögmátrixokról van szó. Különösen akkor előnyös faktorizációs módszert alkalmazni, ha ugyanazon együtthatómátrixszal, de különböző alkalmakkor, különböző jobb oldalakkal is meg akarjuk oldani az egyenletrendszert.

A következőkben néhány konkrét faktorizációs módszert ismertetünk.

LU-faktorizáció

Ennek a felbontási módszernek az a lényege, hogy az A mátrixot $L \cdot U$ alakban írjuk fel, ahol L a főátlóban egyeseket tartalmazó alsóháromszög-mátrix, U pedig felsőháromszög-mátrix. Ekkor a megoldandó egyenletrendszerek:

$$Ly = b$$

és

$$Ux = y.$$

Az L és U meghatározásához jelölje a két mátrix kiszámítandó elemeit l_{ij} ill. u_{ij}. Annak kell teljesülnie, hogy a mátrixszorzást elvégezve az A mátrixot kapjuk, azaz fennálljon a

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ l_{21} & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ l_{n1} & l_{n2} & \dots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & u_{nn} \end{bmatrix}$$

mátrixegyenlőség. A bal oldalon elvégezzük a szorzást, és az elemeket egyenlővé tesszük az A mátrix megfelelő elemeivel.

Példa. Oldjuk meg LU-faktorizációval a

$$\begin{array}{rrrcl} 2x_1 & +x_2 & +x_3 & = & 4 \\ x_1 & +3x_2 & +2x_3 & = & 6 \\ x_1 & +2x_2 & +2x_3 & = & 5 \end{array}$$

egyenletrendszert.
Megoldás: Először meghatározzuk az

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

együtthatómátrix LU-felbontását. Az

$$\begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

egyenlőségnek kell fennállnia. A szorzatmátrix első oszlopát az A első oszlopával összehasonlítva az

$$u_{11} = 2,$$

$$l_{21}u_{11} = 1, \quad \text{azaz} \quad l_{21} = \frac{1}{2},$$

$$l_{31}u_{11} = 1, \quad \text{vagyis} \quad l_{31} = \frac{1}{2}$$

összefüggések adódnak. A második oszlop összehasonlításából

$$u_{12} = 1,$$

$$l_{21}u_{12} = 3, \quad \text{azaz} \quad u_{22} = 3 - \frac{1}{2} \cdot 1 = \frac{5}{2},$$

$$l_{31}u_{12} + l_{32}u_{22} = 2, \quad \text{vagyis} \quad l_{32} = \frac{2}{5} \cdot \left(1 - \frac{1}{2} \cdot 1\right) = \frac{3}{5},$$

és végül a harmadik oszlop összehasonlításából azonnal adódik, hogy

$$u_{13} = 1,$$

$$l_{21}u_{13} + u_{23} = 2, \quad \text{amiből} \quad u_{23} = \frac{3}{2},$$

$$l_{31}u_{13} + l_{32}u_{23} + u_{33} = 2, \quad \text{vagyis} \quad u_{33} = \frac{3}{5}.$$

A keresett tényezők tehát

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & \frac{3}{5} & 1 \end{bmatrix} \quad \text{és} \quad U = \begin{bmatrix} 2 & 1 & 1 \\ 0 & \frac{5}{2} & \frac{3}{2} \\ 0 & 0 & \frac{3}{5} \end{bmatrix}.$$

Az $Ly = b$ egyenletrendszer komponensenként kiírva

$$\begin{array}{llll} y_1 & & = & 4 \\ \frac{1}{2}y_1 & +y_2 & = & 6 \\ \frac{1}{2}y_1 & +\frac{3}{5}y_2 \quad +y_3 & = & 5 \end{array}$$

alakú. Az első, a második és a harmadik egyenletből

$$y_1 = 4, \quad y_2 = 4, \quad y_3 = \frac{3}{5}.$$

Végül az $Ux = y$, azaz a

$$\begin{array}{llll} 2x_1 \quad +x_2 \quad +x_3 & = & 4 \\ +\frac{5}{2}x_2 \quad +\frac{3}{2}x_3 & = & 4 \\ +\frac{3}{5}x_3 & = & \frac{3}{5} \end{array}$$

egyenletrendszer megoldása az utolsó egyenletből kiindulva:

$$x_3 = 1, \quad x_2 = 1, \quad x_1 = 1.$$

7.2. Iterációs módszerek

Az iterációs módszerek azon alapulnak, hogy az $Ax = b$ egyenletrendszert $x = Bx + r$ alakú egyenletrendszerré transzformáljuk, ahol $B \in R^{n \times n}$ és $r \in \mathbb{R}^n$. Vegyük észre, hogy

ekkor a megoldás az $f : \mathbb{R}^n \to \mathbb{R}^n$, $f(x) = Bx + r$ leképezés fixpontja. Az egyenletrendszer iterációs megoldása során olyan vektorsorozatot konstruálunk, amely az f leképezés fixpontjához, azaz a megoldásvektorhoz tart.

A matematika óráról ismert a fixponttétel, amelyet emlékeztetőül felidézünk.

7.2.1 Tétel. *Ha*

- *X teljes normált tér (más szóval Banach-tér),*
- *$f : X \to X$ és $D(f) = X$, azaz f az egész X-en értelmezve van,*
- *továbbá f kontrakció, azaz létezik olyan $q < 1$ szám, amely mellett*

$$\|f(x) - f(y)\| \leq q \cdot \|x - y\| \quad \forall x, y \in X,$$

akkor

- *f-nek létezik egyetlen fixpontja, ezt jelölje x^*,*
- *tetszőleges $x^{(0)} \in X$ elemből indulva az $x^{(k)} = f(x^{(k-1)})$ sorozat konvergens és x^*-hoz tart,*
- *továbbá a sorozat k-adik tagjának x^*-tól vett eltérésére igaz a következő formula:*

$$\|x^{(k)} - x^*\| \leq \frac{1}{1-q} q^k \|x^{(0)} - x^{(1)}\|.$$

Tekintsük most az $f(x) = Bx + r$ függvényt, ahol $B \in \mathbb{R}^{n \times n}$ és $r \in \mathbb{R}^n$. Ekkor $f : \mathbb{R}^n \to \mathbb{R}^n$, tehát nekünk most az $X = \mathbb{R}^n$ eset az érdekes. Ismeretes, hogy $\mathbb{R}^n$ bármelyik tanult vektornormával Banach-tér. Most az f függvény az egész $\mathbb{R}^n$-en értelmezve van. Vizsgáljuk meg, hogy f mikor kontrakció!

Legyen x és y két tetszőleges $\mathbb{R}^n$-beli vektor. Számítsuk ki az $f(x) - f(y)$ különbséget:

$$f(x) - f(y) = Bx + r - By - r = B(x - y).$$

Vegyük mindkét oldal valamely $\|\cdot\|_{\mathbb{R}^n}$ vektornormáját!

$$\|f(x) - f(y)\|_{\mathbb{R}^n} = \|B(x - y)\|_{\mathbb{R}^n} \leq \|B\| \cdot \|x - y\|_{\mathbb{R}^n},$$

ahol $\|B\|$ a B mátrix $\|\cdot\|_{\mathbb{R}^n}$ vektornorma által indukált mátrixnormája. Világos, hogy ha $\|B\| < 1$, akkor az f függvény $q := \|B\|$ mellett kontrakció. Ezért a fixponttétel

értelmében $x^{(k)} \to x^*$. Tehát a konvergenciának elégséges feltétele, hogy valamelyik indukált mátrixnormában $\|B\|$ kisebb legyen, mint 1.

Természetes módon vetődik fel a következő két kérdés:

1. Hogyan tudunk egy $Ax = b$ egyenletrendszert $x = Bx + r$ alakú rendszerré transzormálni?

2. Mikor teljesül, hogy az így kapott B mátrix valamelyik indukált mátrixnormában kisebb, mint 1?

A továbbiakban néhány konkrét módszert ismertetünk.

7.2.1. Jacobi-iteráció

Bontsuk fel az A mátrixot $A = L + D + U$ alakban, ahol L és U alsó ill. felső háromszögű mátrix, a főátlóban nulla elemekkel, D pedig diagonálmátrix. Tegyük fel, hogy D főátlójában nincs nulla elem, azaz D invertálható. Ezzel az egyenletrendszer

$$(L + D + U)x = b.$$

Vigyük át a jobb oldalra az L és az U mátrixot tartalmazó tagokat:

$$Dx = -(L + U)x + b,$$

majd szorozzunk D inverzével:

$$x = -D^{-1}(L + U)x + D^{-1}b.$$

Ezzel sikerült az egyenletrendszert $x = Bx + r$ alakra hoznunk, ahol speciálisan

$$B = -D^{-1}(L + U) \quad \text{és} \quad r = D^{-1}b.$$

A Jacobi-iteráció k-adik lépésében az

$$x^{(k)} = -D^{-1}(L + U)x^{(k-1)} + D^{-1}b$$

összefüggés szerint számolunk.
Érdemes koordinátákkal is kiírni a k-adik iterációs vektort. Jelöljük az A mátrix i-edik

sorában és j-edik oszlopában lévő elemet a_{ij}-vel. Ezzel

$$\begin{aligned} x_i^{(k)} &= (-D^{-1}[(L+U)x^{(k-1)} - b])_i = -\frac{1}{a_{ii}} \cdot ((L+U)x^{(k-1)} - b)_i \\ &= -\frac{1}{a_{ii}} \cdot (((L+U)x^{(k-1)})_i - b_i) = -\frac{1}{a_{ii}} \left(\sum_{j=1, j\neq i}^{n} a_{ij} \cdot x_j^{(k-1)} - b_i \right) \quad i = 1, 2, \ldots, n \end{aligned}$$

Feladatok.

1. Mutassuk meg, hogy a Jacobi-iterációval kapott $B = -D^{-1}(L+U)$ mátrix sornormája pontosan akkor kisebb 1-nél, amikor az eredeti egyenletrendszer A együtthatómátrixára

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

azaz amikor minden sorában a főátlóban lévő elem abszolút értékben nagyobb a többi elem abszolút értékének összegénél. (Ekkor az A mátrixot szigorúan diagonálisan domináns mátrixnak nevezzük.)

2. Tekintsük a

$$\begin{array}{rrrcl} 4x_1 & -x_2 & & = & 0 \\ -x_1 & +4x_2 & -x_3 & = & 6 \\ & -x_2 & +4x_3 & = & 2 \end{array}$$

egyenletrendszert. Legyen a Jacobi-iteráció kiindulási vektora $x^{(0)} := (0,0,0)$. Konvergens-e a Jacobi-iteráció? Ha igen, számítsuk ki az $x^{(1)}$ és $x^{(2)}$ közelítést. Adjuk meg a k-adik iteráció képlethibáját.

Megoldás: Az A mátrix szigorúan diagonálisan domináns, ezért B sornormája kisebb 1-nél, így az iteráció konvergens. Az A együtthatómátrix felbontásában szereplő mátrixok:

$$L = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}.$$

Ezekből

$$B = -D^{-1}(L+U) = \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & 0 \end{bmatrix}, \quad r = D^{-1}b = \begin{bmatrix} 0 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix}.$$

Az első iteráció

$$x^{(1)} = Bx^{(0)} + r = r = \begin{bmatrix} 0 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix}.$$

A második iteráció

$$x^{(2)} = Bx^{(1)} + r = \begin{bmatrix} \frac{3}{8} \\ \frac{1}{8} \\ \frac{3}{8} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{8} \\ \frac{13}{8} \\ \frac{7}{8} \end{bmatrix}.$$

A sornormát, amelyben B-ről tudjuk, hogy kisebb, mint 1, a maximumnorma indukálja. Így a k-adik iteráció képlethibája maximumnormában

$$\|x^{(k)} - x^*\| = \max_i |x_i^{(k)} - x_i^*| \leq \frac{\|B\|^k}{1 - \|B\|} \cdot \max_i |x_i^{(1)} - x_i^{(0)}| = 3 \cdot \left(\frac{1}{2}\right)^k,$$

ahol felhasználtuk, hogy a B mátrix sornormája $\|B\| = \frac{1}{2}$.

7.2.2. Gauss–Seidel-iteráció

A Gauss–Seidel-iterációnál is $L + D + U$ alakban bontjuk fel az A mátrixot, de most az Ux tagot visszük át a jobb oldalra, és $(L + D)$ inverzével szorzunk. Így az

$$x = -(L + D)^{-1}Ux + (L + D)^{-1}b$$

alakhoz jutunk. Azaz most $B = -(L + D)^{-1}U$ és $r = (L + D)^{-1}b$.
Ha a koordinátánénti számolásnál ügyesen csoportosítjuk a műveleteket, akkor elkerülhetjük az $L + D$ mátrix invertálását. Ehhez az

$$Lx^{(k)} + Dx^{(k)} = -Ux^{(k-1)} + b$$

iterációs egyenletben először vonjunk ki mindkét oldalból $Lx^{(k)}$-t, majd szorozzunk D inverzével. Ezzel

$$x^{(k)} = -D^{-1}(Lx^{(k)} + Ux^{(k-1)} - b),$$

amelynek i-edik koordinátája

$$x_i^{(k)} = -\frac{1}{a_{ii}} \cdot (Lx^{(k)} + Ux^{(k-1)} - b)_i = -\frac{1}{a_{ii}} \left(\sum_{j=1}^{i-1} a_{ij} x_j^{(k)} + \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} - b_i \right), \; i = 1, 2, \ldots, n.$$

Vegyük észre, hogy a jobb oldali első összegben a $k+1$-edik közelítésnek csak azon elemei szerepelnek, amelyeket már az előző, $i - 1$-edik lépésben kiszámoltunk.